END TO END DEEP LIP-READING: A PRELIMINARY STUDY

KAMAL THAPA

https://orcid.org/0000-0002-1308-7163

A thesis submitted in partial fulfilment of the requirements of London South Bank University for the degree of Masters by Research

August 2022

Table of Contents

List of Abbreviations

ALR - Automatic Lip-reading

ASR - Automatic Speech Recognition

AVSR - Audio Visual Speech Recognition

BGRU - Bidirectional Gated Recurrent Unit

BLSTM - Bidirectional Long Short Term Memory

CNN - Convolutional Neural Networks

CTC - Connectionist Temporal Connection

DCT - Direct Cosine Transform

E2E - End to End

FCN - Fully Connected Network

GRU - Gated Recurrent Unit

HMM - Hidden Markov Model

HOG - Histogram of Oriented Graphics

LM - Language Model

LRS - Lip-reading Sentences (Dataset)

LRW - Lip-reading in the Wild (Dataset)

LSTM - Long Short Term Memory

ML - Machine Learning

NLP - Natural Language Processing

RBM - Restricted Boltzmann Machine

RNN - Recurrent Neural Network

ROI - Region of Interest

SAR - Sentence Accuracy Rate

SVM - Support Vector Machine

STCNN - Spatio-Temporal Convolutional Neural Network

TCN - Temporal Convolutional Network

VSR - Visual Speech Recognition

WAR - Word Accuracy Rate

WER - Word Error Rate

List of Figures

List of Tables

Acknowledgement

I am deeply grateful to my research supervisor Dr. Daqing Chen for the constant guidance and advice at every stage of this research project. I would like to extend my sincere thanks to London South Bank University, School of Engineering for funding the research degree. My sincere gratitude also goes to lip-reading researchers Dr. Souheil Fenghour and Randa El-Bialy for their insightful comments, suggestions and interesting discussions. A very special thanks to my wife Luisa for her unwavering support and belief in me throughout this project.

Thank you.

Abstract

Deep lip-reading is the combination of the domains of computer vision and natural language processing. It uses deep neural networks to extract speech from silent videos. Most works in lip-reading use a multi staged training approach due to the complex nature of the task. A single stage, end-to-end, unified training approach, which is an ideal of machine learning, is also the goal in lip-reading. However, pure end-to-end systems have not yet been able to perform as good as non-end-to-end systems. Some exceptions to this are the very recent Temporal Convolutional Network (TCN) based architectures. This work lays out preliminary study of deep lip-reading, with a special focus on various end-to-end approaches. The research aims to test whether a purely end-to-end approach is justifiable for a task as complex as deep lip-reading. To achieve this, the meaning of pure end-to-end is first defined and several lip-reading systems that follow the definition are analysed. The system that most closely matches the definition is then adapted for pure end-to-end experiments. Four main contributions have been made: i) An analysis of 9 different end-to-end deep lip-reading systems, ii) Creation and public release of a pipeline[1] to adapt sentence level Lipreading Sentences in the Wild 3 (LRS3) dataset into word level, iii) Pure end-to-end training of a TCN based network and evaluation on LRS3 word-level dataset as a proof of concept, iv) a public online portal[2] to analyse visemes and experiment live end-to-end lip-reading inference. The study is able to verify that pure end-to-end is a sensible approach and an achievable goal for deep machine lip-reading.

[1] https://github.com/thpkml/lrs3_word
[2] https://lsbu-analytics.org/deeplip/playground

1. Introduction

With the advancement in deep neural network technology and increasing compute power, automatic lip-reading (ALR) is evolving from its traditional multi staged process towards a more end-to-end (E2E) approach as shown in Figure 1. While the traditional approach requires a much more involved feature extraction and processing stage, modern E2E approaches make use of neural networks that automatically extract the relevant features. Also, in the non-E2E systems, the features are passed through Support Vector Machines (SVM) for classification or a sequence of features are passed through a Hidden Markov Model (HMM); while E2E systems use deep neural network based backend to process the extracted features. The thesis presents a preliminary study of ALR revolving around the E2E approach. This chapter summarises the aims and objectives of this study, the research questions, methodology, contributions and findings. The meaning of E2E lip-reading is defined in detail in section 2.3.

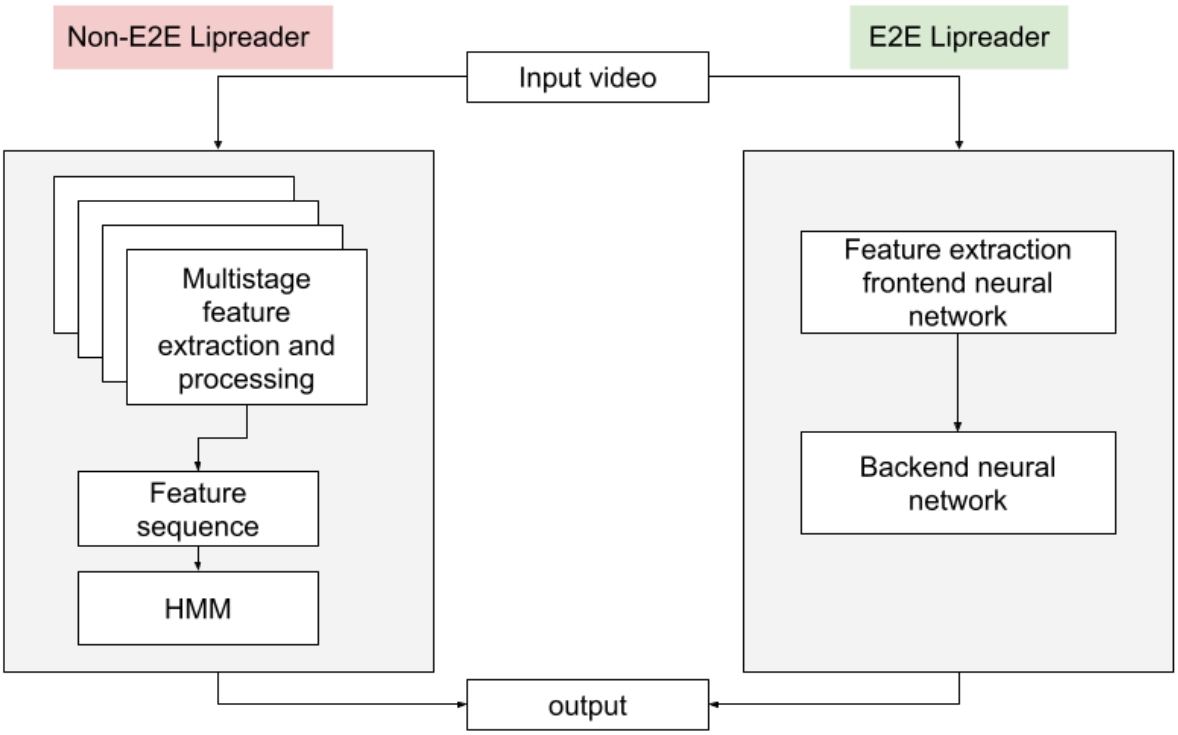


Figure 1. A typical non E2E lip-reading system using a Hidden Markov Model (HMM) (left) vs an E2E system (right)

1.1. Aims and objectives

The study aims to investigate the extent to which E2E deep learning paradigm is applicable to deep lip-reading. To this end, it has the following objectives:

i) Explore the concept of E2E in general and define its meaning in the context of deep learning and deep lip-reading.

ii) Set the deep lip-reading lay-of-the-land beginning with its meaning, applications and challenges.

iii) Survey methods, tools and techniques that have helped advance deep lip-reading from its traditional multistage roots to the modern E2E process.

iv) Evaluate various E2E works in deep lip-reading based on the criteria set for E2E.

v) Conduct experiments on a selected 'E2E' method and analyse the performance and generalisability when trained fully E2E.

1.2. Research questions

The research attempts to answer the following questions:

i) What is the ideal definition of E2E in deep lip-reading?

ii) How are the breakthrough methods, tools and techniques in deep learning, especially in speech and image processing, aiding the E2E approach?

iii) Do the seemingly E2E state-of-the-art works in lip-reading meet the full extent of E2E as defined in this study?

iv) Is the quest for pure E2E deep lip-reading pragmatic or idealistic?

1.3. Methodology

After an analysis and comparative evaluation of 9 different E2E systems with state-of-the-art in deep lip-reading, one of the systems: a TCN based E2E (Martinez *et al*., 2020) is selected for further experimentation. Details of why this specific system was chosen is presented in section 3.6.10. In order to test whether a pure end-to-end training of the whole deep lip-reading system is justifiable, the following tasks and experiments are performed:

i) Retraining the model from the scratch, as opposed to the original work which includes an additional pretraining phase, on the original dataset Lipreading in the Wild (LRW) (Chung and Zisserman, 2016). Observation of its training and test performance on LRW.

ii) Creation of a completely new word-level dataset from a sentence-level LRS3 dataset (Afouras et al, 2018).

iii) Evaluation of original pretrained model on the new more challenging LRS3 dataset to observe how well the E2E model generalises to newer data.

iv) Evaluation of the newly E2E trained model on the new LRS3 word-level dataset.

1.4. Contributions and questions answered

The study has been able to answer the research questions (1.2) at varying levels of success. Additionally, resources and sub-projects created during the project have been made publicly available, when deemed useful, as a part of the contributions as listed:

i) The meaning of E2E lip-reading has been formulated and the pros and cons of this approach are evaluated.

ii) Deep learning methods, tools and techniques that are helpful for the E2E approach have been identified and discussed.

iii) It has been discovered that most alleged E2E lip-reading systems still include a multistage/module process, thus not completely matching the full meaning of E2E. It has also been found in most cases that although the systems could be trained purely E2E, they were not, due to performance lag. However, newer TCN based models have the ability to become purely E2E while still able to set new performance records.

iv) It has been concluded that the quest for fully E2E lip-reading systems is sensible and promising.

Other contributions:

These are contributions not strictly focussed on E2E but can be of good value to lip-reading researchers. The dataset conversion pipeline (section 4.3, 4.4, Figure 27) can be used to create a novel word-level dataset to test the generalizability of E2E lipreading systems trained on other word-level datasets. The online tool can be used for a deeper understanding of visemes and phonemes before using them as classes in lip-reading systems.

v) The pipeline designed for LRS3-sentence to LRS3-word has been made available for public use[3].

vi) A web based tool is created and made publicly available with following functionalities:

   a) A t-SNE (t-distributed Stochastic Neighbour Embedding) based interactive viseme plot to analyse word-similarity clusters in English language.

   b) Text-phoneme-viseme inter-conversion tool

[3] https://github.com/thpkml/lrs3_word
The dataset can be acquired with permission from the owners

c) Live sentence-level lip-reading using user's webcam video

1.5. Organisation of the thesis

The rest of the paper is organised as follows:

- Chapter 2:

  A detailed introduction to ALR is provided including the process and applications. Challenges unique to ALR are laid out and potential solutions discussed. The meaning of 'end-to-end' (E2E) is formalised in the context of lip-reading and the limitations of E2E learning are analysed.

- Chapter 3:

  The chapter begins with a general review of methods, tools and techniques applied in ALR. Each topic is reviewed in the context of E2E lip-reading. A brief review of non E2E approaches to lip-reading is provided. Finally, as the main content of the chapter, 9 different E2E lip-reading systems are thoroughly evaluated.

- Chapter 4:

  The chapter lays out the details of the experiments performed for this research and the analysis of the results.

- Chapters 5, 6:

  Conclude the research and lay out topics for continuation of the research in the future.

2. Automatic lip-reading and the end-to-end approach

ALR is the training and use of computer algorithms to lipread. This chapter provides a detailed insight on the topic starting gradually from the general meaning and process of lip-reading, machine lip-reading and its applications. The E2E approach is then defined both in a general context and in ALR. Lip-reading faces a set of challenges unique to the field. The subsequent sections discuss these challenges, possible solutions and how the advantages and limitations of the E2E approach relate to these challenges.

2.1. Lip-reading: what, why and how

2.1.1. What is lip-reading?

There is more to understanding speech than simply speaking and hearing. Speech information is multimodal. The obvious and the dominant audio, and the often underappreciated visual cues betrayed by lip movement, tongue and teeth. The importance of visual information in speech intelligibility, especially in the presence of audio chatter, was demonstrated by Sumby and Pollack (1954). Speech perception using just the visual information is referred to as lip-reading or Visual Speech Recognition (VSR). In human communication where speakers are visible, the visual information, especially the lip movements, can influence the recognition of speech. A demonstration of this is the following example from a 1976 experiment (McGurk and MacDonald, 1976):

Lip movements in the video ⇒ /ga/

Overlaid audio ⇒ /ba/

Listeners report hearing ⇒ /da/

In the actual experiment, when the lips were saying ‘gaga’ and the voice was saying ‘baba’ the viewer would report hearing ‘dada’; a sound non-existent in the utterance. This phenomenon, now known as ‘McGurk Effect’, proves the importance of lip-reading in speech recognition.

Machine Lip-reading

Machine lip-reading, also known as ALR involves training an algorithm to read lips. Most traditional approaches to machine lip-reading had two clear stages: extraction of visual

features from a silent talking video and processing the sequence of features back into the uttered word/sentence. Features used to be generally handcrafted and purely optical. Statistical models like HMMs (Levinson *et al.*, 1983) were the predominant choice for sequence processing. A seminal work by Goldschen *et al.* (1997) captured geometric features of the oral cavity which were then processed and clustered to generate 'codewords'. Sequences of 'codewords' were then fed into a HMM, the first use of HMM in VSR, to model the temporal dynamics. Movellan (1994) opted for richer representation than handcrafted lip metrics and showed improved performance by modelling images as Gaussian mixtures, although the model also used HMM and was trained to classify only four numeric words 'one' - 'four'. Other common feature extractors for the days of yore include Principal Component Analysis (PCA) (Basu *et al.*, 1999) and Active Appearance Model (AAM) (Matthews *et al.*, 2002) of grey-scaled image frames.

Deep Lip-reading

The progress in deep neural network technology: Restricted Boltzmann Machines (RBMs) (Lee *et al.*, 2007), Convolutional Neural Networks (CNNs) (Krizhevsky *et al.*, 2017) and Recurrent Neural Networks (RNNs) (Hochreiter and Schmidhuber, 1997; Cho *et al.*, 2014) has also been reflected in machine lip-reading. The aptly named 'deep lip-reading' systems started utilising neural networks like RBMs (Ngiam *et al.*, 2011) and CNNs for feature extraction and RNNs like Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho *et al.*, 2014) for sequence processing.

E2E Deep Lip-reading

Earlier deep lip-reading still decomposed the task into two stages: learning the visual features and processing the sequence of features to predict units of speech. Currently, there's been a move towards E2E deep lip-reading where a network is able to train in a single cycle. E2E technique, being the focus of this survey, will be introduced in detail in the section 2.2. and discussed throughout this dissertation.

2.1.2. Why is lip-reading done ? (the applications)

Besides the obvious applications of ALR such as: an aid to increase noise tolerance in audio speech recognition, visual aid to speech perception in face to face communication etc., researchers in the domain often come up with other creative applications: They include:

- Car-phone dialling, considering the engine and road noise (Movellan, 1994)
- A more secure, friendlier, unobtrusive biometric measurement vs retina scan, fingerprints etc (Messer *et al.,* 1999)
- A visual-enhanced computer voice recogniser (Matthews *et al.,* 2002)
- A multimodal authentication tool on top of face and voice (Palanivel and Yegnanarayana, 2008)
- A forensic tool in counter-terrorism and law-enforcement (Theobald *et al.,* 2006)
- Creating speech-driven facial animations (Vougioukas *et al.*, 2018)
- An aide to sign language robotics (Gholipur *et al.*, 2021)
- We suggest a Google Glass like device that does live lip-reading and annotates on the glass screen.

All of these applications demand the ALR state-of-the-art to improve significantly in accuracy as well as computation time for it to be practical. One recent example (Ma *et al.,* 2021) towards this effort is the use of efficient Depthwise-Separable Temporal Convolution (DS-TCN) based models with Knowledge Distillation that runs 8.2 × faster while matching the accuracy of the state-of-the-art in word based prediction.

2.1.3. How is lip-reading done? (the process)

As it is somewhat evident by now, automatic lip-reading involves the following two main stages (Figure 2), regardless of whether the system is trained in multiple phases or E2E. The input video is first passed through a feature extractor which greatly reduces its dimensionality. The extracted features are then used to predict the uttered text e.g. word video to a word. If the ALR system classifies a sequence of extracted features into smaller speech units like phonemes, visemes or characters, such units then need to be combined and aligned properly to reproduce the uttered speech text e.g. sentence video to characters to a sentence.

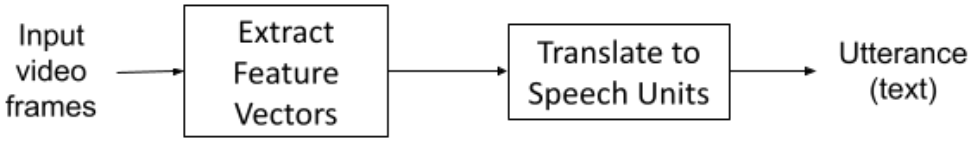


Figure 2. The 2 steps of lip-reading

The following is a list of methodologies for the two stages in chronological order, found in the literature. A detailed discussion on various ALR methods and techniques is covered in later section [Methodologies]:

Table 1: *Various approaches for feature extraction and sequence processing in lip-reading.*

Task	Approach	Year	References
Feature extraction (frontend)	Dynamic time-warping of visual features extracted from the oral cavity	1984	Petajan
	Mapping power-spectra from static images	1989	Yuhas *et al.*
	Optical flow	1989	Mase and Pentland
	Discrete Cosine Transform (DCT) on face Region of Interest (ROI)	2001	Potamianos *et al.*
	Eigenlips	1994, 2016	Bregler and Konig, Wand *et al.*
	Histogram of Oriented Graphics (HOG)	2005, 2016	Dalal and Triggs, Wand *et al.*
	Deep Bottleneck Features (DBF)	2016	Petridis and Pantic
	Fully Connected Networks (FCN)	2017	Petridis et al.
	CNN	2017	Stafylakis and Tzimiropoulos
Sequence Processing (backend)	HMM	1997, 2001	Goldschen *et al.*, Potamianos et al.
	Support Vector Machine (SVM)	2016	Wand *et al.*
	LSTM	2016	Petridis and Pantic

	Temporal CNN	2020	Martinez *et al.*
	Transformer	2018	Afouras *et al.*

2.2. Frontend-backend combinations

Various deep neural network based frontend+backend combinations have been tried for lip-reading in recent years. The choice of frontend and backend networks can determine whether the whole network can be trained E2E. Table 2 lists some of the outstanding works. Most combinations allow for an E2E training but some implement a pre-training stage for faster convergence. A detailed review of the architectures for their E2E nature is provided in section 3.6.

Table 2: *Common DNN-based combinations for feature extractor frontend and classifier backend for lip-reading. A detailed breakdown of DNN and non-DNN combinations can be found in Fenghour et al. (2021)*'s survey.

Frontend	**Backend**	**Examples**
CNN	CNN	(Chung and Zisserman, 2016)
CNN	Long Short Term Memory (LSTM)+ Attention	(Chung and Zisserman, 2016; Lu and Li, 2019)
Autoencoder	LTSM, Bidirectional LSTM (BLSTM)	(Petridis and Pantic, 2017; Petridis *et al.,* 2018)
3-Dimensional CNN (3DCNN) + ResNet	BLSTM, Bidirectional Gated Recurrent Unit (BGRU), Transformer	(Afouras et al, 2018; Stafylakis and Tzimiropoulos, 2017)
3DCNN + ResNet	TCN	(Martinez *et al.*, 2020; Ma *et al.,* 2021)

2.3. End-to-end lip-reading

This section discusses the following:

- The general meaning of 'end-to-end'

- The meaning of ‘end-to-end’ in machine learning
- Defining the meaning of ‘end-to-end’ deep lip-reading

In general, ‘end-to-end’ refers to the way any process is carried out. According to the Cambridge online dictionary, depending on the context it could mean:

- ‘From the very beginning of the process to the very end’
- ‘Including all the stages of a process’

In machine learning however, ‘end-to-end’ has a relatively more specific meaning. It generally refers to a complex learning system with a single model, where there are no intermediate stages. For some, a Machine Learning (ML) system that claims to be ‘end-to-end’ must have these features (Glasmachers, 2017):

- All modules should be differentiable (capable to learn e.g. through gradient descent).
- It should follow an unified training scheme.

There has been a significant push towards E2E ML systems in recent years. Two main drivers of this trend are: the availability of vast quantities of training data in a plethora of domains and the breakthroughs in sophisticated, deep neural network architectures (CNN, LSTM, ResNet, Transformer) able to learn complex E2E relationships using these datasets. Some of the primary motivations behind the E2E ‘revolution’ are:

- An attempt to move from the traditional multistage ML processes towards single Neural Network based processes.
- A quest to let the data speak for itself (an algorithm with very little human bias that can better expose the latent truth in the data).
- Getting the human out of the loop (the learning process).

While unifying a complex multistage ML process enables Neural Networks to learn a more complex mapping, it comes at the cost of explainability of the results. The deeper the network and the purer the E2E process, the darker its ‘black-box’ nature.

An example of why getting the human out of the loop is a good idea is the case of phonemes in automatic speech recognition (ASR). Linguists as well as Natural Language Processing

(NLP) researchers have been considering phonemes as the fundamental/natural units of spoken language. Consequently, when early ASR systems were trained, the learning systems were 'force-fed' phonemes without much consideration whether converting an audio waveform to phonemes to make it 'easier' for the machines was actually helping (Ostendorf and Roukos, 1989). Turns out, letting an E2E NN learn its own feature representations achieves better performance than using phonemes as intermediate classes. Keeping humans in the loop and using handcrafted features also keeps our assumptions and biases in the model. It also takes more time and effort to conduct a ML project. Hence, training a system E2E and letting it learn its own feature representation seems like the best way to let the data speak for itself. One cost to taking humans away from the loop is the taking away also of our hard earned skills, experience and expertise that could in fact save us time and require less data (Glasmachers, 2017). These are conflicting views but nonetheless, the ML momentum is towards E2E.

In deep lip-reading, the expectation from any architecture that claims to be E2E is the direct mapping of visual or audio-visual input into text label sequences as shown in Figure 3. The input to an ALR system is a sequence of vectors each representing an image frame of a video. Usually some form of encoder network maps this high dimensional input into a lower dimensional feature sequence of the same length. These features are then aligned and decoded to produce the output sequence which is different (usually smaller) than the input sequence e.g. 100 image frames to 30 ASCII characters.

