Semi-supervised training of lip reading models using lattice-free MMI

Sonia M Marshall

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2022

Abstract

In this project I focused on training and evaluating models for lip reading, or visual speech recognition, using audio-visual datasets. Since large transcribed audio-visual datasets are costly to build and supervised lip reading models are found to not generalise well to data from a mismatched domain, I investigated how we can make use of the large quantities of untranscribed video data that exist. I generated a multi-genre audio-visual dataset containing 223 hours of training data, and used this to perform semi-supervised training of both acoustic and lip reading models using the LF-MMI training criterion with HMM-DNN based models in Kaldi. While supervised models trained on TED talks did not generalise well to this challenging multi-genre dataset in either modality, through semi-supervised training the word error rate for the acoustic model reduced from 42.1% to 36.8% and for the lip reading model reduced from 93.4% to 91.3%. This improvement, though small, is promising for this first use of the semi-supervised LF-MMI technique in the field of lip reading.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sonia M Marshall*)

Acknowledgements

Many thanks go to my supervisor Peter Bell for the support over the two years of the MInf project, and to Ondřej Klejch, the Kaldi genius, whose help has been invaluable.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Lip reading or visual speech recognition is a variation on typical automatic speech recognition (ASR) where lip movements are analysed to recognise speech. This can be used in different scenarios to typical ASR, for example when the audio is noisy or even unavailable it is still possible to perform lip reading as the visuals are not affected. The combination of the acoustic and visual modalities into audio-visual speech recognition (AVSR) provide very interesting possibilities for improving performance in such cases of noisy audio (Afouras et al., 2018a). Lip reading and AVSR have practical applications for uses such as video captioning or improving hearing aids.

Lip reading involves linguistic challenges such as words that are difficult to distinguish visually as well as the general challenges of image processing: coping with varying background, lighting, speaker appearance and position. Therefore, large quantities of varied data are required to build robust lip reading models. However, generating an audio-visual dataset is a huge task. First of all, hand-transcribing utterances is costly as it requires a lot of time and may require expertise - this is also the case for speech datasets (Sperber et al., 2017). Additionally, having two modalities increases the challenge: not only must the data be transcribed, but the quality of both the audio and video components must be good, and they must be aligned (Chitu and Rothkrantz, 2007). While there are now more large audio-visual datasets available than there used to be (Afouras et al., 2018c; Shillingford et al., 2018; Ephrat et al., 2018), it is still not a huge number of them that are publicly available. We may need to perform lip reading in a different domain to these existing datasets, and in this case it is difficult to generate your own full dataset in the new genre.

We therefore wish to find ways of using the available data to generalise to new domains. A similar problem of lack of transcribed data is found in the field of speech recognition in under-resourced languages, and one solution to this problem is performing semi-supervised training to make use of untranscribed data (Xu et al., 2016). While there is a lack of accurately transcribed video data there is plenty of untranscribed video publicly available, for example on YouTube, so this is a viable method for lip reading. In this project I therefore investigate the benefits of semi-supervised training for lip

reading, using a transcribed dataset of one domain and an untranscribed dataset of a mismatching domain.

It is possible to perform semi-supervised training for the latest end-to-end deep neural network (DNN) ASR models (Weninger et al., 2020), however this involves predicting exact pseudo-transcriptions and filtering out any that are below a confidence threshold. On the other hand, using the more conventional hybrid HMM-DNN systems allows the use of lattices to incorporate uncertainty into the transcriptions, so all data can be used. I therefore use semi-supervised LF-MMI training (Manohar et al., 2018) of HMM-DNN lip reading models in Kaldi for my experiments.

1.2 Previous work

My work in this project builds upon the knowledge and insights gained from my previous research and experiments. In the first year of the project I researched the field of lip reading and the audio-visual datasets that have been produced for this task. I investigated how well a lip reading model trained on one dataset could generalise to an unseen dataset of a mismatched domain.

I used a lip reading model which had been pre-trained by Afouras et al. (2018b) on the LRS2 dataset (Afouras et al., 2018a). I evaluated the performance of this model on the LRS3-TED dataset (Afouras et al., 2018c). It achieved a word error rate (WER) of 80.7% on the LRS3-TED data, compared to 48.8% when evaluated on the LRS2 test data. Although the LRS3-TED dataset is known to be more challenging (see Section 3.1), this still demonstrates poor generalisation of the lip reading model to a mismatching domain. I then improved the performance by doing further supervised training of the lip reading model using a subset of the LRS3-TED training data. This resulted in a reduction in WER to 78.0%.

This DNN lip reading model was built in TensorFlow and was based on the Transformer architecture (Vaswani et al., 2017). As an end-to-end model it takes in the full sequence of input video frames and outputs character probabilities, which are then decoded into the final output sentence. Use of a language model during decoding is optional due to the Transformer architecture which implicitly learns a language model during training.

For further detail on the MInf1 project please read the full report (Marshall, 2021).

1.3 Goals and achievements

Similarly to last year, this project focused on building lip reading models and improving their generalisation to datasets with a mismatching domain. However instead of further pursuing the end-to-end DNN methods, this year I worked with hybrid HMM-DNN models. In particular, I focused on performing semi-supervised training of lip reading models. The goals therefore shifted slightly in comparison to those noted in the Future Work section of the MInf1 report (Marshall, 2021).

The goals of this project were:

- Produce an audio-visual dataset
- Train an initial lip reading model on one dataset and evaluate it on another dataset with a domain mismatch
- Improve performance of the model on the mismatched dataset by experimenting with semi-supervised LF-MMI training
- Combine ASR and lip reading to perform audio-visual speech recognition (if time)

I achieved these goals: I produced a 223 hour Multi-Genre Broadcast (MGB) training dataset from the MGB Challenge dataset (Bell et al., 2015) using the AVSR pipeline developed by Paul et al. (2021). I used SyncNet (Chung and Zisserman, 2016) to extract visual features for use in lip reading models. I trained an acoustic and a lip reading model on the LRS3-TED dataset, and evaluated them on the MGB test dataset. They achieved 42.1% and 93.4% WER respectively, compared to 4.7% and 70.6% on the LRS3-TED test set, demonstrating the large mismatch and poor generalisation between domains. I then performed semi-supervised LF-MMI training of acoustic and lip reading models in Kaldi (Povey et al., 2011), achieving 36.8% and 91.3% WER respectively on the MGB test set. This demonstrated an improvement in performance via semi-supervised training even when the initial model demonstrated such poor generalisation to the MGB dataset.

The original intention at the start of this MInf project was to combine ASR and lip reading into audio-visual speech recognition. However, in the end I continued to focus on lip reading, to fully investigate the possibilities of semi-supervised training. Although I did not implement audio-visual speech recognition, I did make use of acoustic information in the semi-supervised training to improve the lip reading models.

Chapter 2

Technical background

In this project I used Kaldi (Povey et al., 2011), a framework for building HMM-based speech recognition models using weighted finite-state transducers (WFSTs) (Mohri et al., 2008). In this chapter I describe the relevant concepts for how speech recognition is performed in Kaldi, as well as details about time delay neural networks (Waibel et al., 1989), the LF-MMI training criterion (Povey et al., 2016) and the lip synchronisation network SyncNet (Chung and Zisserman, 2016) which I use in my experiments.

2.1 Weighted finite-state transducer speech recognition

Speech can be represented as a sequence of feature vectors, or observations. Hidden Markov models (HMMs) can be used to build acoustic models, in which each state represents a speech unit and the topology defines the allowed transitions between them. In Kaldi, each phone is represented by 3 states. HMMs are generative models, so at each time step in a sequence, one state emits an observation. The HMM can be decoded to find the most probable state sequence given the sequence of observations, and therefore produce a transcription of the input speech. In HMM-GMM models, the observation probabilities of each state are estimated by a Gaussian mixture model (GMM). In HMM-DNN models, these probabilities are estimated by a DNN.

In Kaldi (Povey et al., 2011), an acoustic model is built by first training a monophone HMM-GMM model. This context-independent model is then used to generate an alignment of the training transcripts, to be used to train a context-dependent triphone model. This better model is used to realign the transcripts again, and this repeats several times, creating progressively better models. The training of these models can include transformations of the input data using Linear Discriminant Analysis (LDA) (Somervuo, 2003) and Maximum Likelihood Linear Transform (MLLT) (Gales, 1998), and also speaker adaptive training. In an HMM-DNN system, the final HMM-GMM model is used to generate an alignment which is used to train a DNN.

Each component in the speech recognition system is represented by a WFST (Mohri et al., 2008). The HMM acoustic model is represented by the H WFST which maps HMM states to context-dependent phones. The C WFST maps context-dependent

phones to context-independent phones. The pronunciation lexicon is represented by L, mapping phones to words. Finally, the language model is converted into the grammar G, mapping words to words. These WFSTs are all composed together to form a single decoding graph: $HCLG = (H \circ (C \circ (L \circ G)))$. A sequence of minimization and determinization steps (not shown) are also performed during this composition to reduce the size of the final graph. The $HCLG$ graph can then be used to decode utterances.

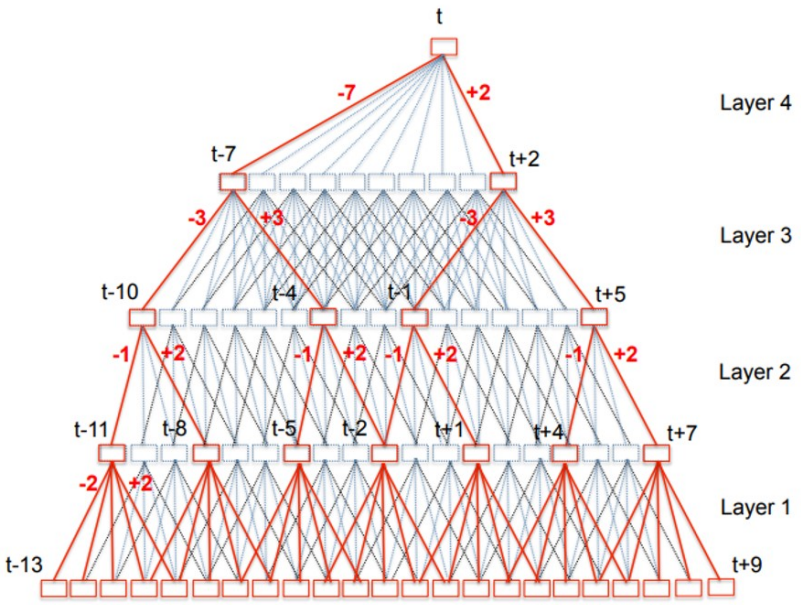
The common metric for speech recognition performance is the word error rate (WER):

$$WER = \frac{S+D+I}{N}$$

where S is the number of substitutions, D the number of deletions and I the number of insertions. S, D and I are computed by comparing recognised transcriptions to their reference transcriptions. N is the number of words in the reference transcriptions.

2.2 Time delay neural networks

Figure 2.1: Example computation in a TDNN with sub-sampling (red) and without (blue) (Peddinti et al., 2015).



Since speech contains long term dependencies it is very useful to be able to model temporal context in speech recognition. There are different approaches to modelling context, I described recurrent neural networks (RNNs) and encoder-decoder architectures in Marshall (2021). In this project I take a different approach, using time delay neural networks (TDNNs) (Waibel et al., 1989).

Subsampled TDNNs (Peddinti et al., 2015) are the standard architecture used in the HMM-DNN models in Kaldi. In the Kaldi nnet3 library the TDNNs are trained using

the LF-MMI (Povey et al., 2016) training criterion, and this is also referred to as the 'chain' model. There are also variations on the plain TDNN, such as the factored TDNN (Povey et al., 2018) and TDNN-LSTMs (Peddinti et al., 2017).

A TDNN is a feed-forward neural network in which hidden units process multiple frames of input at once, referred to as windows of context. Since each layer in the network processes a window of context from the layer below, the higher layers can learn wider temporal dependencies compared to the lower layers. Processing every single hidden unit activation in the network would take considerably more computation and storage compared to a standard DNN. However, since hidden units at neighbouring time steps process overlapping context windows the activations can be sub-sampled to reduce the computation and storage without losing useful information. This is shown in Figure 2.1.

Since TDNNs are a feed-forward network they can be more easily parallelised across GPUs compared to RNNs, so are a good option for modelling wide temporal contexts with reduced computational costs.

2.3 Lattice-free maximum mutual information training

Lattice-free maximum mutual information (LF-MMI) (Povey et al., 2016) is an efficient sequence-discriminative training criterion, which was found to improve performance compared to the frame-based cross-entropy (CE) training criterion. This type of training is suitable for neural networks, for example the TDNNs used in Kaldi. The lattice-free version of MMI (Bahl et al., 1986) takes inspiration from connectionist temporal classification (CTC) (Graves et al., 2006) training and reduces the computational costs of MMI training.

In MMI training, the objective function maximises the likelihood of the sequence of observations given the correct word sequence for the utterance (computed using a numerator), while minimising the likelihood of the observation sequence given all possible word sequences (computed using a denominator).

In LF-MMI the numerators and denominator are represented as *HCLG* WFSTs. The key contribution is that the denominator WFST is trained from scratch using a language model, removing the need to compute lattices representing all possible word sequences using an initial CE trained system. This, and parallelising the computation across GPUs, greatly reduces the computational costs. Numerator WFSTs are produced for each training utterance, using lattices produced by an HMM-GMM model which encode all possible pronunciations of the utterance.

There are some other differences to the standard techniques described in Section 2.1 used to make LF-MMI training more efficient, these are described in Povey et al. (2016).

The LF-MMI training criterion is a highly suitable method to be used in semi-supervised training, as the numerator lattices can be used to encode uncertainties of transcriptions.

2.4 SyncNet

Chung and Zisserman (2016) developed SyncNet, a network primarily built to synchronise audio and video when there is a lag between them. Their novel contribution to this task is that they use only the audio and video to find the synchronisation. They use a convolutional neural network (CNN) with two streams to learn embeddings of the input data: the audio stream processes Mel-frequency cepstral coefficients (MFCCs) and the visual stream processes grayscale images of mouth regions. Two sets of features are output from the final fully connected layers, these are then compared for similarity to determine if the audio and video are synchronised.

The embeddings of the input video and audio learnt by the network can be applied to other related tasks, such as active speaker detection (as used by Paul et al. (2021) in their AVSR pipeline), or lip reading. Chung and Zisserman (2016) achieved state-of-the-art lip reading results on the OuluVS2 dataset (Anina et al., 2015), which is a promising first investigation into using these features. This dataset is however quite small and constrained. In this project we are similarly using the SyncNet visual features to perform lip reading, but on the much more difficult MGB dataset.

Chapter 3

Literature review

In this chapter I describe the audio-visual datasets relevant to this project, as well as give an overview of related work in the field of lip reading and previous approaches to semi-supervised training in speech recognition.

3.1 Datasets

In order to train lip reading models we require audio-visual (AV) datasets containing video clips of speakers' faces. Over the years there has been a shift from more constrained datasets recorded specifically for this purpose (Cooke et al., 2006), towards larger and more diverse datasets produced from existing video content (Afouras et al., 2018a,c; Shillingford et al., 2018). Here I give a brief description of just those relevant to this project - a more detailed overview of the history of audio-visual datasets is given in my MInf1 report (Marshall, 2021).

The LRS3-TED (Afouras et al., 2018c) dataset contains cropped videos of speakers' faces from TED talks, and transcriptions of the videos. This challenging dataset contains approximately 152,000 utterances from over 5000 different speakers, and varying viewpoints as the speaker moves their head. I used this dataset to train supervised acoustic and visual models.

The LRS2 (Afouras et al., 2018a) dataset contains similar video clips with a variety of angles of speakers' faces, but taken from BBC TV news programmes and talk shows. It contains approximately 144,000 utterances, and contains slightly shorter sentences than LRS3-TED.

Soon after the LRS3-TED (Afouras et al., 2018c) dataset was produced, a baseline was set for lip reading performance on it by Afouras et al. (2018a). The best performance was 58.9% WER, achieved using the sequence-to-sequence trained Transformer architecture from Afouras et al. (2018b). Their model fine-tuned on LRS2 data achieved 48.3% WER on the LRS2 test set, demonstrating that for lip reading the LRS3-TED dataset is more challenging than LRS2. Their acoustic models performed better on LRS3-TED (8.3%) compared to LRS2 (9.7%), suggesting that the audio in the LRS3-TED may be cleaner, although the difference is not as noticeable as for the visual aspect.

The Multi-Genre Broadcast (MGB) Challenge dataset (Bell et al., 2015) was created in 2015 for an ASR challenge. It contains the full videos of seven weeks of BBC TV broadcasts of various genres, and their subtitles. The dataset contains training, development and evaluation datasets, and also defines a 'short' training dataset, which is a one week subset of the full training dataset. I used this dataset to generate my own AV dataset. Due to its multi-genre nature, this is an even more challenging dataset than LRS2 and LRS3-TED. This is demonstrated by the results of my supervised acoustic model (trained on LRS3-TED data) which achieved 4.7% WER on LRS3-TED, 12.9% on LRS2 and 42.1% on MGB (see Section 6.1 for more details).

3.2 Lip reading

Lip reading methods have evolved over the years as new developments in deep neural networks have occurred, much in the same way as ASR methods have. Traditionally, handcrafted features were extracted from the video and fed into HMM-based systems to perform speech recognition (Potamianos et al., 1998). More recently, end-to-end trainable DNN systems have been developed, where the network extracts features itself as part of the training process. Common end-to-end DNN architectures include using convolutional neural networks (CNNs) as a front end for visual feature extraction, then long short-term memory (LSTM) networks or other sequence modelling architectures such as Transformers for recognition (Assael et al., 2016; Chung et al., 2017). In between these two extremes are the hybrid HMM-DNN systems, where a DNN replaces the GMM in the traditional HMM-GMM systems (Thangthai et al., 2015). For more in depth background on end-to-end lip reading methods please refer to my MInf1 report (Marshall, 2021).

Examples of traditional methods that can be used to extract visual features from images of faces or lips include geometric feature extraction (Petajan, 1984), appearance-based approaches such as active appearance models (AAM) (Cootes et al., 2001) and image transforms (Potamianos et al., 1998). However, techniques have moved on and nowadays visual features are usually extracted by a DNN such as a CNN, and these features can be used in HMM-based systems as well as in end-to-end models. These features do not require the expense of hand-labelling and are found to improve performance compared to methods such as image transforms (Noda et al., 2014). In this project I use features extracted by a CNN in an HMM-DNN system.

The Kaldi toolkit (Povey et al., 2011) as described in Chapter 2 was built for speech recognition research. Although originally designed for building acoustic models, it can also be used to build visual speech recognition models. Thangthai et al. (2015) investigated lip reading and AVSR using HMM-DNN based models in Kaldi, showing that they improved performance compared to HMM-GMM based systems. However, they used only a small corpus containing utterances from one speaker, and used AAM features, which require hand-labelling. Abdelaziz et al. (2017) performed AVSR using HMM-DNN models in Kaldi, again using the more traditional dimensionality reduction methods to extract features (Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA)).

A more interesting comparison is performed by Yu et al. (2020). They compare hybrid and end-to-end DNNs systems for lip reading and audio-visual speech recognition, motivated by the findings of Lüscher et al. (2019) that hybrid ASR systems are capable of out-performing end-to-end acoustic models. Their hybrid LF-MMI trained TDNN lip reading model built in Kaldi and using visual features extracted by LipNet (Assael et al., 2016) achieves competitive results (48.86%) on the LRS2 dataset compared to the Transformer architectures of Afouras et al. (2018a) which achieved 48.3%. The sequence-to-sequence Transformer model used was the same one I used in last year's project (Marshall, 2021), which I found to be very genre specific, struggling to generalise to a new domain. The LF-MMI trained TDNN from Yu et al. (2020) also outperformed their cross-entropy trained TDNN. These findings are all good motivations for my use of LF-MMI training of HMM-DNN models in Kaldi.

3.3 Semi-supervised training

Semi-supervised training can be used to develop ASR systems using a small quantity of transcribed data and large amount of untranscribed data. This works by training an initial model on the transcribed data, then using this initial model to predict pseudo-labels for the untranscribed data. The initial model can then be trained further using this pseudo-labelled data, which can be seen to improve performance (Lamel et al., 2002). This is useful in any case where there is a lack of transcribed speech data, such as under-resourced languages (Xu et al., 2016), or even new domains in well-resourced languages, as it reduces the time and expense spent creating hand-labelled datasets. Semi-supervised training has been implemented with end-to-end DNN models (Weninger et al., 2020) as well as the traditional HMM-GMM based systems (Lamel et al., 2002).