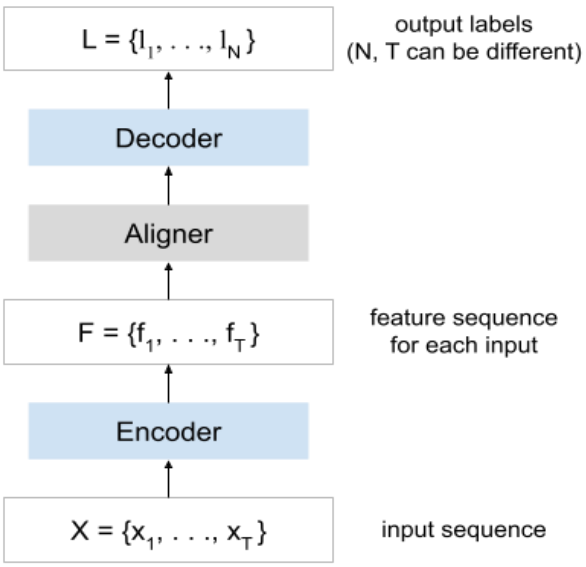


Figure 3: A typical E2E sequence model that takes a sequence of inputs and outputs a sequence. Lip-reading is a special case of a sequence model.
Note: This is a typical representation of an E2E architecture. It is not necessarily the defining structure as many E2E systems do not have a clear encoder-decoder separation as the whole system acts as a single complete structure.

where, x_i is the i^{th} item of input sequence X of length T,
f_i is the i^{th} item of feature sequence F of length T and
l_i is the i^{th} item of label sequence L of length N

Depending on which definition of 'end-to-end' from this discussion the researchers follow, there have been quite a few different flavours of 'end-to-end' ALR systems. Some seemingly end-to-end systems will be individually analysed in detail in section 4.3.

2.4. The limitations of end-to-end learning

While there are several benefits of E2E learning, it has its limitations and potential pitfalls. As the push for E2E is growing, so is the complexity of neural network architectures. It should not be forgotten that current neural networks themselves are modular structures. Hence, is it right to blindly treat every task as best suited for E2E? The limitations of E2E approach are laid out as follows (Glasmachers, 2017):

i) E2E sounds automated and enjoys a high degree of automation but is not fully automated:
    a) Human designers are still required, e.g. to design neural network layers with specific roles (e.g. the design of CNNs for computer vision)
ii) Problem decomposition is at the core of engineering ('Divide-and-conquer')
    a) The E2E paradigm seems to ignore that.
    b) It naively/dangerously assumes that mapping any random initial state to a non-trivial goal state using a simple gradient descent is a straightforward process.
iii) Data requirement grows exponentially for unmodelled interactions between modules.

iv) It is harder to know/ensure that the modules are learning what is intend off them i.e. E2E is a blacker box in terms of Artificial Intelligence (AI) explainability. This stems from the fact that E2E networks are relatively deeper and they take more control away from the human designers (e.g. learn their own features).

v) The efficiency of an E2E system is inversely proportional to the network complexity.
    a) For very complex networks, the E2E approach can fail completely. It is better to train modules one at a time in such cases.
    b) Unsupervised pre-training followed by fine-tuning can solve the issue.

2.5. Addressing the unique challenges in lip-reading and E2E lip-reading

Lip-reading is an extremely difficult task both for humans and machines. Some of the major challenges of the machine lip-reading and corresponding efforts towards their solutions are presented in this section. A discussion of how these challenges affect the E2E approach is also provided. The following challenges are addressed:

- Homopheme
- Dataset challenge
- Video quality
- Network depth
- Context
- Unseen classes
- Lip-reading in non-English
- Speaker dependence/variability

Homophemes

Visemes are the units of spoken language based on unique lip movements made during utterance. They are the visual equivalent of phonemes. 'Homophemes' are words with identical visemes. In the absence of audio information and/or context, these words are almost impossible to distinguish as they produce identical lip movements. They pose one of the toughest challenges in lip-reading. Figure 4 shows an example of a homopheme pair (the words: 'GAME' and 'NAME' which have the same viseme sequence [k, ey, p]) from the lipreading dataset LRS3-TED (Afouras *et al.*, 2018).

a) Word: 'GAME', Visemes: [k, ey, p]

b) Word: 'NAME', Visemes: [k, ey, p]

c) Visemes placed side to side

Figure 4: Frame sequences for utterance of homopheme words a) 'GAME' b) 'NAME' in LRS3-TED (Afouras et al., 2018).

To understand the extent of this challenge, a python script was written to generate homophemes for any given sentence. One output of the script is as follows:

Input sentence: "What is your name"

Sounds/Phonemes in each word:

- *what : [['W', 'AH1', 'T'], ['HH', 'W', 'AH1', 'T']]*
- *is : [['IH1', 'Z'], ['IH0', 'Z']]*
- *your : [['Y', 'AO1', 'R'], ['Y', 'UH1', 'R']]*
- *name : [['N', 'EY1', 'M']]*

Visemses based on these sounds/phonemes:

- *what: [‘W’, ‘AH’, ‘T’]*
- *is: [‘IY’, ‘T’]*
- *your: [‘K’, ‘AO’, ‘W’]*
- *name: [‘K’, ‘Ey’, ‘P’]*

Number of words with visemes identical to each of the given word:

[what, is, your, name] ⇒ [40, 13, 29, 76]

Number of possible sentences that look visibly similar to ‘what is your name’:

1,146,080

A few examples of these 1,146,080 sentences based on the CMU vocabulary (a pronouncing dictionary from Carnegie Mellon University):

- *wass id nohr kepp*
- *rus ede lor gape*
- *wise it corr lamme*
- *whyte ease loehr hemm*
- *wright it corps lam*
- *wythe is horr hamme*
- *weisse ede lore labbe*
- *was id your hemme*
- *what it core nahm*
- *white eat cor heppe*

As it can be seen, the majority of these sentences do not make semantic sense. Nevertheless, it goes on to stress the fact that there are over 1 million word combinations that look almost exactly the same as the sentence ‘what is your name’. Naturally, the problem gets worse for longer sentences. Lip-reading models need the assistance of an external language model or need to learn their own implicit language model (Afouras et al, 2018) in order to narrow down the options through the use of context. The challenge was also tackled by Fenghour *et al.,* (2020) using a Generative Pre-Training transformer (GPT) to learn a language model for visemes to word conversion. The group has also attempted disentangling homophemes using perplexity analysis (Fenghour *et al.,* 2020).

Homophemes pose a great challenge to the E2E approach to lip-reading as it necessitates the learning of the context. This makes the already complex E2E learning even more difficult. Although this is intuitive logic, it needs to be tested. The test would require a homophemic and a non-homophemic dataset to compare the performance of an E2E lip-reading system.

Dataset Challenges

The task of learning to lipread is highly complex. This has been reflected in the usage of huge neural networks with tens of millions of parameters (Afouras et al, 2018) to match the complexity of the task. This generates another problem: the need for training datasets with sizes to match these neural networks. Hence, E2E lip-reading systems require even bigger datasets.

In the early days of ALR, with handcrafted features and light-weight models, the focus was to create datasets in a controlled environment with as little 'noise' as possible, so as not to underfit the models (Matthews *et al.,* 2022; Messer *et al.,* 1999; Sanderson, 2002). An exhaustive, detailed list of audio-visual datasets from over two decades is available in this survey (Fernande-Lopez and Sukno, 2018).

One of the earlier attempts to fill the data void is the CUAVE dataset (Patterson *et al.,* 2002) The creators highlight the fact that researchers were forced to create their own dataset and available datasets suffered from speaker dependency. Their dataset containing 7,000 utterance samples of 36 speakers pronouncing 10 numeric words, despite spurring several researchers in audio-visual speech recognition, is now too limited for modern lip-reading architectures.

GRID corpus (Cooke *et al.,* 2006) focusses on the quality and quantity of audio-visual recordings. With 34,000 high quality clips of 34 speakers uttering English phrases, it is a sizable dataset and hence is still being used in smaller scale experiments. TCD-TIMIT dataset (Harte and Gillen, 2015) mixes some professionally trained lipspeakers among other speakers to see whether they would have an advantage over regular speakers. As expected, the experiments show that they do and by a significant margin. However, a model trained on lipspeakers would not be of much use for ALR applications in the real world. Creators of OuluVS2 dataset (Anina *et al.*, 2015) cite the low counts of speakers, utterances and constrained viewing angles in existing datasets to justify the need for their multi-view dataset with non-rigid mouth motion. Petridis *et al.* (2020) present the performance of a single model on various small-scale datasets. It is one of the commonly used datasets even now. Shillingford *et al.* (2019) mention the lack of open-vocabulary large scale datasets for

visual speech recognition and create Large Scale Visual Speech Recognition (LSVSR) dataset attempts to fill that gap with 3,886 hours' worth of video clips. Their new model based on LSVSR also set a new benchmark that significantly outperforms professional human lip-readers in Word Error Rate (WER).

More recent attempts have been towards creating audio-visual datasets that try to mimic the real world. Hence, they try to incorporate 'defects' like shaky videos, non-frontal faces etc. with the aim of training lip-reading models to be robust towards such defects during inference. To get around the time and cost of hand-labelling, which is one of the biggest challenges in creating large datasets, the group came up with an ingenious pipeline that 'watch' videos, capture the talking heads and use the audio and the subtitles to automatically create the labels. These 'in the wild' datasets have been the basis of most new breakthrough lip-reading models (Chung and Zisserman, 2016; Chung *et al.,* 2017; Stafylakis and Tzimiropoulos, 2017; Afouras et al, 2018; Shillingford *et al.,* 2019; Martinez *et al*., 2020). The BBC-Oxford LRW dataset (Chung and Zisserman, 2016), based on BBC programs, with 400,000 utterances of 500 English words from over 1,000 different speakers is arguably the best available word-level dataset to date. The creators went on to create even larger sentence-level datasets Lip-reading Sentences in the Wild 2 (LRS2) (Chung *et al.,* 2017), also based on BBC programs and LRS3 (Afouras et al, 2018), based on TED talks on YouTube. Table 3 lists some of the large wild datasets.

Lip-reading datasets for non-English language are starting to crop up too, although currently just a handful. Some examples are: the Mandarin dataset LRW-1000 (Yang *et al.,* 2019) and the Romanian language dataset LLRo (Jitaru *et al.*, 2020), just to name a few.

Considering human-labelling effort as one of the biggest obstacles in creating lip-reading dataset, zero-shot learning (Xian *et al.,* 2019; Wang *et al.,* 2019) that is aimed at classifying images in the lack of labelled training data, has also been suggested.

Table 3: *Some large wild datasets currently available for lip-reading*

Name	Year	Segment	Classes	Samples	Vocabulary
LRW	2016	Word	500	478,764	500
LRS2	2017	Sentence	17,428	118,116	41,427
LRS3-TED	2018	Sentence		118,516	51,000
LRW-1000	2018	Word	1,000	718018	1,000

Although, bigger and bigger lip-reading datasets are created and made available recently, the scale is nowhere near e.g. ImageNet on vision tasks with millions of samples, GPT-3 on language tasks with the whole of Wikipedia as dataset. It remains to be seen to what extent a complex E2E lip-reader trained on wild and noisy data like LRS3, would improve if the dataset was much bigger.

Video Quality Challenges

Datasets made up of clean, high quality, frontal-face videos and precise hand labelling do make it easier to train lip-reading models. However, these models will struggle in the wild as most real world videos will have some of the opposite qualities. This experiment (Seymour *et al.*, 2008) is one effort towards comparing the performance of image transformation based feature extractors in a clean version of a dataset and a version which was intentionally corrupted with 'jitters'. Jitters seemed to have a stronger negative effect to feature extraction compared to other defects like blurring. An example of video quality differences is shown in Figure 5. Figure 5.a) shows an image frame of a lipreading video from the GRID dataset (Cooke *et al.*, 2006) where there's very little background noise, the speaker's face is clearly visible and placed centrally in the image. There is also very little head movement between image frames in the video (not possible to depict here). This type of clean input reduces the workload of an ALR system and helps achieve better performances. Figure 5. b) on the other hand is an exact opposite. It is an image frame of one of the lipreading videos from the LRS2 dataset (Chung *et al.*, 2017). It has a noisy background, blurry image, head movement (in the video) and is of a lower definition (quality). Inputs like this pose a lot of challenges to an ALR system and consequently, performances on such 'wilder' datasets are relatively lower.

a) b)

Figure 5. Video quality differences between a) GRID dataset (Cooke et al., 2006) b) In-the-wild LRS2 dataset (Chung et al., 2017)

The extent to which a tiny perturbation in an image can fool a network into completely misclassifying the image was demonstrated by Su *et al.* (2019), where the image was 'attacked' by a noise as small as a single pixel. This effect can multiply in lip-reading where multiple image frames are used for classification.

In the context of E2E lip-reading, video quality is a tricky factor. One needs to decide whether to implement a smaller model on a high quality laboratory dataset or a complex E2E model on a large wild dataset. While E2E systems trained on a high quality dataset do produce good results (Assael *et al.,* 2016), the performance degrades on wild datasets even if the datasets are significantly bigger e.g. LRS3 (Afouras *et al.,* 2018).

Network Depth Challenges

It is an open fact that training very deep neural networks is difficult as well as time and data consuming (Srivastava *et al.*, 2015; He *et al.,* 2016). Most lip-reading architectures are quite deep networks with 10s of millions of parameters (Afouras *et al.,* 2018). Training the networks E2E means learning a very complex mapping between an input video to the output text. This approach necessitates even deeper networks.

To this end, a concept of 'Highway Networks' (Srivastava *et al.*, 2015) has been proposed aimed at easing gradient-based training of very deep networks. Highway Network has been implemented in lip-reading by Xu *et al.* (2018) in their novel LCANet, which achieved a 12.3% improvement on Word Error Rate (WER) over the then state-of-the-art methods on the GRID corpus.

One seminal work on easing the training of deep networks is the ResNet (He *et al.*, 2016) where the network layers are reformulated to learn residual functions with respect to the layer inputs and the resulting network was proven to be more easily optimisable while gaining in accuracy from increased depth. The use of ResNet as a part of the visual frontend has been a common trend in modern lip-reading architectures (Afouras et al, 2018; Assael *et al.*, 2016). The instability of deep networks against small perturbations has also led to the development of 'stability training' technique (Zheng *et al.*, 2018) with proven increase in their robustness in image classification tasks.

Context

One aspect where machine lip-reading systems consistently fall well behind human lip-readers is the use of context (Fernandez-Lopez and Sukno, 2018). The problem of context has been tackled by some with the use of external language models with expensive beam searches. If a network is required to learn the context on its own i.e. its own simplistic language model in an E2E fashion, the sequence processing part of the lip-reading architecture is often a deep attention based network. Transformers are commonly used to learn the context (Afouras et al, 2018). However, this makes the overall architecture quite deep forcing a multistage training strategy involving pretraining of separate modules. This strategy is what we would desire to evolve out of in the quest for purer E2E lip-reading. The human ability to arbitrarily fetch context from seemingly infinite timesteps is something an ALR system can currently only aspire to do.

Unseen Classes

Most current ALR approaches undertake lip-reading as a classification problem where the classes can be individual words, phrases or even sentences. This creates a problem in inference where the model trained on a limited class training set comes across an unseen

class. E.g. a model trained on LRW dataset (Chung and Zisserman, 2016) with 500 word classes is shown a word e.g. 'apple' which is not in the training data.

The most straightforward solution to this issue would be to go out and gather more data with a bigger vocabulary. But advances in ML techniques like zero-shot learning and Generative Adversarial Networks (GANs) have been used to get around the issue. Singla *et al.,* (2020) have been able to improve the accuracy of phrase classification on Oulu dataset by 27% with their novel implementation of zero-shot learning using GANs to generate samples of new classes.

Other Languages

To no one's surprise, the majority of lip-reading experiments are done and hence the resulting progress has been made in the English language. Given the language agnostic applications of lip-reading, progress surely is needed in non-English ALR research. To this end, over the years, whatever progress has been made can be roughly grouped into two different bins: i) creation of datasets (Yang *et al.,* 2019; Jitaru *et al.*, 2020) etc. to train the models directly on the target language ii) use of relatively language independent features like visemes in English language dataset with an expectation of implementing it for a second language (Fenghour *et al.,* 2020).

Zhao *et al.* (2019) have pointed out the fact that even with the availability of target language dataset, Chinese Mandarin poses higher ambiguity to conventional lip-reading architectures because of its tone/pitch-based semantics unlike the word/sentence-based English. Their proposed Cascade Sequence-to-Sequence Model for Chinese Mandarin (CSSMCM), designed to model tones based on visual features and syntactic structure and trained on their own Chinese Mandarin Lip-reading (CMLR) dataset, surpasses the state-of-the-art frameworks. Transfer learning from English dataset trained models to a second language is yet another option (Jitaru *et al.*, 2021).

Languages that pose higher ambiguity might require a deeper network with matching complexity or require a human expert to design intermediate features. Neither of the two

facilitates the E2E approach. A deeper study is required to create novel network architectures that can efficiently train E2E on a highly complex language.

Speaker Dependence/Variability

Speaker dependence refers to the fact that ALR systems do not perform as well with test speakers that were not present in the training data. Speaker variability can mean any or a combination of factors ranging from the differences in lip shape, lip thickness, facial hair etc. to accents and non-lip gestures. It adds to the complexity of the task and makes E2E ALR more difficult.

A detailed survey on speaker dependence in lip-reading is provided by Burton *et al.* (2018). Their experiment shows that their speaker-independent tests, where the speakers in the test dataset are not included in the training dataset, underperforms the speaker-dependent state-of-the-art on the TCD-TIMIT dataset. Another quite common speaker variability issue is the viewing angle of the speaker's face. A model trained on predominantly frontal-face data will not be as effective for faces in various profile view angles. One ingenious solution to this problem is proposed by Cheng *et al.* (2020), where the LRW dataset is augmented with synthetically generated faces at various angles using a 3D Morphable Model (3DMM). Their experiment for word recognition on the LRS2 dataset achieves a 2.55% improvement. Figure 6 illustrates how different the inputs can be for the same output (word 'ABOUT' in this case). There are variabilities present in facial angle, facial hair, presence/absence of glasses, eye movements, shadows, skin tone, frame coverage (% of the frame occupied by the face) etc.

Figure 6: The same word 'ABOUT' uttered by two quite different speakers.

The most natural way to create an E2E speaker-independent ALR model would be to use a dataset that contains all sorts of variabilities. However, such a dataset will have to be very large to allow enough samples for each variation.

2.6. Rationale and criteria for a pure E2E ALR system

Rationale for choosing E2E

If the E2E approach has some obvious limitations, why then should it be pursued? Taking a step back from the finer issues, challenges, solutions etc. and looking at the bigger picture, it is apparent that liberating humans out of the loop for full automation is the machine learning ideal. E2E is an approach in that direction. However, since what is currently available in that front is nowhere near perfection, some caution is necessary. At least for now, an expertly human designed architecture followed by 'full automation' seems to be the way to go for E2E. A detailed discussion of the rationale is provided in section 4.1.

Criteria for pure E2E ALR

We propose that a pure E2E ALR system should meet the following criteria:

i) All modules of the pipeline should be differentiable.

ii) Gradients should flow from one end of the system to the other end.

iii) It should follow a unified training scheme without any:
   a) modular training.
   b) pretraining stages including curriculum learning.

iv) Pre-processing should be kept at a minimum:
   a) Non-intelligent processing such as grey-scaling to reduce input size, cropping the frames to a given size etc. can be included.
   b) Intelligent processing such as the use of externally trained ROI detectors should not be included. The ALR system itself should be able to learn to focus on ROIs through training.

v) If the system works at the sentence-level, it should learn its own language model to align the intermediate outputs such as characters, visemes, phonemes, words etc into sentences. It should not depend on an externally trained language model for the alignment of these speech units.

3. Literature Review

The chapter is divided into the following sections:

i) A review of the feature extraction techniques in ALR
ii) Various classification schemas and how they affect the E2E approach
iii) A review of various tools and techniques used in lip-reading and where applicable, a discussion of how each of them relate to the E2E approach
iv) A discussion on how ALR uses language models to decipher the output
v) A review of a few pre-E2E and non-E2E ALR approaches
vi) A review of several different E2E ALR systems

3.1. Feature extraction in lip-reading: visual and temporal

Extraction of visual features from a talking video is the first stage of ALR. Researchers have tried a variety of techniques in the past in an attempt to best represent the rich input information. Traditional feature extraction techniques in VSR tend to fall into the categories (Dupont and Luettin, 2000) shown in Figure 7:

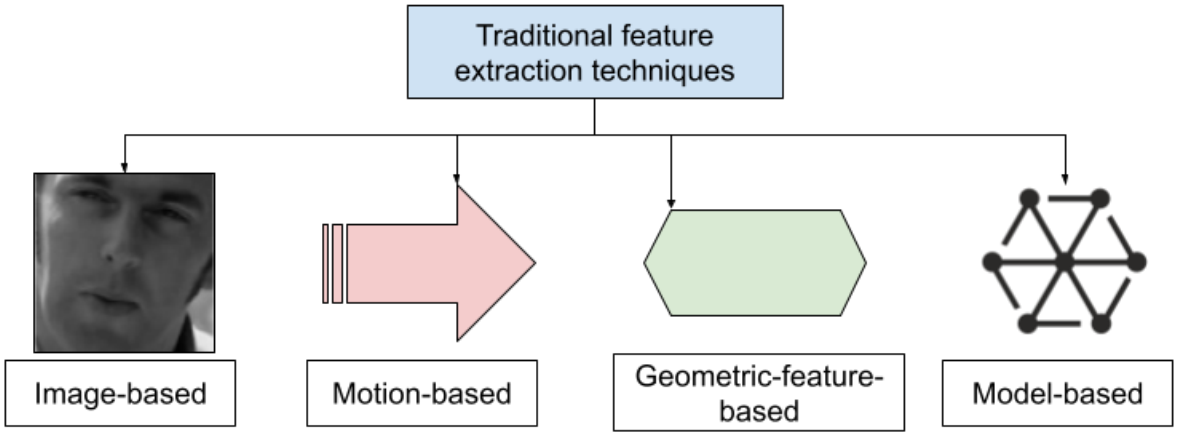


Figure 7: Traditional feature extraction techniques in lip-reading.

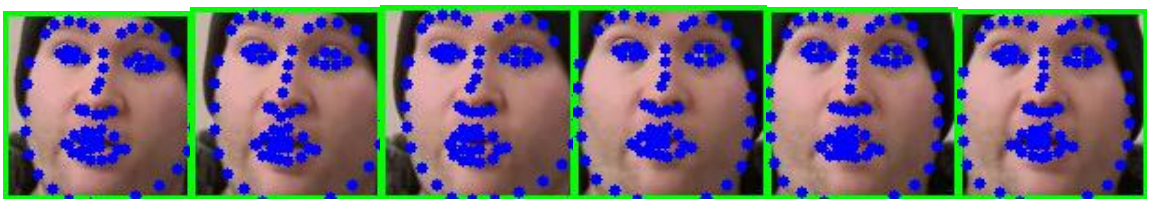
i) Image-based: images of the speaking mouth are used directly at grey-level or after some transformation to create the feature vectors.
ii) Motion-based: assumption that movements in the images during the utterance must be related to speech and hence should be useful for decoding.
iii) Geometric-feature-based: consider metrics like mouth height, width, perimeter/opening etc. to contain important information.

iv) Model-based: create a model of lip contours or other speech articulators into a low dimensional space.