A problem with training using pseudo-transcriptions is that if they are wrong, the model is being trained to recognise utterances incorrectly - and these incorrect transcriptions have more of an effect in systems trained with a discriminative criterion, such as MMI, compared to maximum likelihood estimation (Yu et al., 2007). To reduce this problem, usually utterances below a certain confidence threshold are removed. However, this can mean that the challenging utterances are removed from the training data, leaving the easier utterances and therefore the final model will not be able to cope with any challenging unseen utterances.

Manohar et al. (2018) solve this problem by introducing the LF-MMI (Povey et al., 2016) training criterion to semi-supervised training for ASR. Instead of using pseudo-transcriptions, the untranscribed data is decoded and the entire resulting lattice for each utterance is used in the semi-supervised training - this allows uncertainty to be incorporated into each training utterance, rather than discarding an utterance entirely because its transcription is uncertain. They compared several different methods of generating the lattices, and all were found to give better performance compared to using the best path word sequence (i.e. pseudo-transcriptions). The various methods of generating lattices did not show much difference between themselves - in my project I use their naïve lattice splitting method.

Wallington et al. (2021) demonstrate that in semi-supervised LF-MMI training the quality of the language model used can have more of an impact on performance than the quality of the initial acoustic model. This is promising for the case of semi-supervised lip reading, since in cases where we do not have aligned transcriptions for an audio-visual dataset, we may still have the unaligned video captions and can therefore still generate language models well suited to the target domain even if we have an initial lip reading model that does not generalise well to the domain.

Despite these semi-supervised training techniques being widely used for improving acoustic models, far less research has been done into using these techniques for lip reading or audio-visual speech recognition. In Afouras et al. (2020) they use a similar concept, cross-modal knowledge distillation, where they use transcriptions of unlabeled audio-visual data generated by an ASR system to train a lip reading model. There are instances of untranscribed data being leveraged to learn visual features that better represent the audio-visual data through semi-supervised (Su et al., 2018) and self-supervised (Sheng et al., 2021) learning. While these better features should in turn improve lip reading model performance, this is not the same as directly improving the lip reading models using semi-supervised training. Therefore, as far as I am aware, my project is the first to experiment with semi-supervised training of HMM-DNN based lip reading models using the LF-MMI training criterion.

Chapter 4

Dataset generation

In order to produce an audio-visual dataset I made use of the audio-visual speech recognition (AVSR) pipeline developed by Paul et al. (2021). I input videos from the News, Documentary, Childrens and Drama genres of the MGB Challenge dataset into the pipeline to generate an MGB audio-visual dataset. This chapter describes details of the pipeline, the dataset produced and practical issues encountered in the process.

4.1 Audio-visual speech recognition pipeline

The AVSR pipeline (Paul et al., 2021) takes in a full input video and processes it to output cropped video clips of talking faces. This involves first segmenting the video into scenes using histogram based shot detection, then for each scene facetracks are extracted using the SyncNet (Chung and Zisserman, 2016) face detector. For each facetrack SyncNet is used to detect segments containing an active speaker. This is done by extracting acoustic and visual features (as described in Section 2.4) and computing a confidence score. If the score is above a given confidence threshold the audio likely matches the video, i.e. the speaker is active, and the talking segment is saved as a video clip. These clips make up the final AV dataset.

Later on, in my semi-supervised training experiments, I extract SyncNet features from the videos to use as features for lip reading (this is described more fully in Section 5.1). Since these features were already computed during the dataset generation it would be beneficial for the future to make edits to the pipeline, so that the features are saved as the dataset is being generated, instead of needing to be computed again for the experiments.

The SyncNet confidence score for an audio/video pair should be close to 0 for a video of a non-speaker, while video containing the active speaker should have a high confidence score, with an example score of 7.56 given in Chung and Zisserman (2016). The default confidence threshold set in the AVSR pipeline is 2.5. Before deciding to use this, I compared the output of the pipeline on one programme from each genre using a threshold of 2.0 and 2.5. The lower threshold produced a few extra video clips compared to the 2.5 threshold, as would be expected. These extra clips were not useful, containing facetracks of non-speakers, so I decided to use the default threshold.

As I ran the pipeline I discovered a synchronisation error was occurring so that some clips output by the pipeline had audio that did not correspond the video. In cases where the audio was very short, this resulted in very short video clips being produced (for example, 6 frames long), and in some cases a video was trimmed to 0 frames, which caused an exception to occur. The pipeline was updated by Paul et al. (2021) to resolve these issues in response to my debugging.

4.2 Multi-Genre Broadcast training set

I generated the Multi-Genre Broadcast (MGB) training set by feeding the MGB Challenge training data through the AVSR pipeline. The number of utterances, duration and retention rate for each genre in the MGB training dataset is shown in Table 4.1. The original duration is the number of hours of video data of that genre in the original MGB Challenge training dataset, while the final duration is the duration of the MGB training set produced by the AVSR pipeline. The retention rate is the percentage of the original video that made it through the pipeline. The MGB training set contains 223 hours of data.

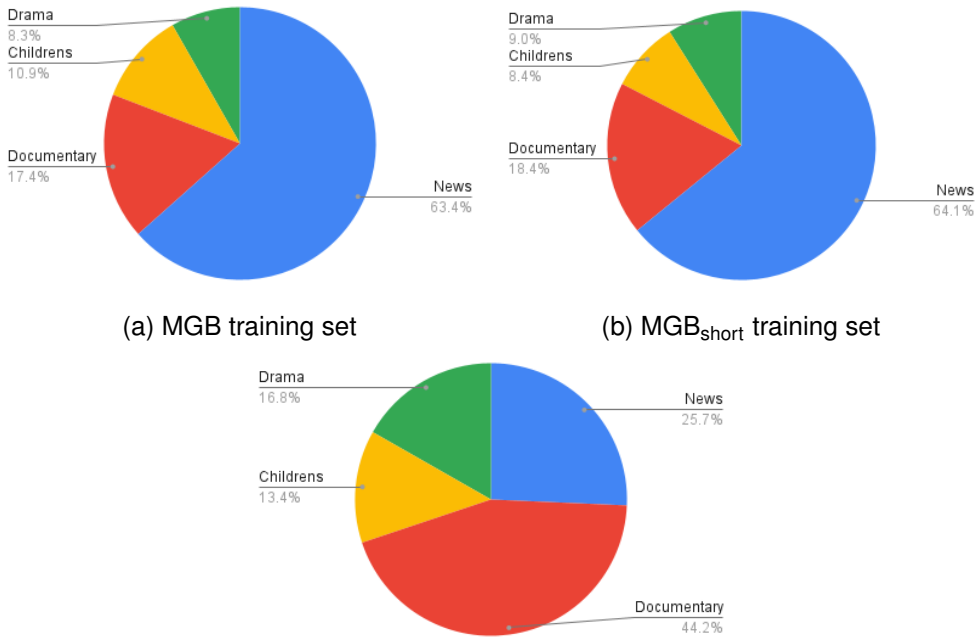
I also generated the MGB short training set, containing 27 hours of data. This is a sample of the full dataset, which I created for the purpose of having a smaller dataset that could be produced and experimented with more quickly than the full dataset. I produced it using a continuous one week sample of the continuous seven weeks of TV broadcasts in the MGB Challenge training set. Since the schedule of TV broadcasts is similar from week to week, the short dataset is a representative sample of the full dataset. This can be seen from the charts in Figure 4.1, which give a genre breakdown of the datasets, and show that the percentages of each genre are very similar between the full and short training sets.

Figure 4.1 also shows that the News genre dominates the training datasets. This means that models trained using these datasets may be slightly biased towards recognising the News genre. This imbalance is partly present because the MGB Challenge dataset contains more hours of News broadcasts than any other genre, and partly because the News genre had a higher retention rate than any other genre. The reason for this higher retention rate is discussed in Section 4.4.

The MGB training dataset contains only video clips, no corresponding transcriptions. I

Table 4.1: Number of utterances, duration and retention rate of the MGB full training set.

Genre	**# utterances**	**Original duration (hours)**	**Final duration (hours)**	**Retention rate (%)**
News	62727	351.55	141.75	40.3
Documentary	29369	214.26	38.84	18.1
Childrens	21952	168.65	24.36	14.4
Drama	26731	107.90	18.47	17.1
All	140779	842.36	223.43	26.5



(a) MGB training set

(b) $\text{MGB}_{\text{short}}$ training set

(c) MGB test set

Figure 4.1: The split of genres that make up the MGB datasets.

use this dataset for semi-supervised training, so transcriptions are not needed. In this case since the BBC subtitles for the MGB Challenge videos are available, it would be possible to create transcriptions. These could be interesting to have in order to check how good the semi-supervised decoding of the training data is compared to reference transcriptions. However, the subtitles are the captions from the original broadcast: they are not hand-transcribed so may not be reliable, and they may not be correctly aligned with the video. Since this would require extra work to align the captions using SyncNet (Chung and Zisserman, 2016) and it was not necessary for this project, I decided not to produce these transcriptions.

4.3 Multi-Genre Broadcast test set

The MGB test set was generated from the MGB Challenge development dataset by Paul et al. (2021) using an older version of the AVSR pipeline. The duration and number of utterances for each genre in the MGB test dataset are shown in Table 4.2, along with the statistics for the LRS2 and LRS3 test sets.

Unlike the training data, the MGB Challenge development dataset includes hand-transcribed and aligned transcriptions, and so this test dataset contains video clips and their corresponding transcriptions. It is to be noted that the original reference transcriptions must still be processed to generate transcriptions for the output videos from the pipeline. The latest version of the pipeline does not process these transcriptions so I could not generate a new test set and used this older version instead.

Figure 4.1c shows the percentage of each genre present in the MGB test set. Noticeably,

Table 4.2: Number of utterances and duration of the LRS2 test set, LRS3 test set and the MGB test set (including a genre breakdown).

Test Set	**# utterances**	**Duration (hours)**
LRS2	1243	0.59
LRS3	1321	0.86
MGB	4244	2.62
News	783	0.67
Documentary	1851	1.16
Childrens	587	0.35
Drama	1023	0.44

the dominant genre in the test set is Documentary, not News as in the training set. In fact, the test set contains less than half the proportion of News clips that are in the training set. This means there is a mismatch in the genre split of the training and test sets. This imbalance is present from the original MGB Challenge development set, which contains more Documentary programmes than programmes of any other genre. If I were to regenerate the test set, I would consider leaving out some of the Documentary videos in order to have a more balanced test set.

Currently there exist only the MGB training and test sets, not a validation set. It would be preferable in future to use this test set as a validation set, and generate a separate test set using the MGB Challenge evaluation dataset.

4.4 Genre differences

Figure 4.2 shows frames from videos in each of the four genres: News, Documentary, Childrens and Drama. Note the variety of backgrounds (indoors and outdoors), lighting and face angles present in just four different programmes.

The four genres used were selected by Paul et al. (2021) as the most interesting ones to work with, since they provide a variety of styles of content. Choosing a selection of genres to work with also reduces the quantity of data to process, which is useful to

Figure 4.2: Example frames from videos in the MGB training set. Source programmes and genres from left to right: BBC London News (News), Around the World in 80 Gardens (Documentary), Blue Peter (Childrens), Doctor Who (Drama).

reduce the time and resources spent running experiments. In a real-world scenario the other genres (Comedy, Advice, etc.) could additionally be utilised to make the most of the available data.

The News genre has a retention rate more than double the retention rate of other genres. This is perhaps unsurprising, since most News programmes consist of a newsreader sitting facing straight into the camera, which is ideal for extracting face tracks. This is in contrast to the Drama genre, which may have more cuts between different filming angles, and the speakers may be moving around more.

The Childrens genre has the lowest retention rate, which may be due to the fact that it contains content such as animated cartoons and puppet shows, where there are not many human faces. The AVSR pipeline does actually extract some facetracks from the animated cartoons. I have kept these facetracks, but it would be worth considering removing them for future projects. They are unlikely to provide useful training data for lip reading, unless the lip movements are animated very accurately. The speech content of the Childrens genre is very varied, visually and acoustically, containing children speaking, adults speaking in an exaggerated child-friendly manner, people singing and even vampires. These types of speech may be difficult for a system trained on more typical adult speech to recognise.

The Documentary genre may not always produce output due to the content of the programme. In one notable case where a programme produced no output this was because it was a nature documentary about wildebeests, with no humans in it.

4.5 Comparison of datasets

The LRS3-TED, LRS2 and MGB datasets are introduced in Section 3.1. Here I discuss in more detail how the datasets differ in content and difficulty.

In this project I use the LRS3-TED and MGB datasets, the deliberate mismatch between the domains of the datasets motivated in Section 1.1. The datasets have a similar size, with MGB containing approximately 145,000 utterances and LRS3-TED containing around 152,000 utterances. While there are benchmarks for AVSR and lip reading on the LRS3-TED dataset, since the MGB dataset was produced in this project, there is no benchmark for performance on this dataset.

When making comparisons between results on both of these datasets, it is important to note that the MGB dataset is more challenging than the LRS3-TED dataset. It contains a variety of genres of TV broadcasts with different visual and acoustic characteristics, as noted in Section 4.4, while LRS3-TED contains only TED talks, which tend to be quite consistent in style with similar indoor backgrounds and speakers often wearing microphones on their faces. The TED talks are likely to have clean audio, lacking background noise and having only one speaker. In comparison, the MGB dataset may contain more background noise or overlapping speakers due to the nature of television programmes, particularly in the Drama genre.

The LRS2 dataset was used in last year's project. It is similar to MGB in some senses as it contains BBC TV news and talk show programmes, but MGB also contains additional

genres which make it more challenging than LRS2. Since some of the genres overlap, the domain mismatch between LRS2 and MGB is more ambiguous than that between MGB and LRS3-TED, and so clear conclusions cannot be drawn by experimenting with LRS2. I therefore use LRS3-TED and MGB in this project.

4.6 Practical issues

Since the dataset generation requires a lot of computation I parallelised the task to be more efficient, splitting up the data into the four genres and running them separately through the pipeline on different GPUs. I was slightly limited by the GPU server I was using, which had around 60GB of memory. The pipeline uses up a lot of memory, so I could not use all four GPUs at once or this would have consumed all the available memory. I therefore used 2-3 GPUs at a time. In order to do this, I changed the naming of some temporary audio and video files in the pipeline code so that the different instances of the pipeline were not all trying to read/write the same temporary files at once.

In order to generate the AV dataset, I first had to convert the original MGB Challenge dataset videos to a frame rate of 25 frames per second (fps) using `ffmpeg` before feeding them into the AVSR pipeline. Generating the full training set took approximately 631 hours, if I had been running it serially (112 hours to convert to 25 fps, 519 hours using the pipeline). However, since I ran it on 2-3 GPUs in parallel, it took less time - around 10 days.

A problem I encountered with the AVSR pipeline that remains unresolved was that very occasionally the process would hang, stuck on generating a temporary audio file for one facetrack within a video. This happened only a handful of times, and by killing that one process that face track could be skipped and the rest of the process continued to run smoothly, generating clips for the rest of the video. The cause of this problem is unknown. It did not occur when the same video file was put through the pipeline on a different server, so it may have been a memory issue specific to the server I was using.

A minor issue I noticed with the MGB Challenge data is that the programme recordings do not cut off immediately after the programmes end. There are therefore some instances where the start of the next programme, which may be classed as a different genre, is included as part of the genre of the previous programme. For example, I found a 60 Second News Blast (genre: News) at the end of an episode of Doctor Who (genre: Drama). Unfortunately there was not much that I could do about this, as it was due to the format of the BBC TV data. However, this is unlikely to cause many erroneously classified clips, since the overlapping segment at the end of the programme is very short in comparison to the length of the overall programme.

Chapter 5

Lip reading implementation

In this chapter I describe the visual features I used for lip reading, the implementation details of supervised and semi-supervised training of acoustic and visual models, the default parameters and the language models used in my experiments, and related practical issues.

5.1 Visual features

Acoustic models typically use MFCC features as input, as these have been found to be a compact and insightful representation of speech. An interesting challenge when designing a lip reading model is choosing visual features which will carry similarly useful information about the speech, extracted from each frame of the video data. As described in Section 3.2, there are various traditional methods of generating visual features to be used as input to HMM-based models, as well as end-to-end models which simply take in the original video as input and extract features within their network. I would like use more insightful modern methods than the traditional ones to extract features to feed into my HMM-DNN system.

In this project I use SyncNet (Chung and Zisserman, 2016) to generate visual features. This network was described in more detail in Section 2.4. The idea behind using these features is that since SyncNet is used to synchronise audio and video, it learns a representation of the images strongly related to lip movement of the speaker. Since lip movement is key to visual speech recognition, this representation should be more insightful compared to the plain pixels of the images, or the traditional handcrafted features.

After running the feature extraction each video clip is represented by a sequence of 512-dimensional feature vectors, one for each frame of the video. These feature vectors are suitable input to the Kaldi models.

A potential issue with using these visual features is that while they are related to lip movement, SyncNet was optimised for a different objective function, namely audio/video synchronisation. It is therefore possible that it is extracting some information relevant only to the synchronisation task and missing some information relevant to

speech recognition. More useful features may be extracted if I were to fine tune SyncNet for the task of lip reading.

In the lip reading experiment by Chung and Zisserman (2016), they used the features output by the final fully connected layer of the visual stream of the SyncNet CNN. However, I use the features output by the final convolutional layer of the CNN before the fully connected layers. The output of this layer should contain slightly more general lip features compared to the final output which was trained to be compared against the audio output for synchronisation.

5.2 Supervised training

I trained supervised acoustic and visual models using scripts initially developed by my co-supervisor Ondřej Klejch. The models were all built using Kaldi (Povey et al., 2011).

The acoustic model takes as input high-resolution 40-dimensional MFCC feature vectors. The visual model takes as input the 512-dimensional visual features described in Section 5.1. The models were trained using the LRS3-TED pretrain and training datasets. The pretrain set contains 118,516 utterances, while the training set contains 30,851 utterances from the trainval set (Afouras et al., 2018c). 1131 utterances were set aside from the trainval set as a validation set, though this was not used in this project.

The acoustic model is built using the standard Kaldi recipe for training HMM-DNN systems, as described in Section 2.1. A series of progressively better HMM-GMM models are trained and the audio and transcripts realigned at each stage, starting with a monophone model, and then several triphone models. The last triphone model is used to produce a final alignment which is used to train a time delay neural network (TDNN).

The visual model builds upon the acoustic model. The last acoustic triphone model is used to generate alignments for the visual features. A TDNN visual model is then trained using these alignments.

The frame subsampling factor is the ratio of frames per second of features trained on to model output. The alignment subsampling factor is the ratio of frames per second of input alignments to model output. The default value for both these parameters is 3. Since the MFCCs are 100 frames per second but the visual features are 25 frames per second, I used a frame subsampling factor of 4 for the acoustic model and 1 for the visual model, so that they both have the same frame rate. The alignment subsampling factor for both models is set to 4. A few other key training parameters are specified in Table 5.1.

5.3 Semi-supervised training

As detailed in Section 3.3, semi-supervised LF-MMI training involves decoding untranscribed training data to produce lattices. I performed semi-supervised decoding of the MGB training set using the supervised LRS3 acoustic and visual models. Due to the choice in subsampling factors as explained in Section 5.2, the lattices produced by either

the acoustic model or the visual model can be used for the semi-supervised training of either acoustic or visual models. I train models using three of these combinations:

- Semi-supervised acoustic models: acoustic models trained using the lattices produced by the acoustic model.
- Semi-supervised visual models: visual models trained using the lattices produced by the visual model.
- Semi-supervised acoustic-visual models: visual models trained using the lattices produced by the acoustic model.

I experimented with semi-supervised models initialised with the supervised LRS3 model and trained further on the MGB training data - I refer to these as LRS3+MGB models. I also trained models from scratch, using only the MGB training data, these are referred to as MGB models.

I created scripts for semi-supervised training of acoustic, visual and acoustic-visual models. I based my implementation on the implementation of semi-supervised LF-MMI training for acoustic models by Manohar et al. (2018)[1][2], adapting the scripts to work with the supervised LRS3 models and for training of visual models.

The general structure of the semi-supervised training script for each modality is the same. First, I extract the acoustic or visual features from the MGB training data, then decode the data using the supervised acoustic or visual model and the semi-supervised decoding language model. The resulting lattices are saved. The denominator WFST is computed. I then generate training examples using the semi-supervised lattices and the training data features. Each utterance is split up into examples containing 150 feature frames, for more efficient minibatch training on GPUs. I train a TDNN model on these examples, either starting from a pretrained input model or from a flat start. Once trained, this semi-supervised model can then be decoded on the MGB test set using the MGB language model.