All four techniques relate to non E2E systems of the pre deep learning era. Human assumptions are used to decide what features the lip-reading systems would find most useful. The evolution of deep neural network architectures have made these techniques redundant to some extent by learning the suitable features automatically. This has enabled E2E training of ALR systems. However, due to the huge volume of information contained in a lip-reading video, even these E2E systems go through various pre-processing stages.

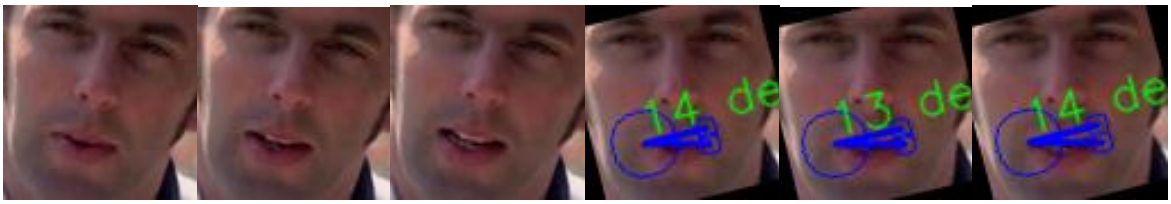
A universal trend to reduce the volume of information in input video is to ignore areas in the image frames outside of the face/mouth region. Cropping a mouth ROI by hand is prohibitively expensive, especially for large datasets. Luckily, there are great tools available to automate that process. DLIB (Davis, 2009) is an excellent open source toolkit that comes with methods and a trained model to detect faces and up to 68 facial landmarks from images. It has been extensively used in ALR for mouth ROI detection as a part of data pre-processing. Figure 8 shows a typical usage of the library on video image frames. As shown in stages a) to e), the original image frames go through a face detector that uses 68 different landmarks for eyebrows, eyes, nose, mouth and jawline. The pixel coordinates for these landmarks can be used to crop various regions of interest in the image e.g. the mouth area for lip-reading. The information can also be used to correct the facial angle relative to the image frame as illustrated in Figures 8.c), d) and e).

a) Original frames

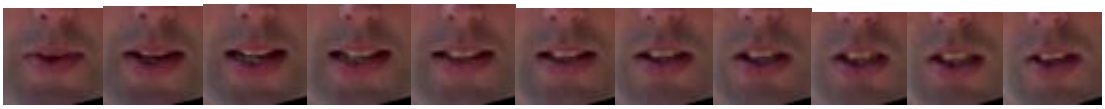


b) Face detected and cropped

c) Face rotated based on lip bisector



d) A more severe case of rotation correction (3 angled faces rotated to vertical)



e) Mouth ROI detected and cropped after rotation

Figure 8: Mouth ROI detection using DLIB (Davis, 2009) python library

Another tool for face detection often found in lip-reading literature is RetinaFace (Deng *et al.*, 2019). However, it only detects 5 facial landmarks and seems to struggle when face size in images is large.

The following sections review various techniques developed in an attempt to best capture the spatial and temporal information present in VSR/AVSR (Audio Visual Speech Recognition) data. The varying levels of success of each technique is also presented.

Zhao *et al.* (2009) introduced the use of Local Binary Pattern (LBP) (Ojala *et al.*, 1994) operator, in an effort to represent both spatial and temporal information from grey-scaled image frames for lip-reading. Their experiment on AVLetters database achieved a new high classification accuracy of 62.8%.

Considering the amount of unnecessary or even adversarial information a talking face video contains relative to the amount of information in the output, a clever technique (Tang *et al.*, 2015) extracts spatiotemporal features from only the significant regions rather than the

whole frame. The proposed method saves a considerable amount of computation time by reducing noise and also is able to achieve a high classification rate on YouTube, KTH and Hollywood2 datasets. While this experiment was performed on a video classification task, it can be used for ALR.

Rekik *et al.* (2014) have implemented HOG and Motion Boundary Histogram (MBH) to extract visual and motion features. Their SVM based system has achieved a good performance on speaker dependent lip-reading.

An attempt towards a compact representation of visual information has been made by Zhou *et al.* (2014) via the use of latent variables that separately encodes:

- variation between speakers;
- variations caused due to the utterance.

In the very related domain of video classification, which also involves the extraction of appearance as well as motion information, Tang *et al.* (2019) cite the high computational cost of using optical flow as the motion information extractor. To solve this, MoNet, a novel network, is proposed, that successfully 'hallucinates' motion from just appearance features without the use of optical flow. This technique has been shown to reduce the computation cost by almost half.

Wu *et al.* (2016) have presented a novel lip descriptor that uses both geometric and appearance information to tackle the issue of varying utterance mannerisms that often lead to false prediction. An advanced face landmark detection method is utilised to generate a set of geometric features.

Evolution of CNNs has given rise to researchers experimenting with their 3-dimensional versions 3DCNNs, now with a time dimension to not only efficiently capture the spatial information like its 2D counterpart, but also the temporal features (Tran *et al.,* 2015). CNNs are also much simpler to understand and easier to train. Weng and Kitani (2019) have taken the 3-dimensional CNNs even further and swapped what is called a conventional

combination i.e. 3DCNN+2D ResNet with their 2-stream 3DCNN to achieve a 5.3% word accuracy improvement on the LRW art.

A video contains two types of information: spatial and temporal. If we assume that there is a lot more information in the spatial dimension compared to the temporal dimension, which is a fair assumption, then why not treat the two information differently via two different channels in the network? The SlowFast (Feichtenhofer *et al.,* 2019) network does exactly that. It passes the visual information on a pathway at a low frame rate to capture spatial semantics, while a faster pathway at a higher frame rate, but lighter with reduced channel capacity, captures motion features. This technique achieves the state-of-the-art in video recognition tasks. Implementation of SlowFast in lip-reading seems promising but is yet to be seen.

3.2. Classification schemas in lip-reading

The choice of a classification schema can determine whether or not an ALR system can be trained E2E. Majority of current lip-reading systems are designed to decode longer speech segments like words, phrases and sentences (Fenghour *et al.,* 2020). A common technique across most of these systems is the use of words or characters as labels for classification (Stafylakis and Tzimiropoulos, 2017; Chung and Zisserman, 2016; Chung *et al.,* 2017; Afouras et al, 2018; Assael *et al.,* 2016). Systems with these class schemas can in theory be trained E2E and some are. Since the use of words as classes can be troublesome for inference of unseen words, few other alternatives to ASCII characters viz. visemes and phonemes have been tried with a fair amount of success. However, the use of visemes/phonemes as classes can require a separately trained decoder e.g. the 'viseme-to-word' converter of Fenghour *et al.* (2020); thus not allowing the whole system to be trained fully E2E. The following is a discussion of different techniques to improve performance using visemes or phonemes as classes.

Quoting the language dependence of visemes, although words, characters and phonemes are relatively highly language dependent, Bastanfard *et al.* (2009) have attempted a viseme classification system for lip-reading of Persian language using their own visemes. The claim for visemes' language dependence has been supported by the fact that Tehrani Persian

language without any accent does not contain any diphthong unlike e.g. English. Eigenlips of each phoneme is used to create the corresponding viseme. The classification result indicates the robustness of their algorithm.

Bear and Harvey (2016) proposed a novel two-pass method where phoneme classifiers were trained on pre-trained viseme classifiers. One possible motivation behind the study is the suggestions from earlier works that phoneme classification can outperform viseme classification under the right circumstances (Howell *et al.,* 2016). Nevertheless, the novel algorithm is able to achieve a better classification accuracy compared to previous results in lip-reading.

Two major ways of classifying visemes have been pointed out (Cappelletta and Harte, 2011):

i) using facial image, lip shapes and other visual speech articulators;
ii) using phonemes to map visemes: which are further divided into linguistic and data-driven methods.

The later approach is much simpler and allows for an easier preparation of a dataset and hence is more commonly used. A detailed comparison (Jachimski *et al.,* 2017) of viseme classification approaches has shown that a combination of geometrical and linguistic features results in a better clustering of newly defined visemes.

Using phonetic visemes as the only cues, Fenghour et al. (Fenghour *et al.,* 2020) have been able to set a new 15% lower WER on the LRS2 dataset. Their ALR system has two stages:

i) input video to visemes classifier;
ii) video to text converter.

Besides, unlike word based systems, the viseme based system seems to be able to recognise words not seen during training. Usage of external text data to train a viseme-text converter is also used by this system (Peymanfard *et al.,* 2022).

A purely phoneme based model has been shown to perform quite well too. This work (El-Bialy *et al.,* 2022) investigates various classification schemas but implements phonemes as class labels to achieve a 10% lower WER on LRS2 dataset.

There seems no apparent winner in the debate between visemes vs phonemes as being the better class label for lip-reading. A comprehensive comparison between the two has been carried out by Thangthai *et al.* (2018) based on the TCD-TIMIT corpus. The results can be summarised into two findings:

i) Phonemes outperform visemes in word accuracy

ii) Visemes achieve higher accuracy at the unit level (at the level/stage of classification of visemes given the images) but suffer in accuracy during sentence/word construction using these predicted visemes (Fenghour *et al.*, 2020).

Figure 9 summarises the various classification schemas found in lip-reading literature. Using video image frames as input, the ALR system can be tasked to classify any of the label types viz. visemes, phonemes, characters, words or sentences. Then depending on the objective, the final output is generated in the required format viz. characters, words, sentences.

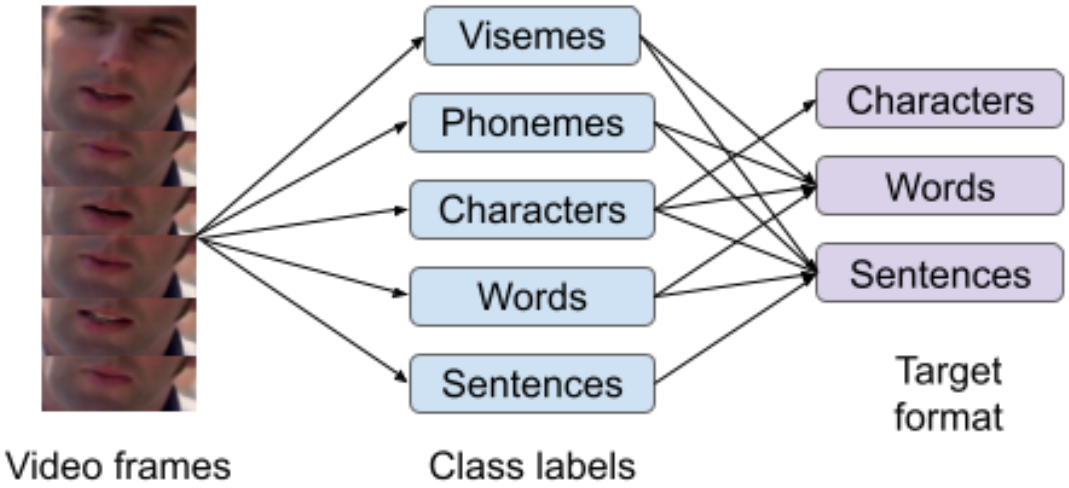


Figure 9: Various classification schemas seen used in lip-reading literature

3.3. Various tools and techniques in lip-reading

3.3.1. Audio-visual fusion

Pure lip-reading only takes cues from visual information like lip movement and other secondary speech articulators. However, research on AVSR is more abundant than pure VSR. This is due to the fact that ASR has a lot more widespread applications compared to somewhat niche VSR. Including visual information has been thoroughly proven to significantly improve speech recognition, especially in noisy audio settings. Having said that,

innovations made in AVSR, sometimes even in ASR, almost always benefit VSR/ALR. With this in mind, the following is a brief survey of audio-visual fusion technologies developed over the years.

By training audio and visual branches of a network separately before fusing them into a new hidden layer and following with yet another deep network trained on the joint feature space, Mroueh *et al.* (2015) have been able to reduce the phoneme error rate by more than 5% compared to audio-only networks. Their pre-processing includes cropping a 64 x 64 pixel mouth ROI and reducing it to 60 using Linear Discriminant Analysis (LDA).

Tao and Busso (2018) observed that when the audio information was clean, adding visual cues did not help much. It rather often negatively affected speech recognition by adding more variability. To this end, a gating layer in the deep neural network was added to diminish the effect of visual information that was not helpful. The system seemed to perform slightly better or at least as good compared to an audio only system with the same high quality audio.

Visual cues were utilised to decipher audio in settings with multiple speakers or presence of background noise (Ephrat *et al.,* 2018). The experiment not only demonstrated this, but also produced a new dataset AVSpeech that contained 1000s of hours of video clips from the internet. The issue of overlapped speech has also been tackled by Yu *et al.* (2020) using their integrated audio-visual network. The system tried to implicitly separate and recognise both modalities with a single cost function and was able to significantly outperform its audio-only baseline using overlapped speech data simulated using the LRS2 dataset.

Figure 10 shows a simplification of the WLAS model (Chung et al., 2017) for sentence-level decoding. The raw audio waveform is converted to Mel Frequency Cepstral Coefficients (MFCC) before feeding it to an RNN. The visual data (image frames) pass through a CNN before also feeding into an RNN. The outputs of the audio and visual RNNs are then concatenated and fed into another RNN and a FCN to produce the output sequence (sentence).

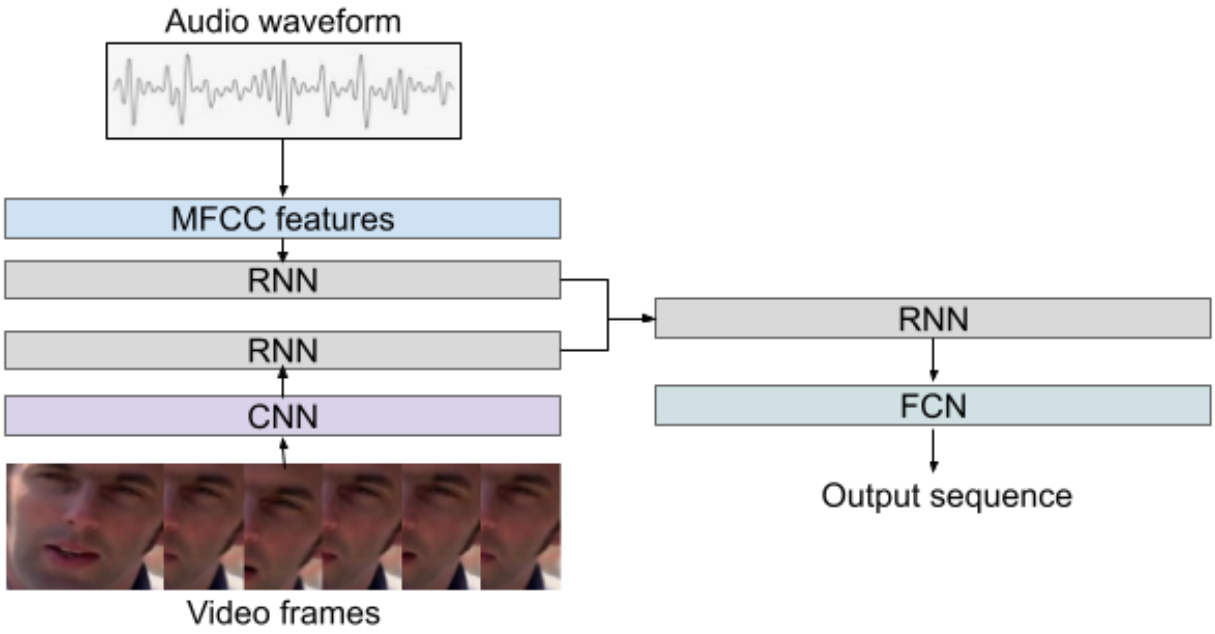


Figure 10: Typical Audio-Visual Fusion method in lip-reading.

We can see that although AVSR and VSR have many things in common and benefit each other, audio-visual fusion is not a primary factor determining the E2E-ness of a lip-reading system.

3.3.2. The use of CNNs

Since their inception in the modern form in the 90s, CNNs have become the backbone of computer vision tasks. ALR makes abundant use of CNNs, first beginning with the frontend as visual feature extractor, then gradually in the classification backend. This section discusses the various types and techniques of CNNs used in lip-reading, their contributions in the progress of this domain as well as how some of them are helping the E2E quest.

CNN's importance in image classification has been known for decades now, but its ability to extract spatiotemporal features was cemented by this seminal work (Karpathy *et al.,* 2014) on video classification.

Besides image classification, ConvNets are also found to be good at sequence modelling. One might be curious as to why ConvNets would be chosen for sequence processing over RNNs like LSTMs or GRUs, which are specifically designed for the purpose. In fact, evidence is piling

in favour of ConvNets for sequence processing tasks. It is useful to note here that as the sequences get longer RNNs become harder to train and suffer from gradients disappearing or exploding over longer range updates. A thorough comparison has been experimented by Bai *et al.* (2019) to see which one is the clear winner for sequence modelling. Their work challenges the conventional default relationship between sequence modelling and RNNs.

ConvNets are very efficient at reducing input dimensions and use a fraction of the number of parameters compared to e.g. a similarly sized FCN. However, deeper networks designed for heavier tasks can still be very computationally expensive. ShuffleNet (Zhang *et al.,* 2018) is a CNN variant aimed at lower computing power devices and hence required the network to be a lot more computationally efficient. To achieve this, it has utilised two novel operations:

- pointwise group convolution;
- channel shuffle.

Experiments on ImageNet (Krizhevsky *et al.*, 2017) classification tasks have shown that while being a lot lighter in terms of parameters, it has comparable accuracy to much bigger ConvNets like AlexNet. This success of ShuffleNet has made them useful in E2E lip-reading as evidenced in (Martinez *et al*., 2020).

One observation that has been made from training very deep ConvNets over the years is that if they have shorter connections between layers close to the input and the output, they train faster and better. DenseNet (Huang *et al.*, 2017) builds onto this idea by connecting all layers to each other as in a feed forward NN. DenseNets have been shown to ameliorate the vanishing gradient problem inherent in very deep networks. The direct connections strengthen the propagation and reuse of features and significantly cut the number of parameters. Chen *et al.* (2020) have made use of DenseNet in lip-reading in combination with residual bidirectional LSTM (resBi-LSTM) for sentence level Mandarin. In an attempt to learn finer level lip movement this work (Chen *et al.,* 2020) proposes yet another ConvNet variant called hierarchical pyramidal convolution (HPConv), which achieves a 1.53% gain on LRW WER state of the art.

Ma *et al.* (2021) raise the LSW state-of-the-art even further with a novel DS-TCN that also runs significantly faster than the original LRW model.

3.3.3. Connectionist Temporal Connection

Lip-reading, like any sequence-to-sequence (seq2seq) task, faces these two common difficulties:

i) alignment of the output;
ii) output post-processing.

To this end, many recent E2E NLP tasks have adopted techniques like Connectionist Temporal Connection (CTC), attention and RNN Transducer (Wang *et al.*, 2019).

The emergence of CTC (Graves *et al.,* 2006) has accelerated the growth in the number of deep neural networks in NLP that are trained E2E by tackling the two difficulties as follows:

i) CTC does not require the segmentation and alignment of labels in the training data as it produces its own alignment of the output labels. It does so by first enumerating all possible hard alignments then aggregating them to generate soft alignments.
ii) Post processing from an intermediate output e.g. phonemes into the final graphemes is not required either as CTC directly outputs the target labels.

Solving these two problems allows the network to be trained E2E by mapping input sequence directly to target sequence. However, CTC inherently assumes during hard alignment that the output labels are independent of each other, thus not being able to learn the context/language model. This necessitates lip-reading systems to make use of an external language model to decipher semantics from the outputs, thus questioning their 'E2E'-ness. Another issue with CTC arises when the output sequences are longer than the input sequence. CTC is not designed for scenarios like such and hence is helpless. Although in lip-reading, since it usually takes one or more frames of an average frame-rate video to articulate a character, input sequences are typically longer than output sequences. This means CTCs don't often come across such scenarios. However, they cannot be totally avoided without human intervention during data cleaning: e.g., an analysis of the LRS2 training set with 45839 video samples shows that in 27 of the samples (0.058%), the number of characters in the label is higher than the number of frames in the video.

Figure 11 illustrates the basics of how CTC collapsing works. For this case, characters are used as the class labels, but the labels can also be phonemes, visemes etc. A character is predicted for each input image frame of a lip-reading video. Repeated characters are merged and a blank token provides a workaround to identify characters that are actually repeated in ground truth labels.

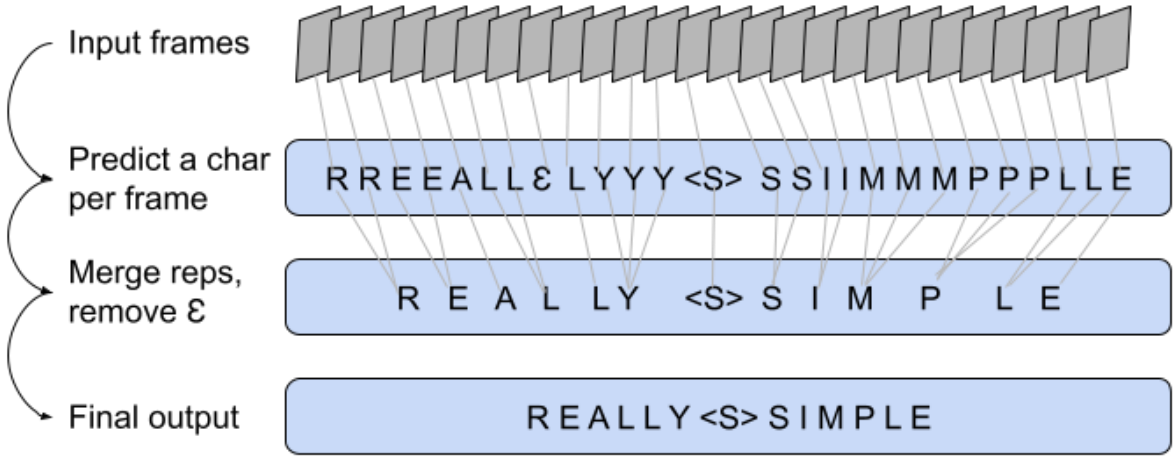


Figure 11. CTC collapsing works by predicting a token for each input (frame in this case). Repeated tokens are merged. Frames with no information will generate a 'blank token': 'Ɛ ' which can be ignored.