Both the semi-supervised acoustic and visual models follow all these stages, with the acoustic model generating lattices using the acoustic supervised model and using those to generate training examples alongside the MFCC features, and the visual model generating lattices using the visual supervised model and using those to generate training examples with the visual features. Since the acoustic-visual model uses the lattices generated by the acoustic model, it only needs to run from the stage where the denominator WFST is produced, and generates training examples using the acoustic lattices and the visual features.

This implementation of semi-supervised training uses the naïve lattice splitting method described in Manohar et al. (2018) to create numerator graphs for the utterances, pruning the lattices with a beam of 4.0 and using an LM-scale of 0.5 to scale the weights of the lattices when generating the numerator graphs. The original script extracts i-vectors

[1] https://github.com/kaldi-asr/kaldi/blob/master/egs/fisher_english/s5/local/semisup/run_100k.sh

[2] https://github.com/kaldi-asr/kaldi/blob/master/egs/fisher_english/s5/local/semisup/chain/tuning/run_tdnn_100k_semisupervised_1a.sh

for speaker adaptation, I removed this as I am not performing speaker adaptation. They also performed speed perturbation to augment the audio data, but I also removed this as it does not make sense in the context of video data. They trained a model from scratch using both the supervised and unsupervised data, whereas I train using only the semi-supervised data either from scratch or from the initial supervised model.

For the semi-supervised acoustic-visual model trained from a flat start I used the same architecture as the supervised visual model described in Section 5.2.

I used the same frame subsampling factor for the semi-supervised models as for the supervised models: 4 for the acoustic models, 1 for the visual models, so that the frame rate for the acoustic and visual features matches. For all the semi-supervised models I set the alignment subsampling factor to 1, since they use the alignments produced by the supervised models which have already been subsampled. A few other key training parameters that differ between the acoustic and visual models are specified in Table 5.1.

5.4 Default model parameters

The default parameters I used for each of the model types is given in Table 5.1. The parameters for the supervised LRS3 models were already tuned previously.

For the semi-supervised LRS3+MGB models, since they build upon a pre-trained model, I decided to perform fewer epochs of semi-supervised training compared to the supervised training. For the semi-supervised LRS3+MGB acoustic model, I chose 4 epochs, which is 2/3 of the 6 epochs used in supervised training. This was not tuned. For the semi-supervised LRS3+MGB acoustic-visual model I tried a few different numbers of epochs - results are shown in the experiment in Section 6.3. The number of epochs which gave the best WER was 12, which also happens to be 2/3 of the supervised epochs (18), making it consistent with the acoustic model. I used the same number of epochs for the semi-supervised LRS3+MGB visual model, since it is based on the same supervised LRS3 visual model. For the semi-supervised MGB acoustic-visual model, since it is a visual model trained from scratch, I used the same number of epochs as the supervised LRS3 visual model (18).

For the semi-supervised training I used the same minibatch size as in the supervised

Table 5.1: Default parameters for training the models.

Training method	Model	# epochs	Minibatch size	Initial learning rate	Final learning rate
Supervised	LRS3 Acoustic	6	256	2.5E-04	2.5E-05
	LRS3 Visual	18	256,128,64,32	2.5E-04	2.5E-05
Semi-supervised	LRS3+MGB (A)	4	256	1.25E-04	2.5E-05
	LRS3+MGB (V)	12	256,128,64,32	1.25E-04	2.5E-05
	LRS3+MGB (AV)	12	256,128,64,32	1.25E-04	2.5E-05
	MGB	18	256,128,64,32	2.5E-04	2.5E-05

training. For visual models I used a variable minibatch size, meaning that the largest possible size is used until there is no longer enough input for a batch that large, at which point the next possible largest size is used. This is represented by a list of sizes: 256,128,64,32.

For the semi-supervised training, I chose an initial learning rate that was half of the initial learning rate of the supervised models, and the same final learning rate. I chose a smaller initial learning rate so that while the model trains on the MGB data it does not forget what it has learned from the LRS3 data. Since the semi-supervised MGB acoustic-visual model is trained from scratch, I gave it the same default learning rates as the supervised models.

5.5 Language models

The language models I used in my experiments are the following:

MGB – a tri-gram language model created using the top 150,000 most frequent words in the vocabulary of the MGB Challenge dataset transcriptions, in addition to transcripts of previous BBC TV programme data, amounting to a total of several hundred million words.

MGB_big – a less pruned version of the MGB language model.

MGB_big_del - the MGB_big language model with a deletion penalty of 1 applied.

MGB_big_del2 – the MGB_big language model with a deletion penalty of 2 applied.

LRS3 – a tri-gram language model trained on the transcriptions of the LRS3-TED training dataset.

In my semi-supervised training experiments I used the MGB, MGB_big, MGB_big_del and MGB_big_del2 language models for the semi-supervised decoding of the MGB training data. For decoding the MGB test data I always used the MGB language model. This was for consistency, so that I could more easily compare the results of semi-supervised training with different semi-supervised decoding language models.

I only used the LRS3 language model to decode the test data in the supervised training experiments, for comparison with the MGB language model.

I applied the deletion penalty by running a script (Fainberg et al., 2019) to convert the *HCLG* decoding graph for the MGB_big language model into a new graph with the penalty applied.

5.6 Practical issues

Currently, utterances longer than 637 frames (25.46 seconds) are skipped during visual feature extraction. This is because any longer clips are too long to fit in memory during the feature extraction. In order to include these clips, I would need to edit the visual feature extraction script to split up the long videos into multiple chunks to be processed

separately. I decided it was not worth changing this during the course of this project, since the number of clips skipped is not high enough to be of concern for data quantities, and additionally, such long utterances would be likely to be difficult for Kaldi to deal with.

When I extracted the visual features for the full MGB training set 3890 utterances were skipped due to being too long. This accounts for 2.8% of the total utterances. It is interesting to note upon closer inspection that the majority of these were from the News genre - with 5.4% of News clips being skipped, compared to less than 1% of clips from each other genre. As mentioned in Section 4.4, the News genre contains more face-on views of speakers and less camera movement than other genres, so it makes sense that longer facetracks can be found in this genre compared to other genres. Drama had the least skipped videos, which makes sense since the filming style of cutting frequently between different angles does not allow for long facetracks.

The visual feature extraction requires a large amount of compute power, so to make this more efficient I parallelised the task across 4 GPUs. I split the data into 4 subsets and processed each individually, after which I recombined the features into one place. It took approximately 4.25 hours to extract the visual features for the full MGB training set.

The semi-supervised decoding of the MGB training data was quite slow, so for the later experiments (the semi-supervised training using the language models with deletion penalty) I changed the lattice pruning beam value used in the decoding script from the default of 8.0 to 4.0. This heavier pruning made it run faster and produced smaller lattices therefore taking up less storage space as well. Since the lattices are pruned with a beam of 4.0 before being used for training anyway, as mentioned in Section 5.3, this change does not affect the results of training.

Once features have been extracted from the data or the lattices or training examples have been generated these can be saved for later use. This makes training many models with varying parameters much more time efficient, as the features, lattices and training examples do not need to be regenerated every time. However, the training examples are quite large so this is not always space efficient, and care must be taken that there is enough disk space to store them. For instance, the acoustic training examples for the full MGB training set take up approximately 8.5GB, the acoustic-visual examples take up approximately 7.5GB and the visual examples take up an enormous 20GB.

Chapter 6

Experiments and results

In this chapter I present my experiments on supervised and semi-supervised LF-MMI training of acoustic and lip reading models and discuss their results.

6.1 Supervised LRS3 models

The aim of this experiment is to compare how the supervised LRS3 models perform on the MGB test set compared to the LRS3 test set, and how the performance differs when decoding using the LRS3 or the MGB language model. Performance on the LRS2 test set is also given for comparison.

The results of the decoding the supervised LRS3 acoustic and visual models are found in Table 6.1.

These results clearly show that the supervised models struggle to generalise to a mismatching domain, since there is a stark difference in performance on the LRS3 test set compared to the MGB test set. This is particularly noticeable in the acoustic model, where 4.7% is achieved on the LRS3 test set, compared to the much worse 42.1% on the MGB test set. The difference is smaller but still quite large for the visual model, 70.6% on LRS3 vs 93.4% on MGB. This demonstrates the motivation to perform semi-supervised training experiments to try and improve this performance on the data from the mismatched domain.

I also decoded on the LRS2 dataset. As explained in Section 4.5, the MGB dataset is a harder dataset than LRS2 and LRS3, due to its multiple genres, and this is demonstrated by the results for the acoustic model. The results on LRS2 (12.9%) are worse than for LRS3, since it is also a mismatched domain, but these results are still much better than on MGB. Interestingly, the visual model does very well on LRS2, getting almost the same result as for LRS3 when the MGB LM is used, although it does worse when the LRS3 LM is used, which makes sense as the MGB LM and the LRS2 dataset both contain BBC data so are well matched.

When comparing the performance between genres in the MGB test set we can see that the News genre performs best. This makes sense, as mentioned in Section 4.2 the News

Table 6.1: The WER of the supervised LRS3 acoustic and visual models, decoded on the LRS3, LRS2 and MGB test sets using the LRS3 and MGB language models.

		Test language model	
Model type	**Test Data**	LRS3	MGB
Acoustic	LRS3	5.1%	4.7%
	LRS2	19.2%	12.9%
	MGB	45.9%	42.1%
	News	35.7%	31.9%
	Documentary	45.4%	42.0%
	Childrens	46.0%	40.6%
	Drama	61.6%	57.6%
Visual	LRS3	71.1%	70.6%
	LRS2	77.6%	70.2%
	MGB	93.4%	93.4%
	News	88.5%	88.7%
	Documentary	94.6%	94.7%
	Childrens	94.3%	94.2%
	Drama	97.0%	96.1%

genre is the dominant genre present in the MGB training dataset. Additionally, the News genre tends to contain more frontal views of faces, and less background noise, therefore making it one of the easier genres from which to recognise speech. The hardest genre is found to be Drama, which again is to be expected, since compared to News it contains more movement of faces and background noise. The Childrens and Documentary genres produce similar results to each other, in between News and Drama. The difference between genres is much greater in the acoustic results compared to the visual results, with a difference of approximately 26% between News and Drama, compared to a difference of approximately 8% for the visual. This indicates that the model may be struggling to deal with the more noisy audio present in the Drama genre that is not in the News genre - which does not affect the visual model since it uses visual features unaffected by background audio. This is one of the reasons why it is interesting to look into audio-visual speech recognition, as it may be able to improve recognition in cases where the acoustic models struggle.