How it works:

$$Frame\ sequence\ (X) = [x_1, \quad x_2, .., x_T]$$
$$Output\ sequence\ (Y) = [y_1, \quad y_2, .., y_V]$$

The frame-sequence length (T) does not necessarily equal the output sequence length (*V),*
In the example shown in Figure 11:

$$first\ alignment = [R, R, E, E, A, L, L, \varepsilon, L, Y, Y, <S>, S, S, I, I, M, M, M, P, P, P, L, L, E]$$
$$Y = [R, E, A, L, L, Y, <S>, S, I, M, P, L, E]$$

CTC allows any similar alignment that ultimately maps to *Y* after the *'merge and drop Ɛ'* process. Hence, other valid alignments could be:

$$[R, R, R, E, A, L, L, \varepsilon, L, L, Y, <S>, S, S, \varepsilon, I, I, M, M, P, P, L, L, L, E]$$
$$[R, E, E, A, A, L, L, \varepsilon, L, Y, Y, <S>, S, S, S, I, M, M, M, \varepsilon, P, P, L, L, E]$$
$$etc.$$

At each input step (*t)* e.g., of a RNN network, the CTC provides a output probability distribution (p_t) over all characters in the set *X* plus the blank character (*Ɛ)*:

$$p_t(a \mid X)$$

Calculated over the outputs: $\{R, E, A, L, Y, < S >, S, I, M, P, \mathcal{E}\}$

For a given *(X, Y)* pair, the CTC loss function can be summarised as:

CTC conditional probability =

marginalisation over the set of all valid alignments (probability for each alignment per time-step)

$$p(Y \mid X) = \sum_{A \epsilon A_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X) \quad (1)$$

where, $p(Y \mid X)$ is the conditional probability, T is the total number of time-steps and A is the alignment.

One problem with this approach is, for the majority of sequence tasks, the possible number of valid alignments can be quite big. This makes calculating the loss very computationally expensive. A solution to that is merging alignments that result in the same output at a given time-step with a dynamic programming algorithm.

3.3.4. RNN Transducer

RNN-transducer (Graves, 2012) also enables E2E training of lip-reading systems by producing alignments like the CTC loss function. It also uses a 'blank' character and calculates and aggregates the probabilities to get the target sequence. But unlike CTC, it does not make the assumption of labels' independence as each state influences the updates of the subsequent state and the output labels. Also unlike CTC, it does not suffer from the problem of output sequence length being greater than the input. The use of RNN-transducers in lip-reading has enabled some performance improvements (Makino *et al.,* 2019). However, RNN-transducers have been known for being hard to train unless they go through a modular pre-training stage. Also it has been pointed out that having to encode the input as a fixed length vector limits its encoding ability (Wang *et al.*, 2019).

Let $x = (x_1, x_2, \ldots, x_T)$ be an input sequence of length T in the set X^* of all sequences over an input space X, and $y = (y_1, y_2, \ldots, y_U)$ be an output sequence of length U in the set Y^* of all sequences over an output space Y. The RNN transducer defines a conditional distribution given by:

$$P(y \in Y^* | x) = \sum_{a \in \beta^{-1}(y)} P(a|x) \quad (2)$$

where $a \in \underline{Y}^*$ are the alignments, and $\beta : \underline{Y}^* \rightarrow Y^*$ removes null symbols from the alignments in Y^*.

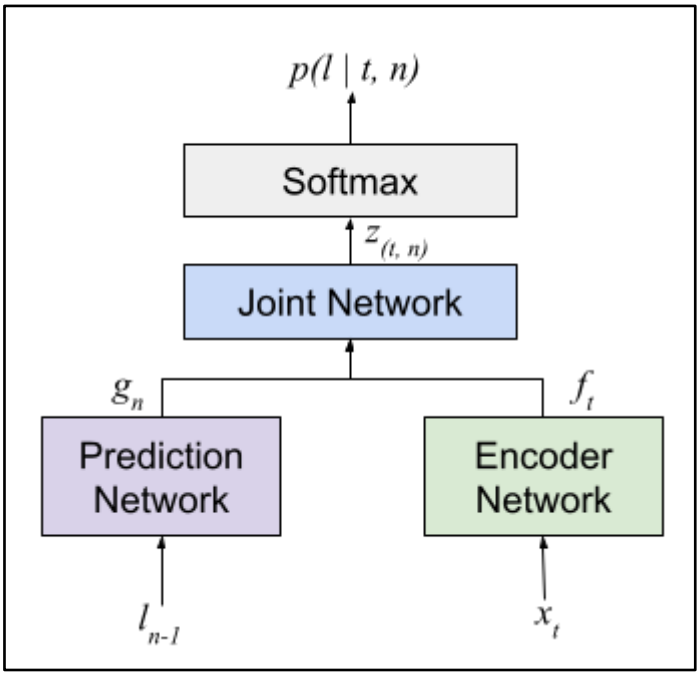


Figure 12: RNN-transducer (Graves, 2012)

The network consists of three sub-networks as shown in Figure 12:

i) Transcription network $(F(x))$: A bidirectional RNN encoder that transcribes input sequence $x = (x_1, x_2, \ldots, x_T)$ into a transcription(feature) sequence $F = (f_1, f_2, \ldots, f_T)$. At input x_t at any time t, the encoder outputs $f_t = F(x_t)$, a $|V| + 1$ dimensional vector.

ii) Prediction network $P(l)$: A RNN decoder network with one input layer, a single hidden layer and an output layer that works as a language model by working out the interdependencies between output labels. The network has a hidden state (h_n) and an output value (g_n) for each label (l_n) in any location $n \in [1, N]$.

$$g_n = P(l_{[1:n-1]}) \quad (3)$$

iii) Joint network $J(f, g)$: Aligns the input and output sequence. For encoder(transcription) output vector (f_t) , prediction output vector (g_n) and label $k \epsilon \underline{Y}$ at output location (n), the output density function e is given by:

$$e(k, t, n) = exp(f_t^k + g_n^k) \quad (4)$$

$$p(k \epsilon \underline{Y}|t, n) = \frac{e(k,t,n)}{\sum_{k' \epsilon V'} e(k',t,n)} \quad ,\forall\ t \epsilon [1, T], \quad n \epsilon [1, N] \quad (5)$$

3.3.5. Attention

Attention mechanism (Bahdanau *et al.*, 2015) greatly supports the E2E approach by enabling the network to learn its own implicit language model. This means a separately trained external language model, which is against the E2E ideal, is not always required.

Attention was originally applied to Neural Machine Translation (NMT) has since been commonly applied (Afouras et al, 2018; Chung *et al.*, 2017; Petridis *et al.*, 2018) to ALR due to the similarity of the two tasks i.e. sequence to sequence modelling. One major benefit of attention based E2E seq2seq is that it does not necessitate the input to be encoded into a fixed length vector. Encoder-decoder structures that utilise the attention mechanism enable implicit soft alignment between the inputs and outputs.

A variant of attention called Transformer (Vaswani *et al.*, 2017) has been successfully used by Afouras et al (2018) to set the state-of-the-art in sentence level lip-reading on LRS2 dataset. It seemingly outperforms their fully convolutional as well as RNN based architectures in WER. Transformer also learns its own implicit alignment and does not depend completely on external language models compared to their other two models, although the use of a language model trained on the LRS2 subtitle corpus does seem to

improve its performance marginally. It is to be noted however that the visual frontend of this model was pretrained before the whole network was trained E2E.

While the CTC assumes conditional independence between output characters, the attention-based model learns its non sequential alignments. In an attempt to force monotonic alignments and avoid the conditional independence assumption, Petridis *et al.* (2018) make use of both the CTC and attention in a hybrid architecture (Watanabe *et al.,* 2017) for lip-reading. Their audio-visual based model trained on LRS2 dataset is able to reduce the WER by 1.3% compared to the audio-only state-of-the-art. Attention in combination with CTC has also been demonstrated to improve computation time and model accuracy by Xu *et al.* (2018) with their novel LCANet architecture.

Figure 13 shows the original Transformer architecture (Vaswani *et al.,* 2017). Input sequences go through a positional encoding stage that adds positional information to each input (as a Transformer does not have recurrence like RNNs). It is then fed to a multi-head self-attention mechanism where the same input embedding acts as key, query and value. This is detailed in Figure 14. The output of the encoder block is fed to the multi-head attention of the decoder as key and query. The previous output of the decoder serves as the value after passing through its own multi-head attention mechanism.

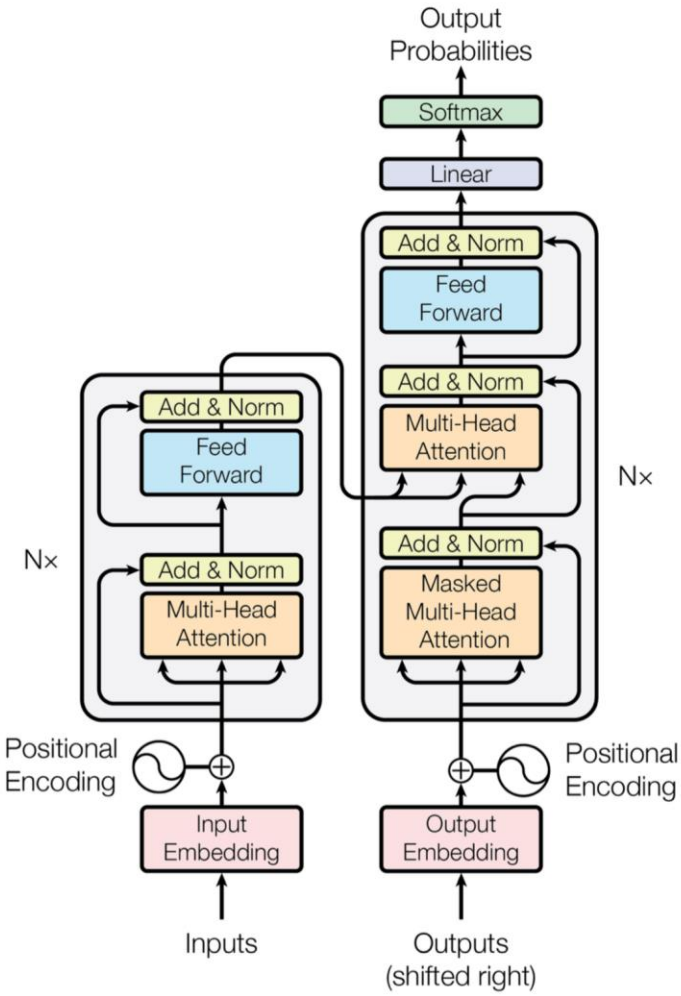


Figure 13. Transformer architecture (Vaswani et al., 2017)

How it works:

While most of the components of the Transformer are common across neural networks, the novel concepts: multi-head attention and positional encoding are summarised as follows:

Let $X_{m\,x\,n}$ be the input matrix.

Key (K), Query (Q) and Value (V) are generated as follows:

$$K = W_k X$$
$$Q = W_q X$$
$$V = W_v X$$

Where W_k , W_q *and* W_v are key, query and value weight matrices, respectively.

A scaled product of *K and Q* is taken as:

$$W_i = \frac{Q_i^T K_i}{\sqrt{k}} \tag{6}$$

Where *k* is the input embedding dimension. Attention scores are calculated as:

$$W_i = softmax(\frac{Q_i^T K_i}{\sqrt{k}})V_i \tag{7}$$

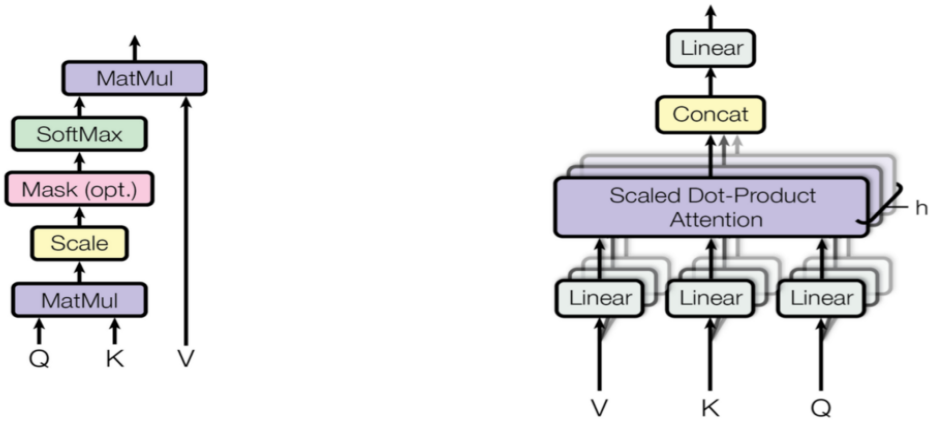


Figure 14. Scaled dot product (left) in a multihead-attention mechanism (right) (Vaswani et al., 2017). The use of multiple attention heads gives the network a greater discriminative power.

The sinusoidal positional encoding has a simple role of providing a sort of position index to the items in the input sequence e.g. words in a sentence input. Without positional encoding, the network is *permutation invariant* e.g. it would not know the difference between a sequence *{'draw,' 'a', 'hand'}* and *{'hand,' 'a', 'draw'}*. The simplest approach to include position information is to one-hot encode the positions of the inputs. The original approach instead uses sinusoidal embeddings.

$$P_{k,2i} = sin(\frac{k}{10000^{2i/d}}), \quad P_{k,2i+1} = cos(\frac{k}{10000^{2i/d}}) \tag{8}$$

where, P is the positional embedding for an input with position k in a sequence of inputs, d is the embedding dimension of the input e.g. word/char or any other token and i is the individual item of the input embedding. Hence, if $d = 4$, w is the current word with

embedding e_w, and e'_w be the final embedding after including the positional embedding based on sinusoidal encoding:

$$e'_w = e_w + [sin(\frac{k}{10000^0}), cos(\frac{k}{10000^0}), sin(\frac{k}{10000^{2/4}}), cos(\frac{k}{10000^{2/4}})] \quad (9)$$

3.3.6. Teacher-student / knowledge-distillation

As it must be apparent by now, the task of lip-reading has been approached via many different angles resulting in many different models with their own strengths and weaknesses. It then should naturally occur to anyone that combining the strengths of multiple models into one would surely create a more effective model. This is in fact possible and has been done (Caruana *et al.,* 2006). Building on this work, Hinton *et al.* (2015) propose Knowledge Distillation of an ensemble of models into a single model with surprisingly good results on MNIST. Teacher-student is a similar concept where a larger teacher model transfers its knowledge to a smaller, faster student model. Teacher-student technique has been successfully implemented in AVSR by Li *et al.* (2019) where a teacher is trained on a large audio-only data. The student is then trained on a smaller audio-visual data to minimise the Kullback Leibler (KL) divergence between its output and the posterior distribution of the teacher.

Knowledge distillation has also worked well in lip-reading (Zhao *et al.,* 2020; Ma *et al.,* 2021; Ren *et al.,* 2021).

3.3.7. Miscellaneous

Other methodologies and techniques include:

- VSR as dysarthric speech (Howell *et al.,* 2016);
- ASR is all you need (Afouras et al, 2020): training VSR models without requiring human labelled ground truth and instead distil from ASR trained model;
- Dilated CNNs: a technique to increase CNNs receptive field while reducing its computational cost;
- Beyond Lips : an effort to extract maximum information from non lip regions (Zhang *et al.,* 2020);
- Mutual Information Maximisation (Zhao *et al.*, 2020).

3.4. The use of language models in lip-reading

Whether a lip-reading model that uses a separately trained external language model (LM), can be considered an 'E2E' model, is debatable. But for the time being, language models seem to be a convenient way to deal with the alignment problem.

For ALR systems that first classify ASCII characters and then look for their optimal alignment using CTC to build the output sentence, a character-aware language model is required. Given the computational complexity of training ALR systems, a beam search of significant width can grind the process almost to a halt. A language model that is character-aware, lightweight as well as efficient, is required. A potential option can be the work from Kim *et al.* (2016). Their CNN-LSTM based character based LM performs on par with the Penn Treebank state-of-the-art despite being 60% lighter in terms of the number of parameters.
Afouras *et al.* (2018) use the language models (Graves and Jaitly, 2014; Maas *et al.,* 2015) with their LSTM+CTC and full-CNN+CTC with a high degree of success in performance but not as much in terms of computation.

In an attention based sequence-to-sequence lip-reading system, the system learns: i) spatiotemporal feature extraction, ii) a language model and iii) alignment mechanism simultaneously (if trained properly E2E). Since the language model only learns from the text labels of the videos, when a separately trained external language model is used during inference, shallow fusion is used. Shallow fusion has been investigated in detail by Kannan *et al.* (2017) and shown to reduce WER by up to 9.1% on ASR. Their work has also been used by Afouras *et al.* (2018), where the system learns an internal language model from the text in the LRS2 training set while the external language model is trained on its superset, a larger LRS2 pretrain set text.

Although LSTM based character-level language models have been successfully used in SR tasks, both ASR and VSR, it's been shown (Al-Rfou *et al.,* 2019) that a deeper attention-based character-level language model outperforms the LSTM or other RNN based LM variants by significant margins. However, the chase for accuracy with the use of heavier and heavier LMs is probably not the way to go for ALR systems striving for E2E.

3.5. Pre/Non end-to-end approaches in lip-reading

In order to get a better overview of the motivation behind E2E approaches, the following first briefly discusses some pre E2E and non E2E approaches to ALR.

Goldschen (1993) presents a lip-reading system that makes use of **oral-cavity features** where 13, mostly dynamic, such features are introduced. The emphasis is given on the pure optical nature of the system and hence the following information is not used:

- Syntax
- Semantics
- Acoustic
- Contextual

The system has two stages:

- Optical processor
- HMM based decoder

The overall system is summarised in Figure 15. The optical processor first extracts the said features and clusters them to generate 'codewords'. A codebook containing codewords from the whole training data is prepared. PCA is used to minimise the feature dimensions. A sentence recognition rate of 25.3% is achieved on TIMIT dataset.

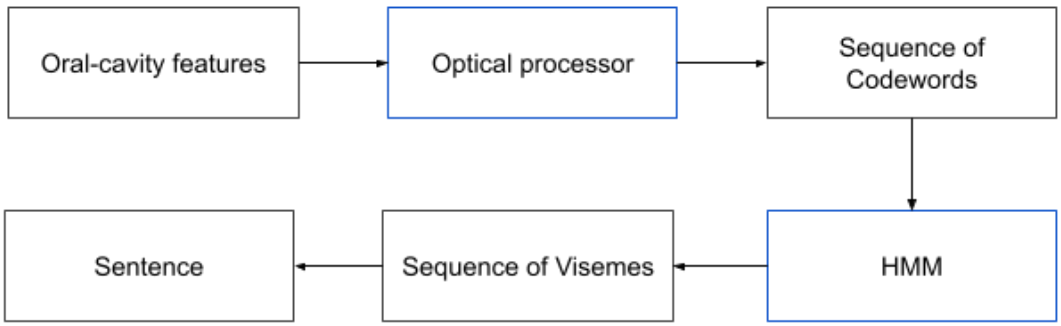


Figure 15: A traditional multistage non-E2E viseme-based ALR system that uses hand-crafted oral-cavity features and codewords sequence Goldschen, (1993).

The concept of **articulatory feature models** used for pronunciation variation in ASR was extended to capture the visually-salient features in VSR by Saenko *et al.* (2005). The model uses SVM feature classifiers to produce inputs to a DBN and compares feature-based and

viseme-based units for inter-feature asynchrony. It was able to outperform models that do not take the asynchrony into account, in limited experiments.

Given the computational overhead due to the redundant information on 'holistic' feature extraction where the whole mouth ROI is used, Lucey *et al.* (2008) attempt a **patch-based analysis** of the ROI in order to find the most informative patches for speech. Their experiment concretises the conventional assumption that central patches of the ROI are the most salient. But the experiment also shows that the holistic features from the entire ROI are more informative than the patches. However, the combination of holistic and patch-based features seemed to improve the performance.

Although most non E2E approaches are traditional, multistage and often pre-deep learning systems, many modern ALR systems have opted for a non E2E approach by design. In an attempt to train a relatively language agnostic lip-reading classifier, Fenghour *et al.* (2020) make use of visemes as classes. The choice of visemes as the classification schema for a sentence level lip-reading has coerced the ALR system to be broken into multiple stages as shown in Figure 16. The Visual Frontend, Viseme Classifier and the Word Detector are separate pretrained NN modules. While in theory, it would not be impossible to train the entire network E2E from scratch, it is unlikely that any significant amount of learning will occur given the long complicated path the gradients flow.

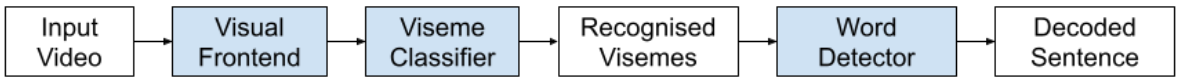


Figure 16. A modern multistage non-E2E viseme based sentence level ALR that first classifies visemes and uses a pretrained viseme-word converter (Fenghour et al., 2020).

The system nevertheless does set a highly improved state-of-the-art WER on LRS2. However, it can be observed that if an ALR system has two-forked aims: language agnosticism (e.g. via viseme usage) and E2E training; a significant rethinking is required for a newer architecture.

3.6. Various end-to-end approaches in lip-reading

This section analyses nine different E2E systems. The systems are investigated for the completeness of their E2E implementation, network architecture, datasets, classification schema and performance. Detailed discussion on their E2E nature is provided in the subsequent sections. Table 4 summarises the comparison. Evaluation of these systems based on the E2E criteria is presented separately in Table 5.

Table 4: *A summary of various E2E lip-reading systems.*

Year	Frontend	Backend	Dataset	Class	Perform-ance	Reference	Is E2E?
2016	FCN	LSTM	GRID	word	WAR 79.6%	Wand *et al.*	E2E
2016	CNN	FCN	LRW	word	WAR 61.1%	Chung *et al.*	E2E
2016	STCNN	BGRU	GRID	ASCII	SAR 95%	Assael *et al.*	E2E*
2017	ResNet	BLSTM	LRW	word	WAR 83%	Stafylakis *et al.*	E2E*
2017	CNN	LSTM	LRS2	ASCII	WAR 23.5%, 49.8%*	Chung *et al.*	E2E, E2E*
2017	FCN + LSTM	BLSTM	OuluVS2	phrase	CA 84.5%	Petridis *et al.*	E2E*
2018	3DCNN + ResNet	BGRU	LRW	word	CA 82% 98%†	Petridis *et al.*	E2E*
2018	3DCNN + ResNet	Transformer	LRS2	ASCII	WAR 50%	Afouras *et al.*	E2E*
2020	3DCNN + ResNet	TCN	LRW	word	WAR 85.3%	Martinez *et al.*	E2E*

* - involves pretraining stages and hence not a purely E2E training scheme
WAR - Word accuracy rate
SAR - Sentence accuracy rate
CA - Classification accuracy for the given class
STCNN - Spatio-Temporal CNN
† - CA of 98% using both audio and video streams. 82% for pure lip-reading (only video)

3.6.1. FCN + LSTM

In this simplistic combination (Wand *et al.*, 2016), feed forwards and LSTM networks are stacked together for a joint E2E training on GRID dataset. Compared to traditional non E2E feature-based approaches viz. Eigenlips and HOGs with SVM classifier, the E2E approach gained an advantage of 11.6% on WAR. The same experiment is also repeated with CNN replacing the FCN but with no improvements on performance. This has been assumed to be due to the small 40x40 mouth ROI containing just enough information.

A major motivation behind the use of LTSM is the inability of SVM to classify sequences. Regardless of the length of the input e.g. a 5 video frame word 'an' and a 10 frame word 'anti'; a fixed vector length is enforced to feed an SVM. LTSMs can take variable length inputs.

Although simplistic, since a unified training approach is used allowing gradients to back-propagate through all the layers of the network, this is a very good example of an E2E deep lip-reading system. However, one downside to a simplistic nature of both the FCN frontend and vanilla LSTM backend as well as the use of a simple lab-generated dataset, is that the performance cannot be reflected in the wild. Nevertheless, it is a good proof of concept for E2E deep ALR.

3.6.2. CNN + FCN

This is the first ALR system to attempt word classification with a large lexicon (LRW). The task of lip-reading is taken as multi-way image classification which shows in the choice of VGG-M (Chatfield *et al.*, 2014) architecture as the base. CNN has been used for its ability to capture spatio-temporal information as evidenced in action recognition. The size and details of the convolution operations can be seen in Figure 17.

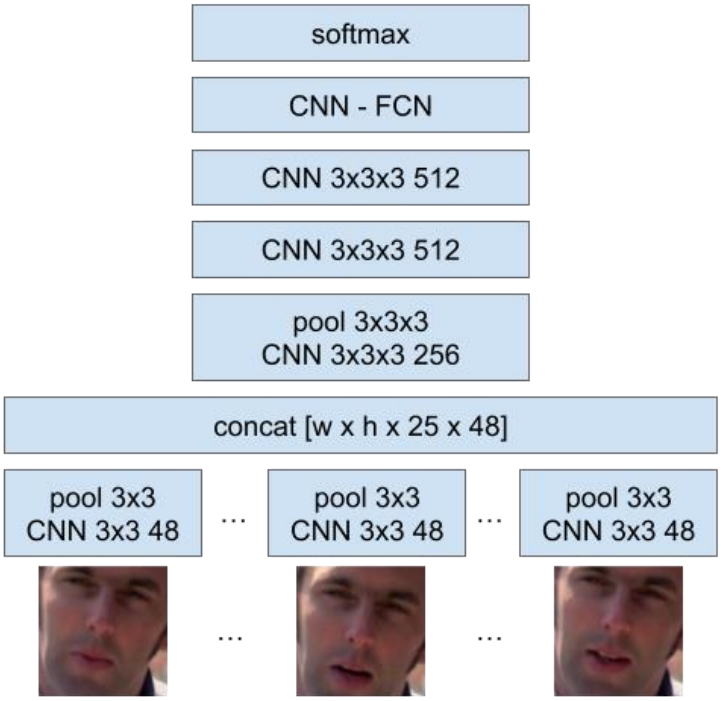


Figure 17: Chung and Zisserman, (2016)'s Multi-Tower architecture

Although not explicitly claimed to be pure E2E this work does train the network E2E. However, the main focus of the work is not on E2E and rather on moving from systems trained on small numbers of utterances in a controlled environment towards a more 'in the wild' data. Moreover, despite the practical limitations of word-level lipreading (discussed in detail in section 4.6.2), the videos in the dataset are made of words that are not uttered in isolation. The frames surrounding the target word contain co-articulation of other words, resembling real world continuous speech. While the ALR system does not use any externally trained ROI detector for a tight registration of the mouth parts, the dataset is already created with the use of such a detector. Not forcing the system to see a narrow ROI does make it robust to variations during inference. But a truly robust E2E system would ideally take videos with a broader view and learn to focus on the speech articulators by itself during training.

3.6.3. STCNN + BGRU LipNet

LipNet is the first 'E2E' sentence level deep lip-reading system. The network structure is illustrated in Figure 18. Image frames from a lip-reading video are input to a spatio-temporal

CNN. The sequence of outputs from this stage are processed by a BGRU. The overall network is trained with the CTC loss at the character level.

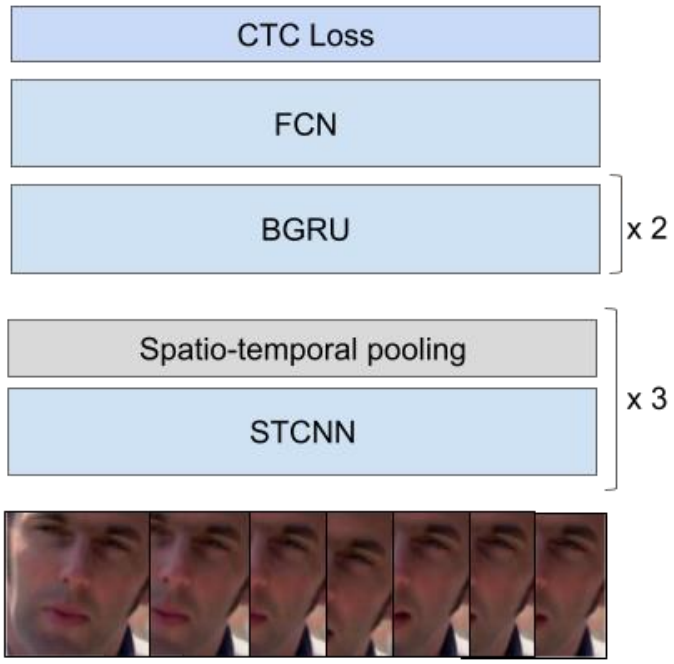


Figure 18: LipNet's STCNN + BGRU architecture block diagram simplified (Assael et al., 2016).

This is one of the major breakthroughs in sentence-level E2E lip-reading. The model emits character level probabilities which are then aligned into a sentence using CTC.

The system is entirely E2E except for the inclusion of a pre-trained face detector during the pre-processing phase and the beam search used on CTC outputs which can be considered a non-E2E post-processing stage. Although the model is able to achieve an Sentence Accuracy Rate (SAR) of 95% on the GRID corpus, it remains to be seen whether the approach can withstand wilder datasets like LRS2 and LRS3. Analysing the generalizability of this E2E architecture on these datasets could be a potential future research topic.

3.6.4. RESNET + BLSTM

This work by *Stafylakis and Tzimiropoulos,* (2017) aims to move away from traditional two-staged ALR to a single stage E2E process. The network, as illustrated in Figure 19, consists of STCNN ResNet combination frontend for feature extraction and a BLSTM backend for sequence modelling. It is trained on LRW dataset and achieves a WAR of 83%, an improvement of 6.8% on the state-of-the-art.

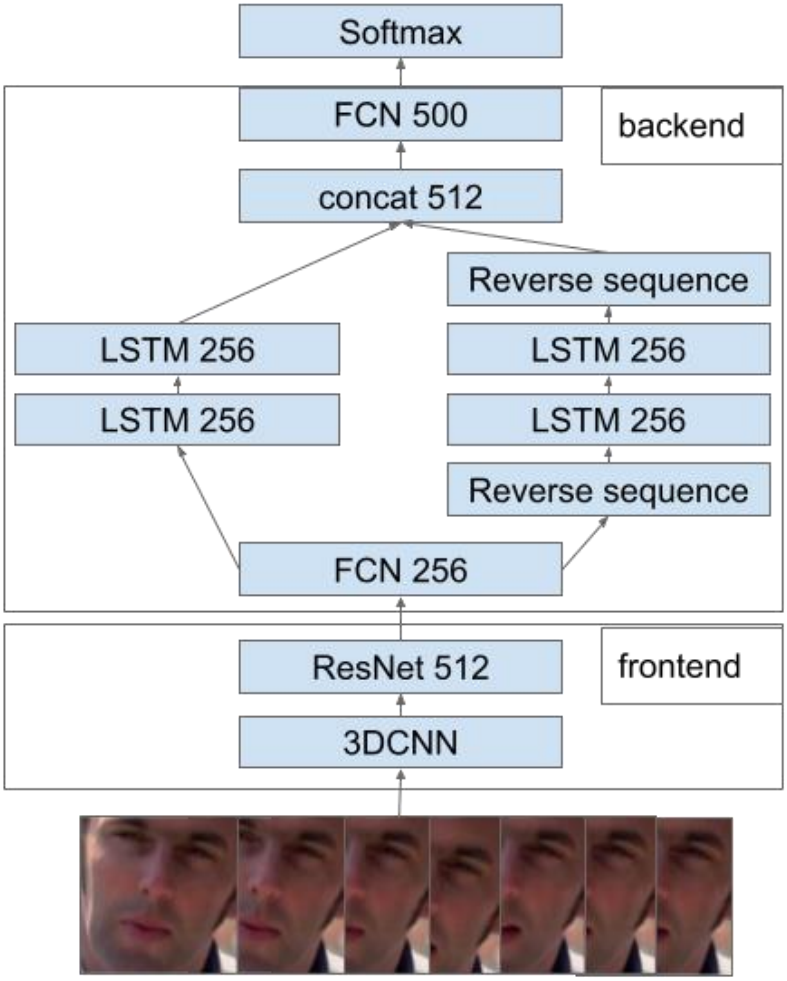


Figure 19: A simplified representation of the ResNet + BLSTM architecture (Stafylakis and Tzimiropoulos, 2017)

Although the model could be trained E2E, given the complexity of the task and the network, a three-stage approach is taken:

i) A TCNN backend is used instead of a BLSTM and trained until convergence

ii) The TCNN backend is replaced with a BLSTM and trained for 5 epochs while freezing the STCNN and ResNet based frontend.

iii) Then the entire network is trained E2E.

The modular training process makes this ALR system diverge away from the E2E ideal. But given the complexity of the architecture, attempting pure E2E training would mean the

gradients will be too weak to train anything significant by the time they propagate back to the frontend. The future of E2E training should aim at solutions to the multi-stage training with simpler but more efficient networks that are better able to learn complex mapping in a single stage process.

3.6.5. CNN + LSTM LRS2

This is another seminal work by Chung *et al.* (2017) in lip-reading, this time with a focus on decoding longer speech sequences: phrases and sentences, from LRS dataset. Experiments are performed with and without the use of audio. A novel multi-module Watch, Listen, Attend and Spell (WLAS) architecture is proposed that operates at character level and is able to learn its own language model to connect them into sentences.

Again, although all modules of the WLAS network are jointly trained E2E, the system first goes through a curriculum learning phase to accelerate learning and to reduce overfitting. This approach has been taken based on the report by Chan *et al.* (2015) which suggests that the longer the input sequence the slower the LSTMs converge, as the decoder finds it quite difficult to extract relevant information from all the input timesteps.

Since curriculum learning introduces a multistage training where sequence length is gradually increased at each stage, whether to consider this experiment a pure E2E can be debated. Also, since the model uses both audio and video input, learning is mostly dominated by the much richer audio information. While this is very helpful in ASR in noisy settings, it is not a pure visual lipreading. When only the visual information is used the Watch, Attend and Spell (WAS) system achieves a WAR of 23.5%. With curriculum learning, scheduled sampling and beam search, the WAR was raised to 49.8%. LSTM uses previous time step ground truth as the next step input during training, while such ground truth is not available during inference. Scheduled sampling uses the previous output at a given sampling rate instead of always using the ground truth. This makes the model better prepared for inference. Also, beam search can be considered a post processing stage that does help improve the performance but makes the system less E2E.

3.6.6. FCN & LSTM + BLSTM

The work, an apparent E2E based on it's title, aims to simultaneously:

i) learn feature extraction from faces;

ii) learn classification using extracted features;

iii) achieve state-of-the-art on 2.

The network as illustrated in Figure 20, consists of two streams: one for the regular mouth ROI image and one from a differential image. In each stream image features are encoded with a FCN and passed on to an LSTM to model temporal dynamics. LSTM outputs from both streams are then fed to a common BLSTM and target classes are predicted using a softmax layer.

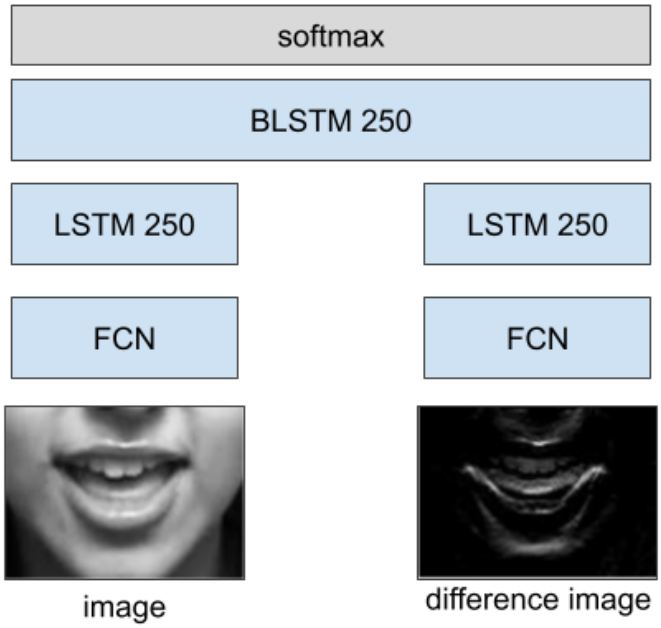


Figure 20: E2E lip-reading with LSTMs using Difference Images (Petridis et al., 2017)

A two-stage training strategy is applied:

i) pre-training the encoding layers with RBM

ii) training the entire network E2E

This approach is reported to help speed up training, but probably does not quite fit the ideal of single-stage E2E training. The system also uses a pre-trained mouth ROI detector (Dlib) before feeding the images frames to the encoding layers. Also the datasets used are quite small and made of extremely simplistic lab-generated videos. The 'difference image'

technique however does seem to contribute towards the 9.7% improvement (phrase recognition rate) on the OuluVS2 dataset baseline. This technique could be incorporated in

other more efficient single stream architectures e.g. 3.6.9. as a second stream as well as with bigger and wilder datasets for a pure E2E experimentation.

3.6.7. 3DCNN & ResNet + BGRU AVSR

Petridis *et al*. (2018) have also applied the E2E BGRU based approach in a fashion similar to Stafylakis *et al*. (2017). However unlike their VSR task, this is an AVSR task and claims to be the first E2E AVSR model to use raw input: raw pixels and raw audio waveform. The use of raw pixels in image classification tasks is not new. However, the using raw audio waveform instead of the conventional Mel Frequency Cepstral Coefficients (MFCC) saves the system one additional processing stage. As discussed earlier, raw inputs to final target output is one of the ideals of an E2E system. In that sense, for an AVSR system, this work stands out as a relatively purer E2E.

However, this system also suffers from the task-network complexity problem and reports poorer performance when trained entirely in E2E. Consequently, training is conducted in multiple stages as follows:

i) Training stream 1 by replacing the backend BGRU with TCNN.
ii) Training stream 2 similarly.
iii) Replacing TCNN backend back with BGRU then freezing streams 1 and 2 to train the BGRU for 5 epochs.
iv) Training the entire network E2E.

The entire network architecture is illustrated in Figure 21, where a ResNet with a 3DCNN input layer processes the video frames in the visual stream while another ResNet processes the waveforms in the audio stream. Feature sequences thus generated are separately processed with BGRU in each stream before being combined into a single stream for classification.

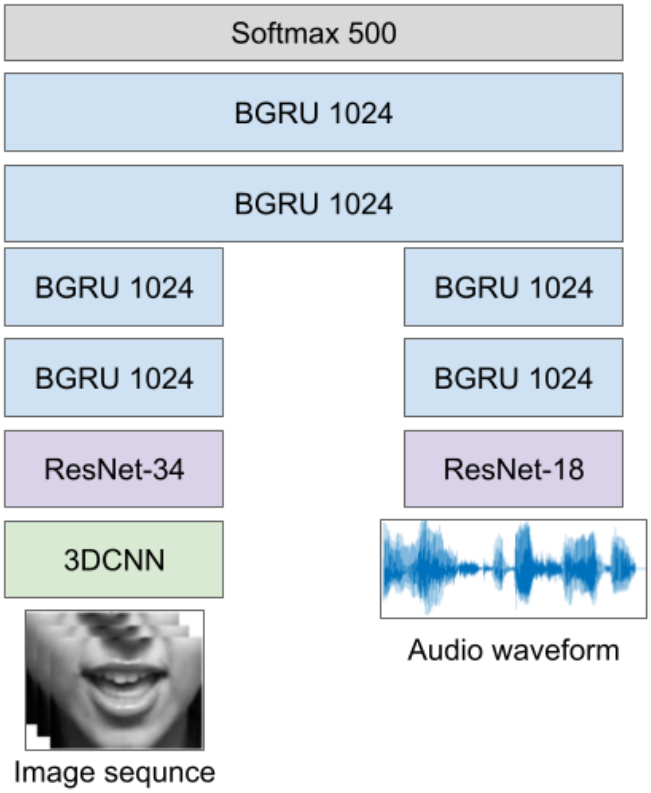


Figure 21: E2E AVSR with BGRU (Petridis et al., 2018)

It can be observed that, the use of RNN-heavy architectures make E2E training less effective due the long distance the gradients need to propagate. The audio stream of the architecture can be considered a more E2E approach as it directly uses raw audio waveforms as input without further conversion. The visual stream could similarly use a raw video of a talking head instead of processed mouth ROI. Following experiments could be performed based on this system:

i) Same network and input but trained entirely E2E.

ii) Raw inputs in both streams with the same network and E2E training.

iii) Experiment 1. and 2., where the BGRUs are replaced by TCNs.

3.6.8. 3DCNN & ResNet + Transformer

This work (Afouras et al, 2018) in fact proposes three NN architectures for lip-reading, which are completely different besides a common frontend/vision module. The vision module is based on (Chung *et al.,* 2017). All architectures are character based. The transformer based model is the best performing among the three. A detailed exploration of the transformer network used in this system is presented in section 3.3.5. Training is conducted as follows:

i) Pretraining the vision module.
ii) Using the trained vision module to generate features from all training data.
iii) Freezing the trained vision module before training the transformer for sequence modelling.

Although the work is highly contributory and sets state-of-the art on LRS2, it does not quite follow the E2E paradigm. To test the effectiveness of this system or its parts in E2E scenarios, following experiments could be run:

i) The same system but trained purely E2E.
ii) Experiment 1., but with visemes, phonemes or words as classes instead of characters.
iii) Experiment 1., with the transformer backend replaced by a TCN.

3.6.9. 3DCNN & ResNet + TCN

For this approach (Martinez *et al*., 2020), a TCN-based architecture based on (Petridis *et al.,* 2018) is proposed where the BGRU backend is replaced with a TCN. The conventional 3-stage cumbersome training scheme is also simplified into a single stage using a Cosine Scheduler (Loshchilov and Hutter, 2017). Their experiment demonstrates that E2E single-stage training from scratch is not only feasible, but also can produce state-of-the-art results. This is evidenced by the models gain of 1.2% and 3.2% on LRW and LRW-1000 WAR state-of-the-art respectively. The superiority of the model mainly lies in its ability to train quickly. Compared to the 3-week training time of the 3-stage BGRU based models, this model can train in 1 week with on-par performance. Since this system forms the basis of our experiments, its details are elaborated in section 4 and the reasons for selecting this system are presented in section 3.6.10.

3.6.10. Comparative evaluation of the E2E systems

All the 9 systems in 3.6.1. - 3.6.9. are compared in Table 5 based on how they meet our E2E criteria set in section 2.3.

Architecture notations:

i) FCN + LSTM = a1
ii) CNN + FCN = a2
iii) STCNN + BGRU = a3
iv) ResNet + BLSTM = a4
v) CNN + LSTM = a5
vi) FCN & LSTM + BLSTM = a6

vii) 3DCNN & ResNet + BGRU = a7 viii) 3DCNN & ResNet + Transformer = a8

ix) 3DCNN & ResNet + TCN = a9

Table 5: *Evaluation of various E2E ALR systems.*

		Words						Sentences			
Criteria for pure E2E based on 2.4.		**a1**	**a2**	**a4**	**a6**	**a7**	**a9**	**a3**	**a5**	**a8**	**Comments**
1	All modules are differentiable	1	1	1	1	1	**1**	1	1	1	
2	Gradients flow from end to end	1	1	1	1	1	**1**	1	1	0	a8 is multistage
3	Has no modular training	1	1	0	0	0	**1**	1	1	0	
4	Has no pretraining stage	1	1	1	1	1	**0**	1	0	0	a5,a8 use curriculum. a9 pretrains on difficult words
5	Has no trained processing module	0	0	0	0	0	**0**	0	0	1	e.g. Dlib for ROI
6	Learns its own language model				0			0	1	1	n/a to word level
	Other evaluation criteria										
7	Uses wild data	0	1	1	0	1	**1**	0	1	1	
8	Lip-reads sentence level	0	0	0	0	0	**0**	1	1	1	
9	Possible to train E2E	1	1	1	1	1	**1**	1	1	0	
10	Training speed.	0	0	0	0	0	**1**	0	0	0	a9 trains relatively significantly faster.
	Overall score	5	6	5	4	5	**6**	6	7	5	

Table 5 evaluates 9 different E2E ALR systems based on 10 different criteria. A score of 1 is given if an architecture meets the criteria, 0 if it does not and blank if the criteria does not apply. Within the constraints of our definitions of an E2E lip-reading system, the ALR system 3.6.9. (a9) seems to tick the most boxes and hence can be considered a great example of such. The following are the reasons the experiments of this project are based on this system:

- The project currently aims to run word-level experiments and 'a9' is word-level.
- It trains much faster (80 epochs in a week) than other similarly scored word-level systems.
- It uses the efficient and currently trending TCNs for sequence processing.
- It is simple to run the system entirely in E2E with minimal changes (i.e. no pre-training on difficult words).
- The system is able to set a new baseline on LRW dataset despite being an E2E.

4. Proposed experiments

This chapter lays out the details of the experiments performed to test the effectiveness of pure E2E ALR. Various E2E systems were analysed and compared for this purpose (section 3). The E2E experiments are based on the architecture by Martinez *et al.* (2020) (section 3.6.9.). The model is retrained for a few epochs and makes use of an additional dataset. Some experimental settings are tweaked in order to test the E2E concept. It should be reiterated that: designing a new architecture or setting a state-of-the-art is not within the current scope of this study but could form the aims for future research.