Comparing the WER when decoding using the different language models, the MGB language model outperforms the LRS3 language model on all test data for the acoustic model. For the visual model, the MGB LM performs better or equally to the LRS3 LM for the LRS3 test data and the full MGB test set - though it performs very slightly worse on some individual genres. The improvements in WER produced by using the MGB LM compared to LRS3 LM were less than 1% for the LRS3 test set, and for the MGB data using the visual model, but using the acoustic model on the MGB data gave a much more noticeable improvement of between 3.4-5.4%.

One might expect that since the supervised model is trained on LRS3 data, an LRS3 language model would be more suitable than the MGB language model, particularly when decoding on the LRS3 test data. However, the better performance achieved using

the MGB language model may be because the MGB model was trained on a larger vocabulary than the LRS3 model. In conclusion, the MGB language model is used in future experiments to decode the test data since it gave better performance.

The main results to note from this experiment are the 93.4% on the MGB test set with the visual model and 42.1% with the acoustic model. These are the results I am aiming to improve upon using semi-supervised training.

6.2 Language models

In this experiment I decoded the MGB test data using the supervised LRS3 acoustic and visual models and a variety of language models: MGB, MGB_big, MGB_big_del and MGB_big_del2. See Section 5.5 for details of each model. These language models were later used for semi-supervised decoding, not test decoding, so the aim of this experiment was to perform a sanity check to see how the big language models and deletion penalty affect the decoding of MGB data.

The results of this experiment are found in Table 6.2.

As expected, using the MGB_big LM decreased the WER compared to the more pruned MGB LM. This was more noticeable in the acoustic model, which improved by 2.8%. In comparison, the visual model improved by only 0.7%. This improvement motivates the use of the MGB_big LM for semi-supervised decoding, to improve the lattices produced and therefore improve the final model.

The WER breakdown shows that the visual model has a very high deletion rate - similar to the number of substitutions, and almost half of the overall WER - and almost no insertions. This means almost half of each utterance is being deleted. In an ideal ASR system we expect to see the majority of errors being substitutions, and fewer deletion and insertion errors. The acoustic model has a better breakdown - the deletions are lower than the substitutions, but are still double the insertions, meaning that utterances are still being predicted to be shorter than they are.

Table 6.2: The WER breakdown into percentages of substitution, deletion and insertion errors of the supervised LRS3 acoustic and visual models, decoded on the MGB test set using varying language models.

		Test language model			
Model type	**Metric**	MGB	MGB_big	MGB_big_del	MGB_big_del2
Acoustic	WER	42.1%	39.3%	38.4%	39.4%
	Substitutions	20.0%	18.0%	18.4%	19.5%
	Deletions	15.3%	14.2%	11.8%	9.9%
	Insertions	6.8%	7.1%	8.2%	10.0%
Visual	WER	93.4%	92.7%	92.3%	98.6%
	Substitutions	44.9%	46.5%	44.9%	65.9%
	Deletions	46.7%	43.9%	45.3%	22.1%
	Insertions	1.8%	2.3%	2.2%	10.7%

When a model is prone to deletion, we expect that the semi-supervised training will increase this effect, and the final model trained on the semi-supervised lattices which are prone to deletion will be even more prone to deletion. Therefore, to help reduce the deletion error rate in the final model, it seems a good idea to introduce a deletion penalty into the language model used for semi-supervised decoding.

Prior to using the deletion penalty in the semi-supervised decoding, I decoded the supervised models with the MGB_big_del and MGB_big_del2 LMs on the test data to check what effect it had. The deletion penalty of 1 gave promising results, decreasing the deletion rate and the overall WER for the acoustic model compared to using MGB_big. The deletion penalty of 2 decreased the deletion rate even more, but due to an increase in substitution and insertion errors, the overall WER increased, compared to using MGB_big. For the visual model the results were less consistent, with the deletions actually increasing when the deletion penalty was 1.

The conclusion from this experiment is that it is worth using the larger language model in the semi-supervised training, particularly since it improves the acoustic model more, and the main experiment in this project is the semi-supervised acoustic-visual model which involves semi-supervised decoding using the acoustic model. Additionally, it is worth paying attention to how the deletion rate changes for semi-supervised models compared to the supervised models, and experimenting with the deletion penalty to mitigate any increase in deletions.

6.3 Epoch tuning

The aim of this experiment was to determine the number of epochs to use for training the semi-supervised LRS3+MGB acoustic-visual models.

This experiment was carried out using the MGB language model for semi-supervised decoding and test decoding. The default minibatch size and learning rates given in Table 5.1 were used.

The results of this experiment are found in Table 6.3.

Table 6.3: The WER of the semi-supervised LRS3+MGB acoustic-visual model trained using varying number of epochs.

# epochs	WER
4	92.2%
6	92.0%
12	91.9%

12 epochs provided the lowest WER of 91.9% and so in future experiments in training semi-supervised LRS3+MGB acoustic-visual models, this is the number of epochs used. I also used this number of epochs for the semi-supervised LRS3+MGB visual models, since they are similar since they are based upon the same initial model.

6.4 Semi-supervised LRS3+MGB models

The aim of this experiment was to improve performance on the MGB test set by training semi-supervised acoustic, visual and acoustic-visual models, using the supervised LRS3

Table 6.4: The WER of the semi-supervised LRS3+MGB acoustic, visual and acoustic-visual models decoded on the MGB test set. Models trained using the default parameters, and varying the language model used for semi-supervised decoding of the training data.

	Semi-supervised decoding language model			
Model	MGB	MGB_big	MGB_big_del	MGB_big_del2
LRS3+$\text{MGB}_{\text{short}}$ (A)	40.5%	39.8%	**39.3%**	40.2%
LRS3+MGB (A)	37.8%	37.3%	**36.8%**	37.7%
LRS3+$\text{MGB}_{\text{short}}$ (V)	**95.2%**	-	-	-
LRS3+MGB (V)	**95.0%**	-	-	-
LRS3+$\text{MGB}_{\text{short}}$ (AV)	94.0%	93.9%	93.9%	**93.4%**
LRS3+MGB (AV)	91.9%	91.8%	91.7%	**91.4%**

Table 6.5: The WER breakdown into percentages of substitution, deletion and insertion errors of the semi-supervised LRS3+MGB acoustic, visual and acoustic-visual models, decoded on the MGB test set using varying language models for semi-supervised decoding of the training data.

		Semi-supervised decoding language model			
Model	**Metric**	MGB	MGB_big	MGB_big_del	MGB_big_del2
LRS3+MGB (A)	WER	37.8%	37.3%	36.8%	37.7%
	Substitutions	16.1%	17.2%	17.1%	18.4%
	Deletions	15.7%	13.2%	11.8%	10.3%
	Insertions	6.0%	7.0%	7.9%	9.0%
LRS3+MGB (V)	WER	95.0%	-	-	-
	Substitutions	33.1%	-	-	-
	Deletions	60.9%	-	-	-
	Insertions	0.9%	-	-	-
LRS3+MGB (AV)	WER	91.9%	91.8%	91.7%	91.4%
	Substitutions	39.0%	40.4%	36.7%	42.9%
	Deletions	51.5%	49.8%	53.7%	46.5%
	Insertions	1.4%	1.6%	1.3%	1.9%

models as an initial model and training further using the MGB training set. This experiment used the default parameters specified in Table 5.1, and various different LMs were used for semi-supervised decoding.

The experiments were run using both the full and short training set, the results are given in Table 6.4. The breakdown of the WER into substitution, deletion and insertion errors for the models trained using the full training set is given in Table 6.5. The breakdown of WER for each genre for the best performing models is given in Table 6.6, and a full breakdown is given in Tables A.1 and A.2.

All the semi-supervised acoustic models improved upon the supervised acoustic model result of 42.1%. The most basic model, trained using the short training set and using the MGB LM achieved 40.5%, which is a promising start from training on just the small amount of MGB training data with the smallest language model. This was brought

Table 6.6: The WER of the best semi-supervised LRS3+MGB acoustic, visual and acoustic-visual models decoded on the MGB test set, broken down by genre. Models trained using the default parameters.

	Model		
Test data	LRS3+MGB (A)	LRS3+MGB (V)	LRS3+MGB (AV)
MGB	36.8%	95.0%	91.4%
News	27.2%	91.9%	84.5%
Documentary	37.4%	96.0%	93.6%
Childrens	34.6%	94.8%	91.0%
Drama	50.8%	97.1%	95.6%

down a few more percentage points by using the full training set, achieving 37.8%. Using the larger MGB_big LM to decode the training data took much longer compared to the MGB LM, but it was worth the extra time as it decreased the WER even more, and further to that, adding the deletion penalty of 1 gave the best result of 36.8%.

From the genre breakdown in Table 6.6 note that we still see a general genre trend as seen with the supervised models, where the News genre has the best WER while the Drama genre has the worst, and this difference is larger in the acoustic models compared to the visual or acoustic-visual models. The lowest WER achieved in the whole set of experiments was for the News genre with the semi-supervised LRS3+MGB acoustic model using the deletion penalty of 1, which achieved 27.2% (compared to 31.9% with the supervised model). The lowest WER achieved by a visual model was 84.5% on the News genre by the semi-supervised LRS3+MGB acoustic-visual model using deletion penalty of 2 (compared to 88.5% with the supervised model).

The semi-supervised LRS3+MGB visual model performed worse than the supervised visual model, with a WER of 95.0% compared to 93.4%. This is not surprising, since the initial visual model has such a high WER that the semi-supervised decoding of the MGB data does not produce useful training examples. The supervised visual model has roughly equal substitution errors and deletion errors (and very low insertion errors), and the semi-supervised training exacerbates this resulting in the highest deletion rate seen, 60.9% (almost double the substitution rate). This means that that more than half of the words in utterances are being left out. This is why I perform the acoustic-visual experiments, the idea being to use the acoustic model to decode the semi-supervised training data, rather than the visual model, since it performs much better.

I did not run the visual experiments with the MGB_big language model, because the semi-supervised decoding of the training data with the visual model would create enormous lattices that take up too much memory and time to create. Additionally, they would be unlikely to give any useful result in the end, based off the result using MGB.

The semi-supervised LRS3+MGB acoustic-visual models trained using the full MGB training set improved upon the supervised visual model by 2.0%, achieving 91.4%. Similarly to the acoustic model, the larger language models improved the results slightly compared to the MGB LM. When trained only on the short training set the models achieved worse or equal performance compared to the visual model.

Looking at the breakdown of substitutions, insertions and deletions in Table 6.5 we can see that the semi-supervised training increases the quantity of deletions. Note that similar trends were seen for the short and full training sets, so only the breakdown for the full set is included in this report. For the semi-supervised LRS3+MGB acoustic model decoded using the MGB LM, although the overall WER decreases compared to the supervised model, the number of deletions increased. This brings the number of deletions much closer to the number of substitutions, compared to the supervised model.

Looking to the semi-supervised acoustic-visual model, the deletions also increase compared to the supervised model. Notably, however, the deletions do not increase as much as they do for the semi-supervised visual model. This makes sense since the model use to decode the lattices (the supervised acoustic model) had a lower deletion rate compared to the supervised visual model, which was used to decode the lattices for the semi-supervised visual model.

Generally the deletion penalty was effective in reducing the number of deletions, with the higher penalty of 2 having a stronger effect - but due to an increase in insertion and substitution errors this did not always result in an overall decrease in WER, e.g. for the acoustic model the deletion penalty of 1 gave the best results. For the acoustic-visual model the deletion penalty of 2 was more effective in reducing the WER compared to the penalty of 1, reducing it almost exactly back to the deletion rate of the supervised model. Therefore, in the next experiment I only used the deletion penalty of 2.

In the next experiment to tune the learning rate of the semi-supervised acoustic-visual model I decided to use only the full MGB training set. The idea of using the short set was to be faster, but in the end it did not take too long to train using the full set, and the MGB short training set did not give promising results for this type of model.

6.5 Learning rate tuning

The aim of this experiment was to tune the initial and final learning rate for the semi-supervised LRS3+MGB acoustic-visual model to try to improve upon the WER achieved using the default learning rates in the previous experiment.

The results of this experiment are given in Table 6.7 and the full genre breakdown is given in Table A.3.

The best result, 91.3% was achieved using an initial learning rate of 2.5E-04 and final learning rate of 2.5E-05 and the MGB_big_del2 language model. The best learning rates were different for each semi-supervised language model. The tuning only improved the models by up to 0.2% compared to the defaults.

I did not perform any learning rate tuning for the semi-supervised LRS3+MGB acoustic models, as the focus of the project was on the visual models and there was not time to carry out comprehensive experiments in all directions, although it would have been interesting for completeness to investigate if there is any improvement to be gained in the acoustic model by changing the learning rate.

Table 6.7: The WER of the semi-supervised LRS3+MGB acoustic-visual models decoded on the MGB test set, tuning over initial and final learning rates and varying the language model used for semi-supervised decoding of the training data.

Semi-supervised decoding language model	**Initial learning rate, final learning rate**					
	2.5E-04	1.25E-04	1.25E-04	1E-04	2.5E-05	1E-05
	2.5E-05	2.5E-05	1.25E-05	1E-05	2.5E-06	1E-06
MGB	92.1%	**91.9%**	92.1%	92.0%	92.1%	92.6%
MGB_big	91.8%	91.8%	91.9%	**91.6%**	92.0%	92.2%
MGB_big_del2	**91.3%**	91.4%	91.5%	91.5%	91.7%	91.6%

6.6 Semi-supervised MGB models

The aim of this experiment was to train a semi-supervised acoustic-visual model from scratch, using only the semi-supervised MGB training data, and to compare its performance to the supervised LRS3 visual model and the semi-supervised LRS3+MGB acoustic-visual models. The default parameters specified in Table 5.1 were used, except for the learning rates which were tuned.

The results of this experiment are given in Table 6.8 and the full genre breakdown is given in Table A.4.

The best WER achieved by this model was 92.0%, using initial learning rate 5E-04, final learning rate 5E-05 and the MGB_big_del2 LM. This is an improvement compared to the supervised visual model result of 93.4%. It is not as good as the result of the semi-supervised LRS3+MGB acoustic-visual model, which achieved 91.3%. However the result is still promising, since it uses no supervised data whatsoever and still gets very close to the results of the LRS3+MGB models which are additionally pretrained on the whole LRS3 training data.

In this experiment, using the deletion penalty of 2 improved the results, but only by up to 0.3%.

Note that in this experiment I only used the full training data set. I decided that for a model being trained from scratch the short dataset (containing 27 hours) was too small to be useful.

Table 6.8: The WER of the semi-supervised MGB acoustic-visual models decoded on the MGB test set, tuning over initial and final learning rates and varying the language model used for semi-supervised decoding of the training data.

Semi-supervised decoding language model	**Initial learning rate, final learning rate**				
	1E-03	5E-04	2.5E-04	1E-04	1E-05
	1E-04	5E-05	2.5E-05	1E-05	1E-06
MGB	92.7%	92.6%	**92.3%**	93.0%	94.8%
MGB_big	92.4%	92.3%	**92.1%**	92.6%	94.5%
MGB_big_del2	92.6%	**92.0%**	92.1%	92.3%	94.2%

For the future, it would be interesting to experiment with different model architectures which use more layers, as this may improve the performance. It would also be interesting to train a semi-supervised MGB acoustic model from scratch, to compare how it performs against the semi-supervised LRS3+MGB acoustic model.

Chapter 7

Conclusions

7.1 Summary of results

I produced a 223 hour MGB training dataset (and a 27 hour short subset) from the MGB Challenge dataset, containing video clips from the News, Documentary, Childrens and Drama genres, and used this in my semi-supervised training experiments.

The summary of the best WER achieved for each type of model on the MGB test set after experimenting with various language models and parameters is given in Table 7.1.

Comparing the performance of supervised LRS3 models on the LRS3-TED dataset and the MGB dataset, we can see that there is a much higher WER on the MGB dataset (4.7% -> 42.1% for acoustic, 70.6% -> 93.4% for visual) - demonstrating that the models struggle to generalise to the mismatching domain. Despite this poor performance on MGB data from the initial model, the semi-supervised training methods were successful in improving performance on this challenging dataset.

The acoustic semi-supervised training showed good signs of improvement, with the WER decreasing by 5.3% to 36.8%. The visual semi-supervised training showed worsening WER (95.0%), while the acoustic-visual training improved to 91.3%, demonstrating the benefits of using information from both the acoustic and visual domains. Although the improvement of the semi-supervised LRS3+MGB acoustic-visual model was quite small (2.1%), the fact it improved at all is promising for the concept of using semi-supervised LF-MMI training for lip reading models. The semi-supervised MGB

Table 7.1: The overall best WER on the MGB test dataset for each modality and type of model training.

Supervised	**Acoustic**	**Visual**	
LRS3	42.1%	93.4%	
Semi-supervised	**A**	**V**	**AV**
LRS3+MGB$_{\text{short}}$	39.3%	95.2%	93.4%
LRS3+MGB	36.8%	95.0%	91.3%
MGB	-	-	92.0%

acoustic-visual model trained from a flat start gave a similar result (92.0%), but the pretrained LRS3+MGB model gave the best result.

Using larger language models and incorporating a deletion penalty during semi-supervised decoding was found to be beneficial. However, it is to be noted that the visual models still have a much higher deletion rate than would be expected in an ideal ASR system.

7.2 Future work

Since the semi-supervised pipeline has been shown to improve lip reading performance but not as much as hoped, it is worth investigating ways to improve these models. A key improvement would be to improve the utility of the visual features used. As described in Section 5.1 the SyncNet feature extraction is tuned for the task of synchronisation. It would be therefore be beneficial to fine tune the output of SyncNet to the task of lip reading.

Since the semi-supervised acoustic model performs much better than the supervised LRS3 acoustic model, in future I would experiment with using the semi-supervised acoustic model to decode the training data again. This should produce better training examples, which could be used to train a better semi-supervised acoustic-visual model. This would be a form of incremental semi-supervised training, which has been found to be beneficial in ASR (Khonglah et al., 2020). It may be useful to train semi-supervised MGB acoustic models from scratch for this, so far I only trained semi-supervised LRS3+MGB acoustic models.

It is also worth investigating different model architectures for the visual models. Currently, the acoustic models have more layers than the visual models, so adding more layers to the visual models may provide some improvement.

It would also be interesting to try and combine the acoustic and visual features into one model to perform AVSR. Since the acoustic models perform much better than the visual models, this is very likely to improve the visual models. However, much more intriguing is the possibility of using the visual model to improve the acoustic model in some select circumstances where the audio is noisy. Combining these two streams in Kaldi however may be challenging.

Finally, a few smaller items of further work mentioned previously in this report are:

- Modifying the AVSR pipeline to save the SyncNet visual features during dataset generation
- Generating a new MGB test dataset with the latest version of the AVSR pipeline, and balancing the genre split in the dataset
- Creating an MGB validation set
- Generating transcriptions for the MGB training dataset to be used for comparison to the semi-supervised experiments

Bibliography

A H Abdelaziz et al. NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In *Interspeech*, pages 3752–3756, 2017.

T Afouras, J S Chung, A Senior, O Vinyals, and A Zisserman. Deep Audio-Visual Speech Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018a.

T Afouras, J S Chung, and A Zisserman. Deep Lip Reading: a comparison of models and an online application. In *INTERSPEECH*, 2018b.

T Afouras, J S Chung, and A Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018c.

T Afouras, J S Chung, and A Zisserman. ASR is All You Need: Cross-Modal Distillation for Lip Reading. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147, 2020. doi: 10.1109/ICASSP40776.2020.9054253.

I Anina, Z Zhou, G Zhao, and M Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–5. IEEE, 2015.

Y M Assael, B Shillingford, S Whiteson, and N De Freitas. LipNet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

L Bahl, P Brown, P de Souza, and R Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52, 1986. doi: 10.1109/ICASSP.1986.1169179.

P Bell, M J F Gales, T Hain, J Kilgour, P Lanchantin, X Liu, A McParland, S Renals, O Saz, M Wester, and P C Woodland. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693, 2015. doi: 10.1109/ASRU.2015.7404863.

A G Chitu and L JM Rothkrantz. Building a data corpus for audio-visual speech recognition. *Proceedings of Euromedia 2007*, pages 88–92, 2007.

J S Chung and A Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

J S Chung, A Senior, O Vinyals, and A Zisserman. Lip Reading Sentences in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

M Cooke, J Barker, S Cunningham, and X Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120 5 Pt 1:2421–4, 2006.

T F Cootes, G J Edwards, and C J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

A Ephrat, I Mosseri, O Lang, T Dekel, K Wilson, A Hassidim, W T Freeman, and M Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. *ACM Trans. Graph.*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201357.