4.1. Rationale

E2E systems are one of the ideals of machine learning systems and AI in general. With decades of progress and hundreds of great works in the field of deep lip-reading, good E2E systems are still rare to find. One major hesitation seems to come from the push towards setting a state-of-the-art performance. This is evidenced by several great architectures, with a potential to train E2E, defaulting to multi-stage training with the sole aim of achieving a percentage gain. While this is by no means a censure to the convention of setting new records on benchmarks, as such records are a prime movers of the frontier, the fact remains that yet another prime objective of AI, i.e., E2E is pushed a bit into the shadows. It could probably be claimed that, had the primary drive been towards an E2E system and its perfection, records on benchmarks would surely have followed along. With this in mind, and the details covered in the coming sections, the proposed experiment has the following objectives/characteristics:

i) Training a pure E2E deep lip-reading system.
ii) Word based classification schema for simplicity.
iii) Use of new in-the-wild dataset LRS3 to compare the performance of the original E2E model.
iv) No pretraining stage in order to stay within the definition of pure E2E training.

4.2. Architecture

Figure 22 shows a TCN with 4 layers with a receptive field of 16. Receptive field is a hyperparameter that needs to be set apriori. Since the input length is equal to the output length in a TCN, the receptive field is determined by the number of layers and the kernel size/dilation factor.

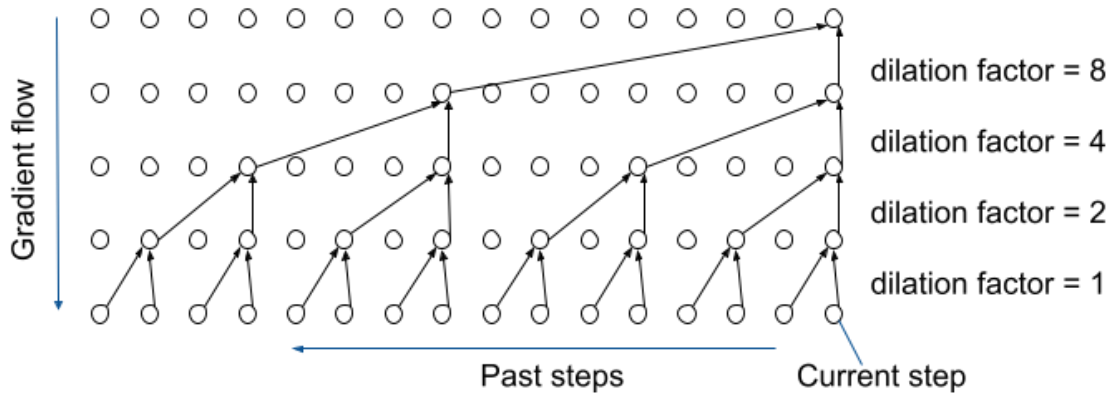


Figure 22: A 4 layered TCN

For the network in Figure 22:

$$num\ layers\ (l)\ =\ 4$$
$$kernel\ size\ (k)\ =\ 2$$
$$receptive\ field\ (R)\ =\ 2^{l}(k\ -1) \tag{10}$$
$$R\ =\ 2^{4}(2-1)\ =\ 16$$

Thus one advantage of using TCNs is the convenience with which a desired receptive field can be set based on the task. The example shows a *causal* TCN where the output is determined based solely on the past input steps. TCNs can also be designed to be *non-causal* by simply allowing each output to look at future timesteps when available and required. In Martinez *et al*., (2020), the *non-causal* variant is used since the whole input sequence is known beforehand. The network consists of multiple blocks where the stride size is calculated based on the block index.

$$stride\ (s)\ =\ 2^{i-1} \tag{11}$$
$$where, i\ =\ Block\ index$$

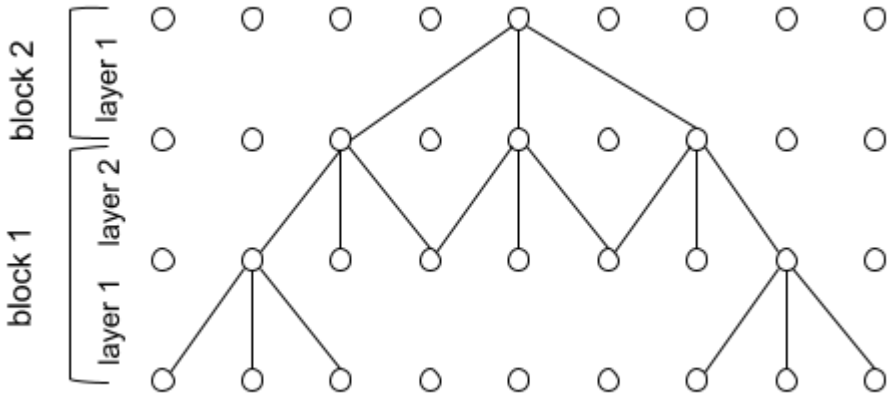


Figure 23: TCN network used by Martinez et al., (2020)

The experiment implements one of the several proposed network variants viz. the Multiscale TCN (MS-TCN) as shown in Figure 24. Since the activations at any given layer have the same temporal receptive field, in order to enable the network to see temporal information of different lengths, MS-TCN implements 3 different TCNs each with a different kernel size. The number of kernels is made to be a function of the number of TCN branches used.

$$Given, Channel\ dimensionality\ =\ C$$
$$Number\ of\ TCN\ branches\ =\ n$$
$$number\ of\ kernels\ =\ k$$
$$k\ =\ \frac{C}{n} \tag{12}$$

This is illustrated in Figure 23, where n = 3.

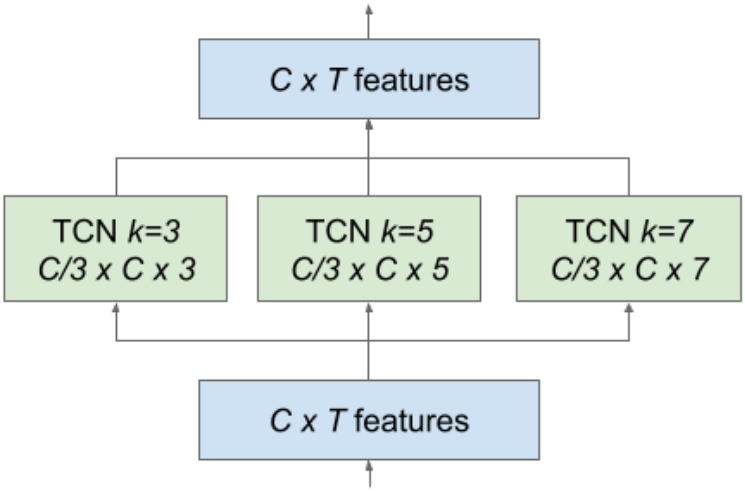


Figure 24: Multiscale TCN with 3 TCN streams (Martinez et al., 2020)

The overall architecture is illustrated in Figure 25. It consists of 3DCNN and ResNet-18 combination as the frontend to extract spatiotemporal features, which are then fed to a MS-TCN to model the sequence. Finally a softmax layer is used with cross-entropy loss.

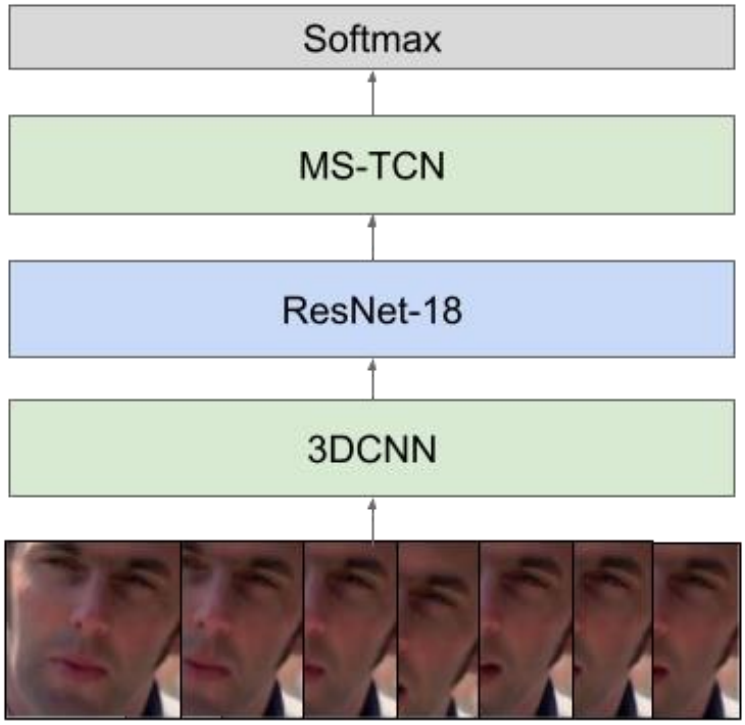


Figure 25: TCN based lip-reading model (Martinez et al., 2020)

4.3. Datasets

The primary dataset in our experiments is Lip-reading Sentences in the Wild 3 (LRS3-TED), a collection of sentence clips from TED videos on YouTube. The LRW dataset is also used for retraining the model from scratch. Table 6 and Table 7 summarise the main characteristics of the two datasets.

Table 6: *LRW dataset statistics. The dataset is word based and contains around 1000 videos (samples) per word in the training set. There are 500 classes (unique words), i.e. the vocabulary size = 500.*

Set	Classes (words)	Samples	Samples per Class
Train	500	478764	800 - 1000
Validation	500	25000	50
Test	500	25000	50

Table 7: *LRS3-TED dataset statistics. The dataset is created by trimming sentence clips from full TED videos. Each full video includes one speaker. E.g., there are 5090 speakers in the pretrain set.*

Set	Full Videos	Sentence clips	Words	Vocabulary
Pretrain	5090	118,516	3,900,000	51,000
Train-Val	4004	31,982	358,000	17,000
Test	412	1,321	10,000	2,000

Further details of the two datasets are included in appendix A.: Datasets.

Table 8: *Statistics of the* ***newly created*** *word-level dataset from LRS3-TED (Afouras et al, 2018).*

Set	Classes (words)	Samples	Samples per class
Train	500	142817	134 - 2241
Val	500	56454	36 - 894
Test	500	6022	3 - 36

Table 8 shows the 500 words shortlisted version of the whole word-based conversion. The sample frequency (number of samples per class) varies for different words. E.g. for the training set, some words have 134 videos while some have as many as 2241 videos.

Conversion of sentence-based LRS3 dataset into word-based:

Since the experiments are word-based, the sentence-level LRS3 dataset first needs to be adapted as such, akin to LRW dataset. Figure 26 details the dataset conversion and the pre-processing stages. As it can be seen from tables 7 and 8, there's a lot of difference in both the structure, as well as the content of files between the two datasets. Two new word-level datasets are created from the sentence-level LRS3 dataset:

i) A smaller word-wise split containing only 30% of LRS3 pretrain set (Trainval set is not used as it is a subset of pretrain set.).

ii) A full size set using 100% of the LRS3 pre-train set.

The pretrain set is chosen for the creation of the new word-level datasets because it contains the timestamps for each word in the videos.

The LRS3, hence also the word-level dataset created from it, are much more challenging compared to the LRW dataset:

- Although LRW is a 'wild' dataset, it has been curated with special focus on maintaining an almost equal number of samples for each word. On the other hand, LRS3 is wilder still, and contains sentences of variable lengths with no special regards to word choice. Hence, when sentence videos from LRS3 are clipped by word to create word videos, the sample frequency of each word is as wild/natural as is normally spoken (TED talks in this case). E.g. The word 'about' appears 1277 times while the word 'abilities' appears only 149 times in the videos. This is a more natural representation of spoken English in the real world. However, severe imbalance in class frequency is avoided as that will require specialised techniques to achieve good classification results for minority classes. This is achieved by limiting the number of words to 500, equal to LRW. Only 500 words with the most numerous samples are selected from over 39,000. Hence, despite some difference in the number of samples for each class, there is not any particularly minor class.
- Another challenge that is uncovered during the conversion is the fact that a lot of words do not get enough utterance duration and hence do not have enough number of frames for the lip-reading model. The model expects a minimum of 5 frames as its 3D convolution kernel is 5 x 7 x 7. If variable length temporal augmentation of the frames is allowed during runtime, i.e. random removal of a few frames to improve model robustness to 'jitter', even more frames are required. This issue is resolved by yet more filtering: by minimum duration of 0.5s during the creation of videos and by minimum number of frames as 10. This is in stark contrast to a more organised LRW where each video has 29 frames with the target word lying roughly in the middle frames.
- Speaker variability and other intra-class differences also seem to be a lot greater in LRS3 compared to LRW from manual comparison of random samples of clips from the two datasets. Head-pose and movement being the most common. While this variability helps build a more robust model, it also means a lot more data is required for training.

4.4. Pre-processing

Martinez *et al*., (2020) have taken precomputed face landmarks set for LRW to generate mouth ROIs. Further transformation and augmentation happens during the training run.

Since the word-split version of LRS3 is used, the pipeline as shown in Figure 26, is created for the purpose. Runtime transformations and/or augmentations are the same as in the original work.

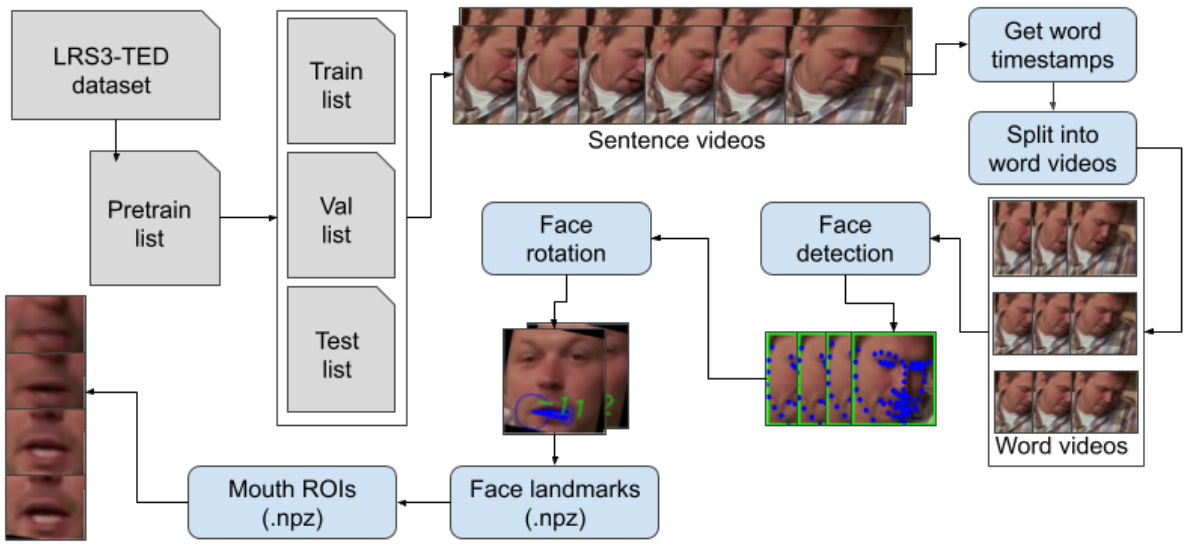


Figure 26: Dataset conversion (LRS3 to word-level) and pre-processing pipeline.

Face rotation:

Dlib's 68 face landmarks are used to rotate an angled face back to vertical as shown in Figure 27.

Lip corners: $A = (x_1, y_1)$ and $B = (x_2, y_2)$ obtained from landmarks 48 and 54.

A horizontal drawn through *A* and a vertical line drawn through *B* intersect perpendicularly at $C = (x_2, y_1)$.

The correction angle to rotate θ, is calculated as:

$$tan\theta = \frac{BC}{AC}$$

$$\theta = tan^{-1}\left(\frac{(y_1 - y_2)}{(x_2 - x_1)}\right) \quad (13)$$

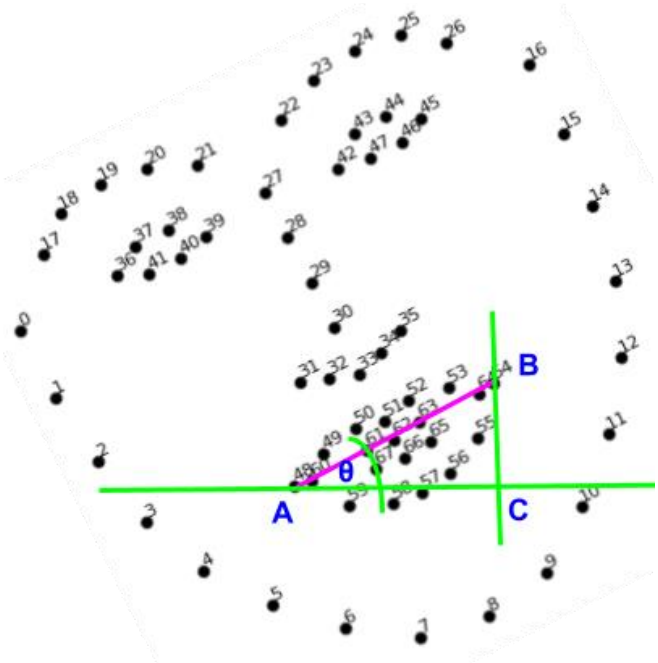


Figure 27: Face rotation using Dlib landmarks. See appendix for code snippet.

4.5. Experiments

The following experiments were planned towards E2E training as well as testing the robustness of the pretrained models (Martinez *et al.*, 2020; Assael *et al.*, 2016; Afouras *et al.*, 2018) on newer data. While some of the planned experiments were completed, some of the experiments, being outside the current scope of this study, will form the topics of our continued research in the domain.

Experiments conducted:

TCN based word-level E2E model (Martinez *et al.*, 2020):

i) Training a model for 7 epochs on LRW E2E with no pretraining on the 'difficult' words. While the pretraining seems to make a slight improvement in the performance, as reported in the original work, it does not meet the 'single, unified training' of an ideal E2E approach. The training was stopped after 7 epochs as the system behaved predictably (which can be seen in the learning curves Figure 28 and the test accuracy in Table 9 and there was no further point to make or record to set.

ii) Testing the new model on:

a) LRW test set;

b) LRS3-Word test set.

Experiments planned for the future:

Follow up experiments to further consolidate the effectiveness of the E2E approach are listed in section 6 : 'Recommendations for future work'.

4.6. Results and Discussion

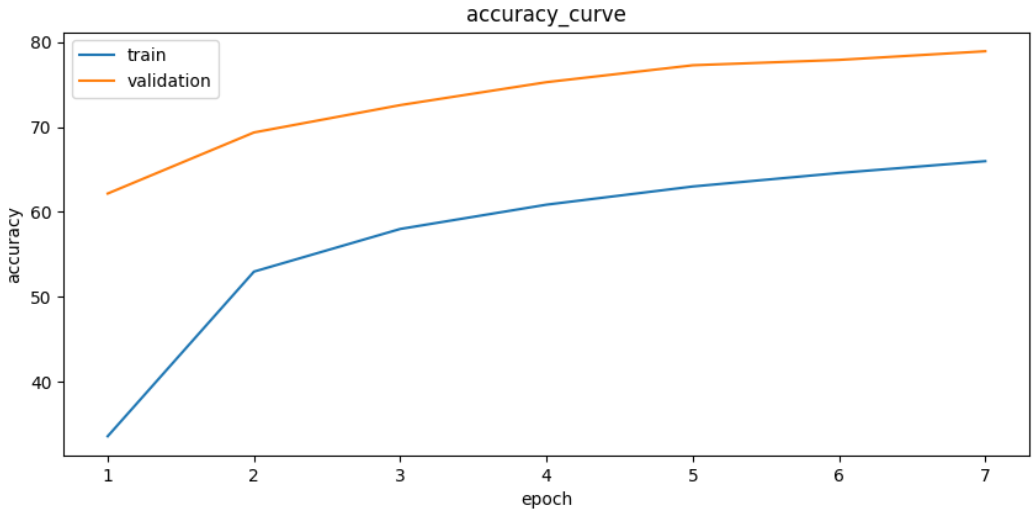
4.6.1. Performance

Table 9. *Experiments and results (* - fully pretrained models)*

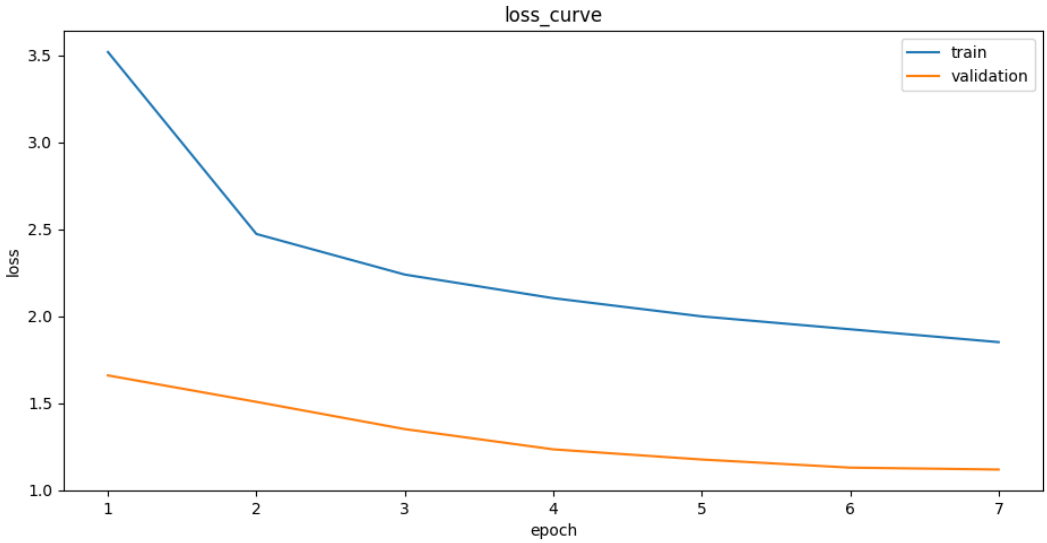
Experiment	Dataset	Epochs	Mode	Loss	WAR	Architecture
1	LRW	7	train	1.8506	65.98%	resnet_18_mstcn
1.1.	LRW		validation	1.1184	78.92%	resnet_18_mstcn
2	LRW		test	1.1299	78.57%	resnet_18_mstcn
3	LRS3		test	10.8272	0.17%	resnet_18_mstcn
4	LRW	80*	test	0.4760	87.94%	resnet_18_mstcn
5	LRS3	80*	test	9.1031	0.20%	resnet_18_mstcn

i) Resnet_18_mstcn is the best performing of several models from (Martinez *et al.*, 2020).

Figure 28 shows the training and validation losses and accuracies of the system over 7 epochs of training.



a)



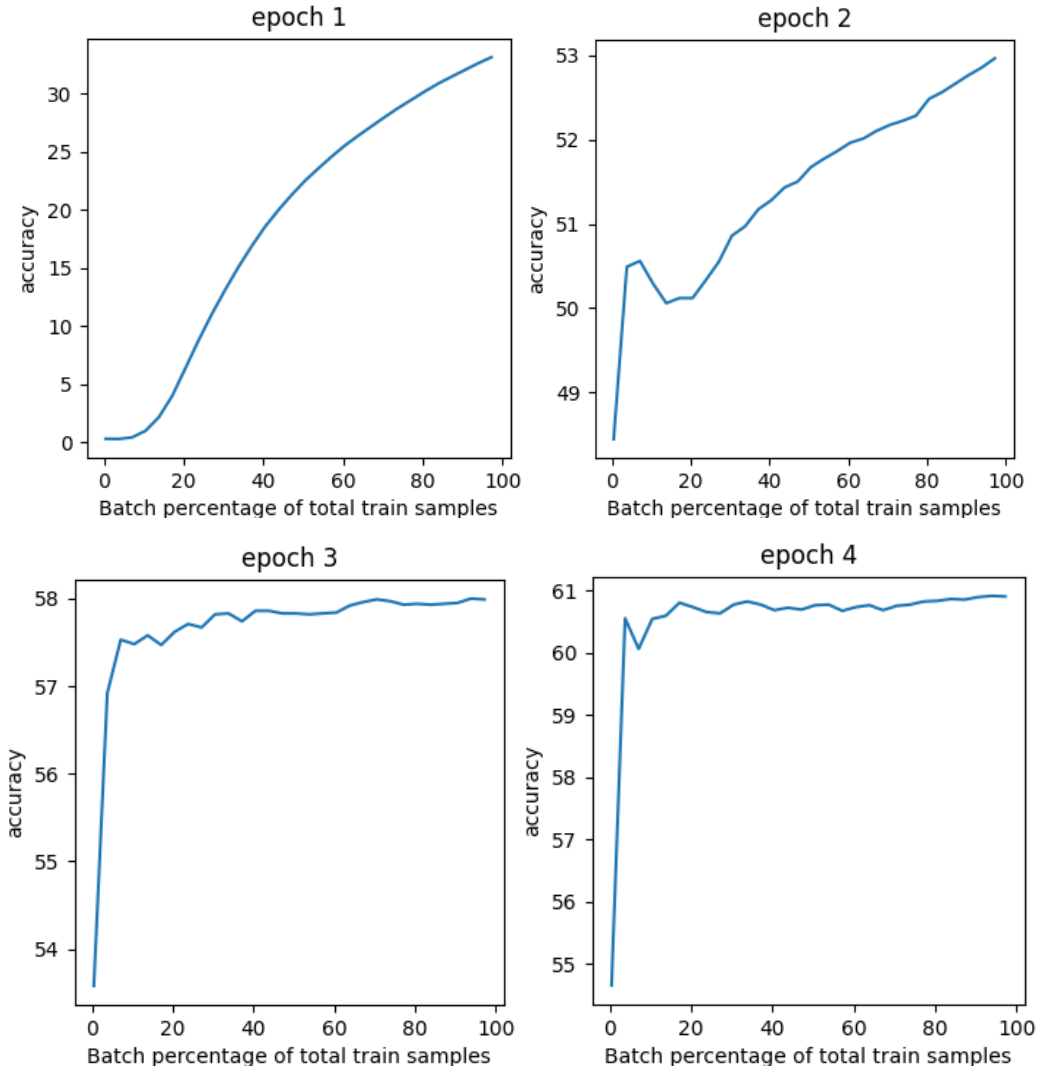
b)

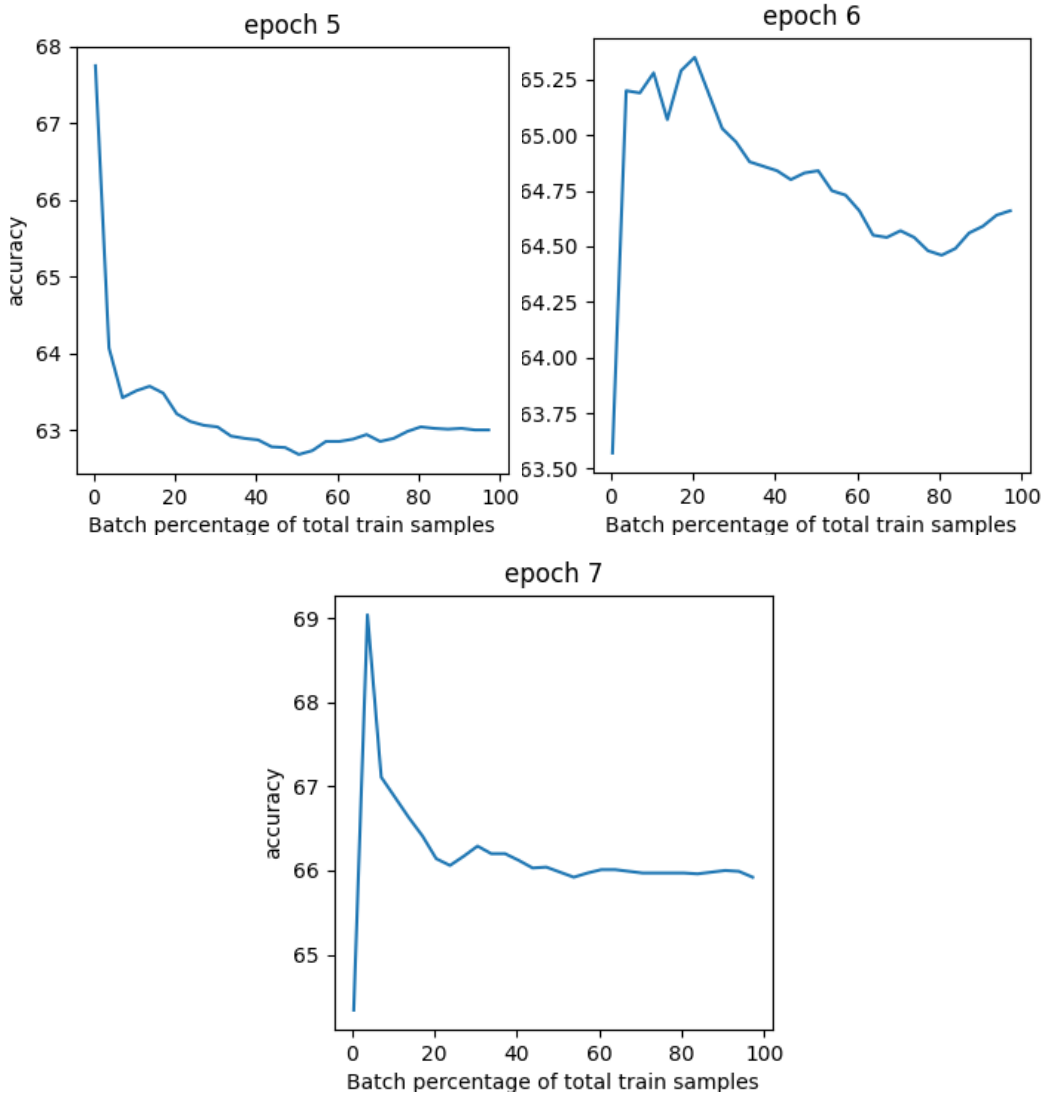
Figure 28. Accuracy (a) and loss (b) curves of the E2E system over 7 epochs

The learning curves in Figure 28, are not presented to suggest that the model was sufficiently trained in only **7 epochs**. Training until convergence is not the aim of this study as the original system is used almost as is and hence there's no baseline to be improved. Since the system is almost unchanged compared to the original, except for the exclusion of pretraining on difficult words, the purpose of the training was to confirm that the system trains predictably. The pretraining stage was excluded as it is the only non-E2E aspect of the system as seen in Table 5 (besides the use of a pre-trained pre-processor). The predictability

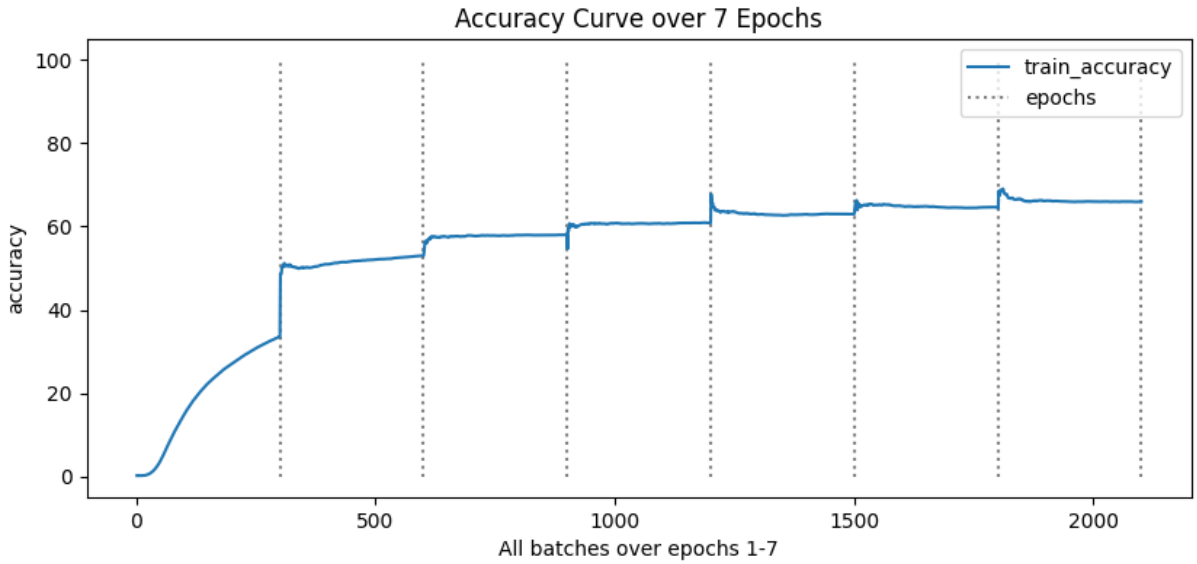
of the training can be already seen from 7 epochs and thus a considerable amount of time was saved.

Since the model was only trained for 7 epochs, the plots in Figures 29 illustrate batch by batch progress in training accuracy in %. The first batch size 32 and the remaining updates are logged after every 1600 samples. The plots show batch/accuracy both in percentage.




epoch 5
epoch 6
epoch 7
accuracy
Batch percentage of total train samples

a)


Accuracy Curve over 7 Epochs
train_accuracy
epochs
accuracy
All batches over epochs 1-7

b)

Figure 29: a) Accuracy over batch progress for each of the 7 epochs of training. b) Combined view of training accuracy over 7 epochs.

The decision for num_epochs = 7 was taken based on the following factors: the time taken for training (approx. 36 hours for 7 epochs) as well as the tapering off of the accuracy as seen in the plots. This is not to suggest that the training further e.g. all 80 epochs would not have improved the accuracy. However, that would have taken in excess of 21 days per experiment. Besides, it would be outside our objectives, given the current scope.

ii) The test WAR of 78.57 % over the training accuracy 65.98% needs further investigation. One possible explanation is the fact that most test videos in LRW are handpicked for quality.

iii) When tested on 500 label word-version of LRS3, the pretrained model (80 epochs) and the 7 epochs model achieves a very low WAR of 1.7% and 2% respectively. The primary reasons for such a low performance has been attributed to:

   a) The extreme variability in video frame numbers compared to the constant 29 frames in LRW (despite the use of variable length training of the original model).
   b) Unseen classes, as not all words in the LRW and LRS3-Word videos are the same.
   c) The quality of the videos in LRS3-Word is often significantly lower than that in LRW.

However, the test accuracy (WAR) of a fully trained model on the original LRW dataset does attain 87.9% as mentioned in the literature.

4.6.2. Limitations of using words as classes

Using words as class labels simplifies the networks, gets rid of the complexity of alignment and learning/use of language models. For such a network, a word is just a label and has no linguistic value that needs to be learned. A thus simplified network is relatively easier to train E2E. While this helps our push towards E2E, it is of little practical value.

For example, the TCN model (Martinez *et al*., 2020) uses the word labelled LRW dataset. Despite setting the state-of-the-art on the dataset and being able to train the model almost purely E2E, the model can only distinguish between 500 English words. A vocabulary of size 500 is not nearly enough for any significant NLP task. This is not saying that the model could not be trained on a dataset with more words. It could simply be trained with the same procedure and settings on more words. But each new word would require enough samples in the dataset i.e. around 1000 videos per word by LRW scale. Thus the first problem in building a full English vocabulary word classifying ALR is creating the dataset. Such a dataset would have to have over 171,000,000 labelled lip-reading videos compared to around 500,000 in the LRW dataset i.e. around 240 times larger than the LRW dataset. LRW is already quite a large dataset in the current scale. The second problem with a full vocabulary word based ALR is that, even if a complete dataset is laboriously created, the ALR trained on such a dataset would still only be able to identify individual words. This is not very useful for continuous speech tasks.

One possible solution would be to break down a continuous speech video into word chunks before feeding into a word-based ALR and concatenate the outputs. But the system would have no way to learn the contextual relationship between words. That would have to be done as a part of post processing e.g. via the use of a language model. This would make the system overly complicated and also non E2E.

Hence, the focus of lip-reading research should be towards building E2E sentence level ALR systems. This will be one of the main topics of our future research.

5. Reflection

5.1. A review of the research aims and objectives.

The level at which this study has been able to achieve the aims and objectives set out in the beginning of this study (section 1.1.), are as follows:

i) The study has successfully explored the concept of E2E in general and defined its meaning in the context of deep learning and deep lip-reading.

ii) Deep ALR has been successfully introduced with its meaning, applications, challenges and solutions.

iii) Various methods, tools and techniques in the related domains have been thoroughly reviewed for their significance in ALR within the context of E2E approach.

iv) Various E2E ALR systems have been reviewed and comparatively evaluated based on the set criteria for E2E.

v) Pure E2E experiments have been carried out with some level of success in justifying the practicability of E2E ALR.

5.2. Answering the research questions.

This section re-visits the proposed research questions (section 1.2.) and discusses to what extent the study and the experiments were able to answer them.

i) What is the ideal definition of E2E in deep lip-reading?

The criteria for pure E2E lip-reading was defined (section 2.4) based on differentiability of the modules of an architecture, continuous flow of gradients, unified training process, level of pre-processing and postprocessing.

ii) How are the breakthrough methods, tools and techniques in deep learning, especially in speech and image processing, aiding the E2E approach?

Through a review of such methods, tools and techniques (section 3.3), the impact of their advances in the more general fields of speech and image processing was found to be very significant for the progress in ALR.

iii) Do the seemingly E2E state-of-the-art works in lip-reading meet the full extent of E2E as defined in this study?

It was discovered from a review of several different E2E systems (section 3.6) that most of the systems are not completely E2E when judged based on the newly set criteria for E2E.

iv) Is the quest for pure E2E deep lip-reading pragmatic or idealistic?

The full E2E experiment and its performance on the test set of a fairly wild dataset proves that at least for word-level ALR systems, pure E2E is a pragmatic approach. However, the generalisability of the E2E model is questionable as its performance seems to falter on an unseen dataset. The performance lag on new data could very well be a general issue in ML and has very little do with the E2E approach. This needs to be clarified with further controlled experiments. Also, since the experiments in this study are word-based, whether the pure E2E approach is pragmatic for sentence-level ALR remains to be verified.

Note: The confusion matrices of the word classes are provided in Appendix B. Due to the large number of classes (500, one for each word label), the matrices had to be broken down into chunks for better visibility. Appendix C provides the top-5 predictions of the system for all 500 word labels.

6. Conclusion

The meaning and purpose of the E2E concept is outlined in the context of machine learning in general and in lip-reading specifically. A survey of the E2E approaches to lip-reading is provided along with an overview of pre/non E2E techniques that directly or indirectly led to the development of recent E2E trends in the field. Upon individual analysis of several E2E approaches, it is discovered that most works still incorporate non-E2E methods like pretraining and multistage training in order to improve performance. However, the most recent among such approaches seem to be getting ever so close to the complete definition of E2E. Most of the objectives of this study have been achieved and its questions answered. The experiments conducted suggest that although E2E approaches have evolved enough to set new state-of-the-art, the models still need to be a lot more robust to newer data in order to be applicable in the real world. Future work will include the completion of all the planned experiments for a more thorough analysis of a few chosen works. Another future work will include incorporating the information obtained through this survey into developing a purer novel E2E lip-reading architecture.

Recommendations for future work:

The project was able to achieve most of its outlined objectives and answer most of the research questions as reflected in section 5. However, to obtain a more complete understanding of the E2E in the full spectrum of lip-reading, following experiments are planned for the future as a follow up to the experiments conducted:

i) TCN based word-level E2E
   a) Full training and test of a new model on the new LRS3-Word dataset.

ii) BGRU based E2E sentence-level model (Assael *et al.,* 2016):
   a) Train purely E2E on LRS2 vs the original, much easier GRID corpus
   b) Test on LRS2, LRS3 for generalisation capability

iii) Transformer based sentence-level model (Afouras et al, 2018):
   a) Train on original LRS2 but E2E instead of original multistage pretraining approach.
   b) Investigate techniques/modifications that could facilitate E2E learning.

Besides these experiments with specific goals, the following studies with broader objectives in the topic are recommended for future work:

iv) Measuring the impact of homophemes in E2E lip-reading: This can be done by comparing the performance of E2E ALR systems on homophemic and non-homophemic videos. A dataset can be split into several versions, each with varying percentages of homophemes. Various ALR systems can then be trained and tested on each of these versions for performance comparison.

v) Application of the latest advances in neural network research to create a novel, pure E2E lip-reading architecture. The tools, tips and techniques discussed throughout this study can be brought together where possible to combine their strengths into a single architecture.

vi) Extending the investigation to other classification schemas. Design experiments to test the effectiveness of the E2E approach on ALR systems that are based on visemes, phonemes, characters etc.

vii) Testing the effectiveness of the pure E2E approach in a more challenging sentence-level ALR system. Given the limitations of word-level lip-reading (section 4.6.2.) a move towards the more useful sentence-level E2E lip-reading is required.

It is our firm belief that these works will help clear some obstacles and answer some questions in the path towards the development of an end-to-end lip-reading system that is more accurate and can perform sentence-level predictions. Such a system will become a significant milestone towards the higher goal of the domain, i.e. a reliable and practical ALR system capable of continuous decoding. An investigation into homopheme, which is one of the biggest obstacles in lip-reading, can provide us with insights into improving the accuracy. A study of sentence-level lip-reading with various classification schemas and more efficient networks can pave the path for a more continuous lip-reading.

References

Afouras, T., Chung, J.S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. *ArXiv*, abs/1809.00496.

Afouras, T., Chung, J.S. and Zisserman, A. (2018). Deep Lip Reading: a comparison of models and an online application, *ArXiv*, abs/1806.06053.

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

Al-Rfou, R., Choe, D., Constant, N., Guo, M. and Jones, L. (2019). Character-Level Language Modeling with Deeper Self-Attention. AAAI.

Anina, I., Zhou, Z., Zhao, G. and Pietikäinen, M. (2015). OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis, *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-5, doi: 10.1109/FG.2015.7163155.

Assael, Y., Shillingford, B., Whiteson, S. and Freitas, N.D. (2016). LipNet: End-to-end Sentence-level Lipreading, *arXiv*: Learning.

Bahdanau, D., Cho, K. and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Bai, S., Kolter, J.Z. and Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv*, abs/1803.01271.

Bastanfard, A., Aghaahmadi, M., Abdi, A., Fazel, M. and Moghadam, M. (2009). Persian Viseme Classification for Developing Visual Speech Training Application. 1080-1085. 10.1007/978-3-642-10467-1_104.

Basu, S., Neti, C., Rajput N. et al. (1999). Audio-visual large vocabulary continuous speech recognition in the broadcast domain, *IEEE Third Workshop on Multimedia Signal Processing (Cat. No.99TH8451)*, pp. 475-481, doi: 10.1109/MMSP.1999.793893.

Bear, H.L. and Harvey, R. (2016). Decoding visemes: Improving machine lip-reading, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009-2013.

Bregler, C. and Konig, Y. (1994). Eigenlips for robust speech recognition, *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. II/669-II/672 vol.2, doi: 10.1109/ICASSP.1994.389567.

Burton, J., Frank, D., Saleh, M., Navab, N. and Bear, H.L. (2018). The speaker-independent lipreading play-off; a survey of lipreading machines. *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 125-130.

Cappelletta, L. and Harte, N. (2011). Viseme definitions comparison for visual-only speech recognition, *19th European Signal Processing Conference*, pp. 2109-2113.

Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O. (2015). Listen, attend and spell. *arXiv preprint* arXiv:1508.01211.

Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets, *In: Proc. BMVC*.

Chen, H., Du, J., Hu, Y., Dai, L., Lee, C., & Yin, B. (2020). Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention. *ArXiv*, abs/2012.14360.

Chen, X., Du, J. and Zhang, H. (2020). Lipreading with DenseNet and resBi-LSTM, *Signal, Image and Video Processing*, 14, 981-989.

Cheng, S., Ma, P., Tzimiropoulos, G., Petridis, S., Bulat, A., Shen, J. and Pantic, M. (2020). Towards Pose-Invariant Lip-Reading, *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4357-4361.

Cho, K., Merrienboer, B.V., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. SSST@EMNLP.

Chung, J.S., and Zisserman, A. (2016). Lip Reading in the Wild. *ACCV*.

Chung, J. S., Senior, A., Vinyals, O. and Zisserman, A. (2017). Lip Reading Sentences in the Wild, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444-3453.

Cooke M., Barker J., Cunningham S. and Shao X. (2006). An audio-visual corpus for speech perception and automatic speech recognition, *J Acoust Soc Am.*, 120(5 Pt 1):2421-4. doi: 10.1121/1.2229005. PMID: 17139705.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

Davis E. King. (2009). Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* 10 (12/1/2009), 1755–1758.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I. and Zafeiriou, S. (2019). RetinaFace: Single-stage Dense Face Localisation in the Wild. *ArXiv*, abs/1905.00641.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition, *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141-151, Sept. 2000, doi: 10.1109/6046.865479.

El-Bialy, R., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O. and Li, B. (2022). Developing Phoneme-based Lip-reading Sentences System for Silent Speech Recognition.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K.W., Hassidim, A., Freeman, W.T. and Rubinstein, M. (2018). Looking to listen at the cocktail party, *ACM Transactions on Graphics (TOG)*, 37, 1 - 11.

Feichtenhofer, C., Fan, H., Malik, J. and He, K. (2019). SlowFast Networks for Video Recognition, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6201-6210.

Fenghour, S., Chen, D., Guo, K, et al. (2021). Deep Learning-Based Automated Lip-Reading: A Survey, *in IEEE Access*, vol. 9, pp. 121184-121205, 2021, doi: 10.1109/ACCESS.2021.3107946.

Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Disentangling Homophemes in Lip Reading using Perplexity Analysis, *ArXiv,* abs/2012.07528.

Fenghour, S., Chen, D., Guo, K. and Xiao, P. (2020). Lip Reading Sentences Using Deep Learning With Only Visual Cues, *IEEE Access*, vol. 8, pp. 215516-215530, doi: 10.1109/ACCESS.2020.3040906.

Fernandez-Lopez, A., & Sukno, F.M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput*., 78, 53-72.