J Fainberg, O Klejch, S Renals, and P Bell. Lattice-Based Lightly-Supervised Acoustic Model Training. In *Proc. Interspeech 2019*, pages 1596–1600, 2019. doi: 10.21437/Interspeech.2019-2533.

M JF Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

A Graves, S Fernández, F Gomez, and J Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

B Khonglah, S Madikeri, S Dey, H Bourlard, P Motlicek, and J Billa. Incremental semi-supervised learning for multi-genre speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7419–7423, 2020. doi: 10.1109/ICASSP40776.2020.9054309.

L Lamel, J Gauvain, and G Adda. Unsupervised acoustic model training. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–877. IEEE, 2002.

C Lüscher, E Beck, K Irie, M Kitza, W Michel, A Zeyer, R Schlüter, and H Ney. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation. In *INTERSPEECH*, 2019.

V Manohar, H Hadian, D Povey, and S Khudanpur. Semi-Supervised Training of Acoustic Models Using Lattice-Free MMI. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4844–4848, 2018. doi: 10.1109/ICASSP.2018.8462331.

S Marshall. Developing tools for audio-visual speech recognition. MInf1 Thesis, The University of Edinburgh, https://soniammarshall.github.io/MInf1_report.pdf, 2021.

M Mohri, F Pereira, and M Riley. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer, 2008.

K Noda, Y Yamaguchi, K Nakadai, H G Okuno, and T Ogata. Lipreading using convolutional neural network. In *fifteenth annual conference of the international speech communication association*, 2014.

A S Paul et al. Developing tools for audio-visual speech recognition. BSc Thesis, The University of Edinburgh, 2021.

V Peddinti, D Povey, and S Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*, 2015.

V Peddinti, Y Wang, D Povey, and S Khudanpur. Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, 25(3):373–377, 2017.

E D Petajan. *Automatic Lipreading to Enhance Speech Recognition (Speech Reading)*. PhD thesis, University of Illinois at Urbana-Champaign, USA, 1984.

G Potamianos, H P Graf, and E Cosatto. An image transform approach for HMM based automatic lipreading. In *Proceedings 1998 International Conference on Image Processing.*, pages 173–177 vol.3, 1998. doi: 10.1109/ICIP.1998.999008.

D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, et al. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

D Povey, V Peddinti, D Galvez, P Ghahremani, V Manohar, X Na, Y Wang, and S Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Interspeech*, pages 2751–2755, 2016.

D Povey, G Cheng, Y Wang, K Li, H Xu, M Yarmohammadi, and S Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747, 2018.

C Sheng, M Pietikäinen, Q Tian, and L Liu. Cross-modal Self-Supervised Learning for Lip Reading: When Contrastive Learning meets Adversarial Training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2456–2464, 2021.

B Shillingford, Y Assael, M W Hoffman, T Paine, C Hughes, U Prabhu, H Liao, H Sak, K Rao, L Bennett, et al. Large-scale visual speech recognition. *arXiv preprint arXiv:1807.05162*, 2018.

P Somervuo. Experiments with linear and nonlinear feature transformations in HMM based phone recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, volume 1, pages I–52. IEEE, 2003.

M Sperber, G Neubig, J Niehues, S Nakamura, and A Waibel. Transcribing against time. *Speech communication*, 93:20–30, 2017.

R Su, X Liu, and L Wang. Semi-supervised Cross-domain Visual Feature Learning for Audio-Visual Broadcast Speech Transcription. In *INTERSPEECH*, pages 3509–3513, 2018.

K Thangthai, R W Harvey, S J Cox, and B Theobald. Improving lip-reading performance for robust audiovisual speech recognition using DNNs. In *AVSP*, pages 127–131, 2015.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

A Waibel, T Hanazawa, G Hinton, K Shikano, and K J Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989. doi: 10.1109/29.21701.

E Wallington, B Kershenbaum, P Bell, and O Klejch. On the learning dynamics of semi-supervised training for ASR. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, pages 716–720. International Speech Communication Association, 2021.

F Weninger, F Mana, R Gemello, J Andrés-Ferrer, and P Zhan. Semi-supervised learning with data augmentation for end-to-end ASR. *arXiv preprint arXiv:2007.13876*, 2020.

H Xu, H Su, C Ni, X Xiao, H Huang, E S Chng, and H Li. Semi-Supervised and Cross-Lingual Knowledge Transfer Learnings for DNN Hybrid Acoustic Models Under Low-Resource Conditions. In *INTERSPEECH*, pages 1315–1319, 2016.

J Yu, S Zhang, J Wu, S Ghorbani, B Wu, S Kang, S Liu, X Liu, H Meng, and D Yu. Audio-visual recognition of overlapped speech for the LRS2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020.

K Yu, M JF Gales, and P C Woodland. Unsupervised training with directed manual transcription for recognising Mandarin broadcast audio. In *Interspeech*, pages 1709–1712. Citeseer, 2007.

Appendix A

WER broken down by genre

Table A.1: The WER of the semi-supervised LRS3+MGB acoustic and acoustic-visual models decoded on the MGB test set, broken down by genre. Models trained using the default parameters, and varying the language model used for semi-supervised decoding of the training data.

		Semi-supervised decoding language model			
Model	**Test data**	MGB	MGB_big	MGB_big_del	MGB_big_del2
LRS3+MGB$_{\text{short}}$ (A)	MGB	40.5%	39.8%	**39.3%**	40.2%
	News	29.7%	28.9%	28.9%	29.3%
	Documentary	41.5%	40.8%	39.9%	41.0%
	Childrens	37.2%	37.5%	37.8%	39.1%
	Drama	55.0%	54.8%	53.8%	54.1%
LRS3+MGB (A)	MGB	37.8%	37.3%	**36.8%**	37.7%
	News	27.8%	27.5%	27.2%	27.7%
	Documentary	38.9%	38.0%	37.4%	38.4%
	Childrens	35.0%	34.9%	34.6%	35.5%
	Drama	51.3%	50.8%	50.8%	51.6%
LRS3+MGB$_{\text{short}}$ (AV)	MGB	94.0%	93.9%	93.9%	**93.4%**
	News	88.5%	88.8%	88.4%	88.2%
	Documentary	95.6%	95.5%	95.3%	95.1%
	Childrens	94.6%	94.5%	94.4%	93.9%
	Drama	97.3%	97.1%	97.1%	96.6%
LRS3+MGB (AV)	MGB	91.9%	91.8%	91.7%	**91.4%**
	News	85.1%	85.2%	85.3%	84.5%
	Documentary	94.1%	93.8%	93.8%	93.6%
	Childrens	91.9%	91.9%	91.3%	91.0%
	Drama	96.3%	95.9%	95.7%	95.6%

Table A.2: The WER of the semi-supervised LRS3+MGB visual models decoded on the MGB test set, broken down by genre. Models trained using the default parameters, and using the MGB language model for semi-supervised decoding of the training data.

	Model	
Test data	LRS3+MGB$_{\text{short}}$ (V)	LRS3+MGB (V)
MGB	95.2%	95.0%
News	92.3%	91.9%
Documentary	96.2%	96.0%
Childrens	95.2%	94.8%
Drama	97.1%	97.1%

Table A.3: The WER of the semi-supervised LRS3+MGB acoustic-visual models decoded on the MGB test set, broken down by genre, tuning over initial and final learning rates and varying the language model used for semi-supervised decoding of the training data.

Semi-supervised decoding language model	**Test data**	**Initial learning rate, final learning rate**					
		2.5E-04	1.25E-04	1.25E-04	1E-04	2.5E-05	1E-05
		2.5E-05	2.5E-05	1.25E-05	1E-05	2.5E-06	1E-06
MGB	MGB	92.1%	**91.9%**	92.1%	92.0%	92.1%	92.6%
	News	85.4%	85.1%	86.0%	85.4%	85.3%	86.5%
	Documentary	94.2%	94.1%	93.6%	94.0%	94.5%	94.6%
	Childrens	92.3%	91.9%	92.5%	92.2%	92.2%	93.1%
	Drama	96.2%	96.3%	96.5%	96.5%	96.1%	96.5%
MGB_big	MGB	91.8%	91.8%	91.9%	**91.6%**	92.0%	92.2%
	News	85.2%	85.2%	85.2%	84.9%	85.3%	85.6%
	Documentary	93.8%	93.8%	93.8%	93.7%	94.2%	94.2%
	Childrens	92.0%	91.9%	92.5%	91.7%	92.3%	92.7%
	Drama	96.3%	95.9%	96.2%	96.2%	96.0%	96.3%
MGB_big _del2	MGB	**91.3%**	91.4%	91.5%	91.5%	91.7%	91.6%
	News	84.6%	84.5%	84.6%	85.0%	85.4%	85.3%
	Documentary	93.3%	93.6%	93.5%	93.4%	93.5%	93.3%
	Childrens	91.5%	91.0%	91.3%	91.1%	92.0%	91.9%
	Drama	95.7%	95.6%	95.6%	95.9%	96.2%	96.0%

Table A.4: The WER of the semi-supervised MGB acoustic-visual models decoded on the MGB test set, broken down by genre, tuning over initial and final learning rates and varying the language model used for semi-supervised decoding of the training data.

Semi-supervised decoding language model	**Test data**	**Initial learning rate, final learning rate**				
		1E-03	5E-04	2.5E-04	1E-04	1E-05
		1E-04	5E-05	2.5E-05	1E-05	1E-06
MGB	MGB	92.7%	92.6%	**92.3%**	93.0%	94.8%
	News	86.9%	86.5%	86.1%	87.1%	91.0%
	Documentary	94.4%	94.6%	94.2%	94.7%	95.9%
	Childrens	92.8%	92.6%	92.4%	93.6%	95.0%
	Drama	97.0%	96.4%	96.7%	96.8%	97.6%
MGB_big	MGB	92.4%	92.3%	**92.1%**	92.6%	94.5%
	News	86.6%	86.2%	85.5%	86.7%	90.4%
	Documentary	94.2%	94.3%	94.3%	94.7%	95.6%
	Childrens	92.7%	92.0%	92.1%	92.4%	94.6%
	Drama	96.5%	96.5%	96.4%	96.5%	97.5%
MGB_big_del2	MGB	92.6%	**92.0%**	92.1%	92.3%	94.2%
	News	86.5%	85.5%	85.7%	86.3%	89.9%
	Documentary	94.3%	93.8%	94.0%	94.1%	95.2%
	Childrens	92.4%	92.2%	92.4%	92.9%	94.5%
	Drama	96.4%	95.9%	96.0%	96.2%	97.3%