Gholipour, A., Taheri, A. and Mohammadzade, H. (2021). Automated Lip-Reading Robotic System Based on Convolutional Neural Network and Long Short-Term Memory, *Social Robotics: 13th International Conference*, ICSR, Proceedings. Springer-Verlag, Berlin, Heidelberg, 73–84. https://doi.org/10.1007/978-3-030-90525-5_7

Glasmachers, T. (2017). Limits of End-to-end Learning. ACML.

Goldschen, A. (1993). Continuous automatic speech recognition by lipreading.

Goldschen, A.J., Garcia, O.N., Petajan, E.D. (1997). Continuous Automatic Speech Recognition by Lipreading. *In: Shah, M., Jain, R. (eds) Motion-Based Recognition. Computational Imaging and Vision*, vol 9. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-8935-2_14

Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning*, ICML '06, Association for Computing Machinery, New York, NY, USA, 369–376. https://doi.org/10.1145/1143844.1143891

Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks. *ArXiv*, abs/1211.3711.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks, *in Proceedings of the 31st International Conference on International Conference on Machine Learning* - Volume 32, ser. ICML.

Harte, N. and Gillen, E. (2015). TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech, *in IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603-615, doi: 10.1109/TMM.2015.2407694.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

Hinton, G.E., Vinyals, O. and Dean, J. (2015). Distilling the Knowledge in a Neural Network, ArXiv, abs/1503.02531.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9, 1735-1780.

Howell, D., Cox, S. and Theobald, B. (2016). Visual units and confusion modelling for automatic lip-reading, *Image and Vision Computing*, Volume 51, Pages 1-12, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2016.03.003.

Huang, G., Liu, Z. and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261-2269.

Jachimski, D., Czyżewski, A. and Ciszewski, T.E. (2017). A comparative study of English viseme recognition methods and algorithms, *Multimedia Tools and Applications*, 77, 16495-16532.

Jitaru, A. C., Abdulamit, S. and Ionescu, B. (2020). LRRo: a lip reading data set for the under-resourced romanian language*, Proceedings of the 11th ACM Multimedia Systems Conference. Association for Computing Machinery*, New York, NY, USA, 267–272. https://doi.org/10.1145/3339825.3394932

Jitaru, A. C., Ştefan, L. D. and Ionescu, B. (2021). Toward Language-independent Lip Reading: A Transfer Learning Approach, *International Symposium on Signals*, *Circuits and Systems (ISSCS)*, 2021, pp. 1-4, doi: 10.1109/ISSCS52333.2021.9497405.)

Kannan, A., Wu, Y.,Nguyen, P. et al. (2017). An analysis of incorporating an external language model into a sequence-to-sequence model, *arXiv preprint* arXiv:1712.01996, 2017.

Karpathy, A.,Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

Kim, Y., Jernite, Y., Sontag, D. and Rush, A. (2016). Character-aware neural language models. *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI'16). AAAI Press, 2741–2749.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM 60*, 84-90. https://doi.org/10.1145/3065386

Lee, H., Ekanadham, C., and Ng, A. (2007). Sparse deep belief net model for visual area V2. *NIPS.*

Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035-1074, April 1983, doi: 10.1002/j.1538-7305.1983.tb03114.x.

Li, W., Wang, S., Lei, M., Siniscalchi, S.M. and Lee, C. (2019). Improving Audio-visual Speech Recognition Performance with Cross-modal Student-teacher Training, ICASSP 2019 - *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6560-6564.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts, *Int'l Conference on Learning Representations*.

Lu, Y. and Li, H. (2019). Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory, *Appl. Sci.*, vol. 9, no. 8, p. 1599.

Lucey, P., Potamianos, G. and Sridharan, S. (2008). Patch-Based Analysis of Visual Speech From Multiple Views.

Ma, P., Martinez, B., Petridis, S. and Pantic, M. (2021). Towards Practical Lipreading with Distilled and Efficient Models. *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7608-7612.

Maas, A. L., Xie, Z., Jurafsky, D. and Ng, A. Y. (2015). Lexicon-free conversational speech recognition with neural networks, *in Proceedings the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Makino, T., Liao, H., Assael, Y., Shillingford, B., García, B., Braga, O. and Siohan, O. (2019). Recurrent Neural Network Transducer for Audio-Visual Speech Recognition, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 905-912.

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading Using Temporal Convolutional Networks, ICASSP 2020 - *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6319-6323.

Mase, K. and Pentland, A. (1989). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22, 67-76.

Matthews, I., Cootes, T. F., Bangham, J. A. et al. (2002). Extraction of visual features for lipreading, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198-213, doi: 10.1109/34.982900.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature 264*, 746-748. https://doi.org/10.1038/264746a0

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maître, G. (1999). XM2VTSDB: The Extended M2VTS Database.

Movellan, J. R. (1994). Visual Speech Recognition with Stochastic Networks, NIPS.

Mroueh, Y., Marcheret, E. and Goel, V. (2015). Deep multimodal learning for Audio-Visual Speech Recognition, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2130-2134.

Ngiam, J., Khosla, A.,Kim, M. et al. (2011). Multimodal deep learning, *Proc. 28th Int. Conf. Mach. Learn*., (ICML), pp. 1–8.

Ojala, T., Pietikäinen, M. and Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, vol. 1, pp. 582 - 585.

Ostendorf, M. and S. Roukos, (1989). A stochastic segment model for phoneme-based continuous speech recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1857-1869, doi: 10.1109/29.45533.

Palanivel, S. and Yegnanarayana, V. (2008). Multimodal person authentication using speech, face and visual speech, *Comput. Vis. Image Underst.109*, 44–55. https://doi.org/10.1016/j.cviu.2006.11.013

Patterson, E.K., Gurbuz, S., Tufekci, Z. et al. (2002). Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus. *EURASIP J. Adv. Signal Process*, 208541, https://doi.org/10.1155/S1110865702206101

Petajan, E. (1984). Automatic lipreading to enhance speech recognition (speech reading).

Petridis, S., and Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2304-2308.

Petridis, S., Li, Z. and Pantic, M. (2017). End-to-end visual speech recognition with LSTMS, *2017 IEEE International Conference on Acoustics, Speech and Signal*

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G. and Pantic, M. (2018). End-to-end Audiovisual Speech Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6548-6552.

Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G. and Pantic, M. (2018). Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. *2018 IEEE Spoken Language Technology Workshop* (SLT), 513-520.

Petridis, S., Wang, Y., Li, Z. and Pantic, M. (2017). End-to-end Multi-View Lipreading. *ArXiv*, abs/1709.00443.

Petridis, S., Wang, Y., Ma, P., Li, Z. and Pantic, M. (2020). End-to-end Visual Speech Recognition for Small-Scale Datasets. *Pattern Recognit. Lett.*, 131, 421-427.

Peymanfard, J., Mohammadi, M.R., Zeinali, H. and Mozayani, N. (2022). Lip reading using external viseme decoding, *2022 International Conference on Machine Vision and Image Processing (MVIP)*, 1-5.

Potamianos, G., Neti, C., Iyengar, G. et al. (2001). A Cascade Visual Front End for Speaker Independent Automatic Speechreading, *International Journal of Speech Technology 4*, 193–208, https://doi.org/10.1023/A:1011352422845 *Processing (ICASSP)*, 2592-2596.

Rekik, A., Ben-Hamadou, A. and Mahdi, W. (2014). A New Visual Speech Recognition Approach for RGB-D Cameras, Campilho, A., Kamel, M. (eds) Image Analysis and Recognition. ICIAR 2014. Lecture Notes in Computer Science(), vol 8815. Springer, Cham. https://doi.org/10.1007/978-3-319-11755-3_3

Ren, S., Du, Y., Lu, J. et al. (2021). Learning from the Master: Distilling Cross-modal Advanced Knowledge for Lip Reading, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320-13328, doi: 10.1109/CVPR46437.2021.01312.

Saenko, K., Livescu, K., Glass, J. and Darrell, T. (2005). Production domain modeling of pronunciation for visual speech recognition, *Proceedings. (ICASSP '05). IEEE International*

Conference on Acoustics, Speech, and Signal Processing, pp. v/473-v/476 Vol. 5, doi: 10.1109/ICASSP.2005.1416343.

Sanderson, C. (2002). The VidTIMIT database, *Tech. rep*., IDIAP.

Seymour, R., Stewart, D. and Ming, J. (2008). Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *Image Video Process*, 9 pages, https://doi.org/10.1155/2008/810362

Shillingford, B., Assael, Y., Hoffman et al. (2019). Large-Scale Visual Speech Recognition. *ArXiv*, abs/1807.05162.

Singla, Y.K., Sahrawat, D., Maheshwari, S., et al. (2020). Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition. *AAAI*.

Srivastava, R.K., Greff, K. and Schmidhuber, J. (2015). Highway Networks. *ArXiv*, abs/1505.00387.

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining Residual Networks with LSTMs for Lipreading. *ArXiv*, abs/1703.04105.

Su, J., Vargas, D.V. and Sakurai, K. (2019). One Pixel Attack for Fooling Deep Neural Networks, *IEEE Transactions on Evolutionary Computation*, 23, 828-841.

Sumby, W. H. and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America 26*, 212.

Tang, X., Bouzerdoum, A. and Phung, S. L. (2015). Video Classification Based on Spatial Gradient and Optical Flow Descriptors, *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371319.

Tang, Y., Ma, L. and Zhou, L. (2019). Hallucinating Optical Flow Features for Video Classification, *ArXiv*, abs/1905.11799.

Tao, F. and Busso, C. (2018). Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290-1302, doi: 10.1109/TASLP.2018.2815268.

Thangthai, K., Bear, H.L. and Harvey, R. (2018). Comparing phonemes and visemes with DNN-based lipreading, *ArXiv*, abs/1805.02924.

Theobald, B. J., Harvey R., Cox, S. J. et al. (2006). Lip-reading enhancement for law enforcement, *Proc. SPIE 6402, Optics and Photonics for Counterterrorism and Crime Fighting II*, 640205 (28 September 2006); https://doi.org/10.1117/12.689960

Tran, D., Bourdev, L., Fergus, R. et al. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks, *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, USA, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need, ArXiv, abs/1706.03762.

Vougioukas, K., Petridis, S., & Pantic, M. (2018). End-to-end Speech-Driven Facial Animation with Temporal GANs. *BMVC*.

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

Wang, D., Wang, X. and Lu S. (2019). An Overview of End-to-end Automatic Speech Recognition, *Symmetry*. 2019; 11(8):1018. https://doi.org/10.3390/sym11081018

Wang, Y., Yao, Q., Kwok, J.T., & Ni, L.M. (2019). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *arXiv*: Learning.

Watanabe, S., Hori, T., Kim, S. et al (2017). Hybrid CTC/attention architecture for end-to-end speech recognition, *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253.

Weng, X. and Kitani, K. (2019). Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading, *BMVC*.

Wu, P., Liu, H., Li, X. et al. (2016). A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion, *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326-338, doi: 10.1109/TMM.2016.2520091.

Xian, Y., Lampert, C.H., Schiele, B. and Akata, Z. (2019). Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 2251-2265.

Xu, K., Li, D., Cassimatis, N. and Wang, X. (2018). LCANet: End-to-end Lipreading with Cascaded Attention-CTC. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 548-555.

Yang, S., Zhang, Y., Feng, D. et al. (2019). LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1-8.

Yu, J., Zhang, S., Wu, J. et al. (2020). Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6984-6988.

Yuhas, B.P., Goldstein, M.H. and Sejnowski, T.J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27, 65-71.

Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6848-6856.

Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020). Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition* (FG 2020), 356-363.

Zhao, G., Barnard, M. and Pietikainen, M. (2009). Lipreading With Local Spatiotemporal Descriptors, *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254-1265, Nov. 2009, doi: 10.1109/TMM.2009.2030637.

Zhao, X., Yang, S., Shan, S., & Chen, X. (2020). Mutual Information Maximization for Effective Lip Reading. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 420-427.

Zhao, Y., Xu, R., Wang, X., Hou, P., Tang, H. and Song, M. (2020). Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers, *AAAI*.

Zhao, Y., Xu, R. and Song, M. (2019). A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading*, Proceedings of the ACM Multimedia Asia*.

Zheng, S., Song, Y., Leung, T. and Goodfellow, I.J. (2016). Improving the Robustness of Deep Neural Networks via Stability Training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4480-4488.

Zhou, Z., Hong, X., Zhao, G. and Pietikäinen, M. (2014). A Compact Representation of Visual Speech Data Using Latent Variables, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 1-1, doi: 10.1109/TPAMI.2013.173.

Zhou, Z., Zhao, G., Hong, X., & Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image Vis. Comput*., 32, 590-605.

Appendix

A. Datasets

Details of the datasets used in the experiments.

LRW file structure

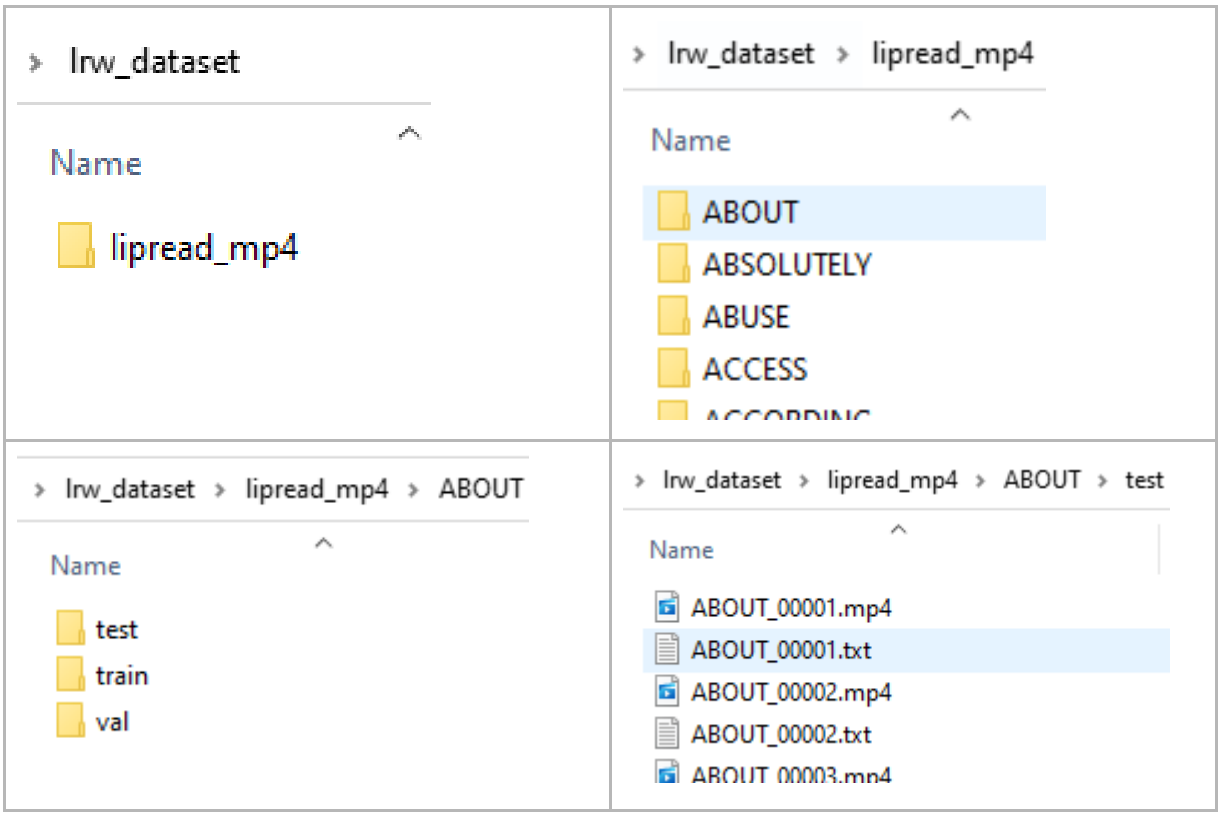


Figure 30. a): LRW file structure

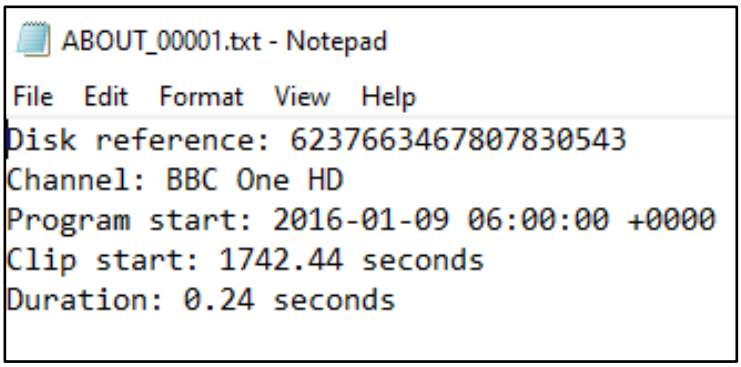


Figure 30. b): Sample .txt file metadata for each word .mp4 file clip

LRS3-TED file structure

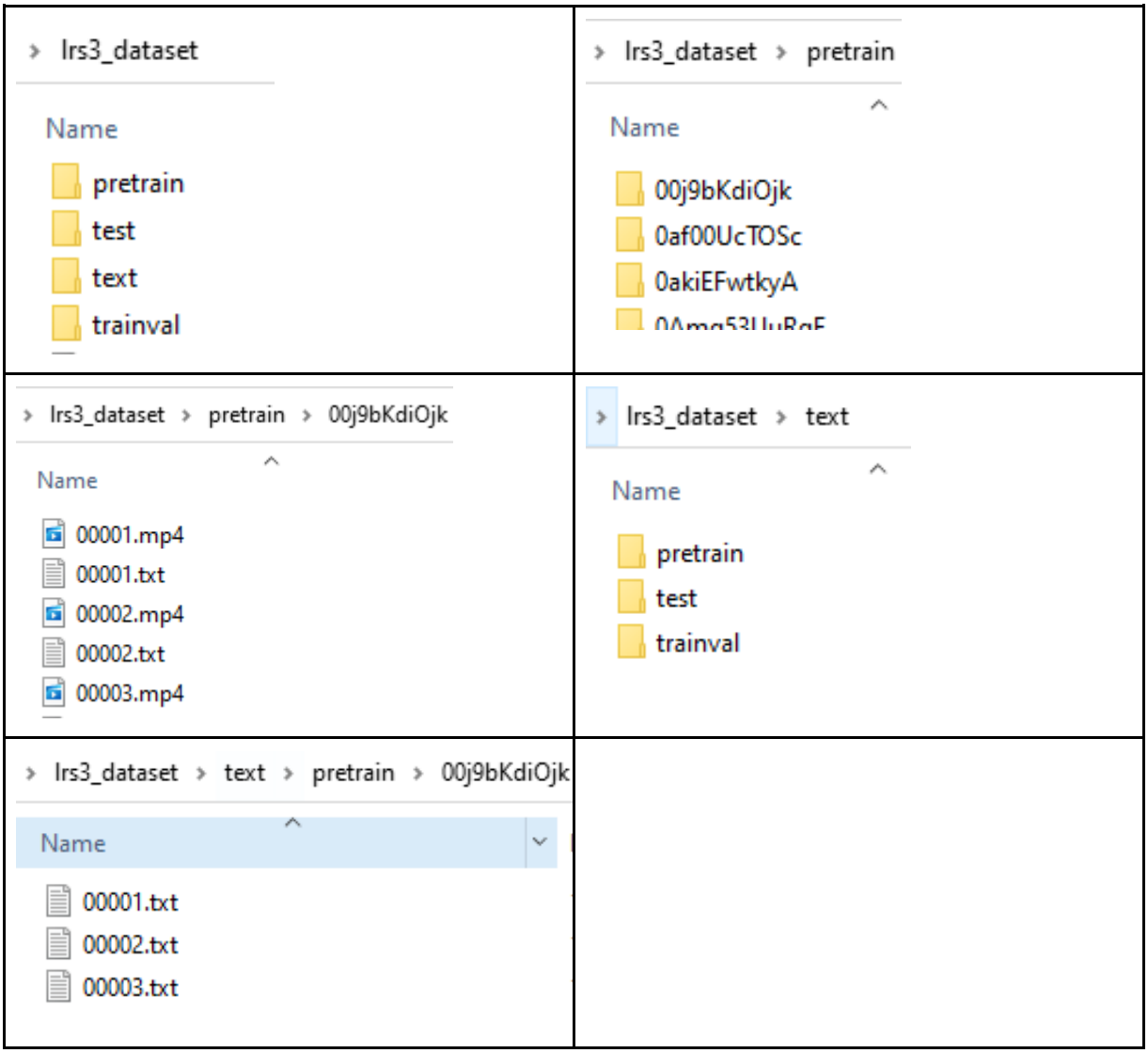


Figure 30. c): LRS3-TED dataset file structure

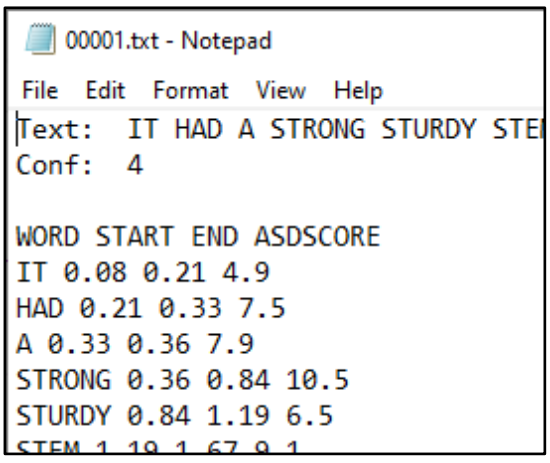


Figure 30. d): LRS3-TED sample label text file

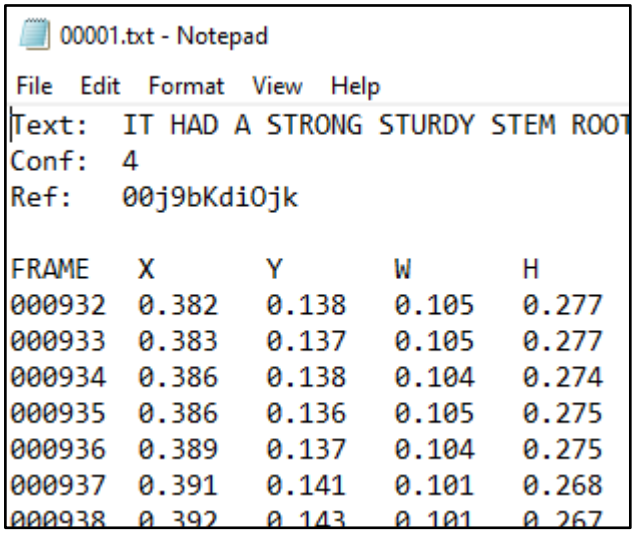


Figure 30. e): LRS3-TED sample text file with label and face bounding box coordinates.

B. Confusion matrices

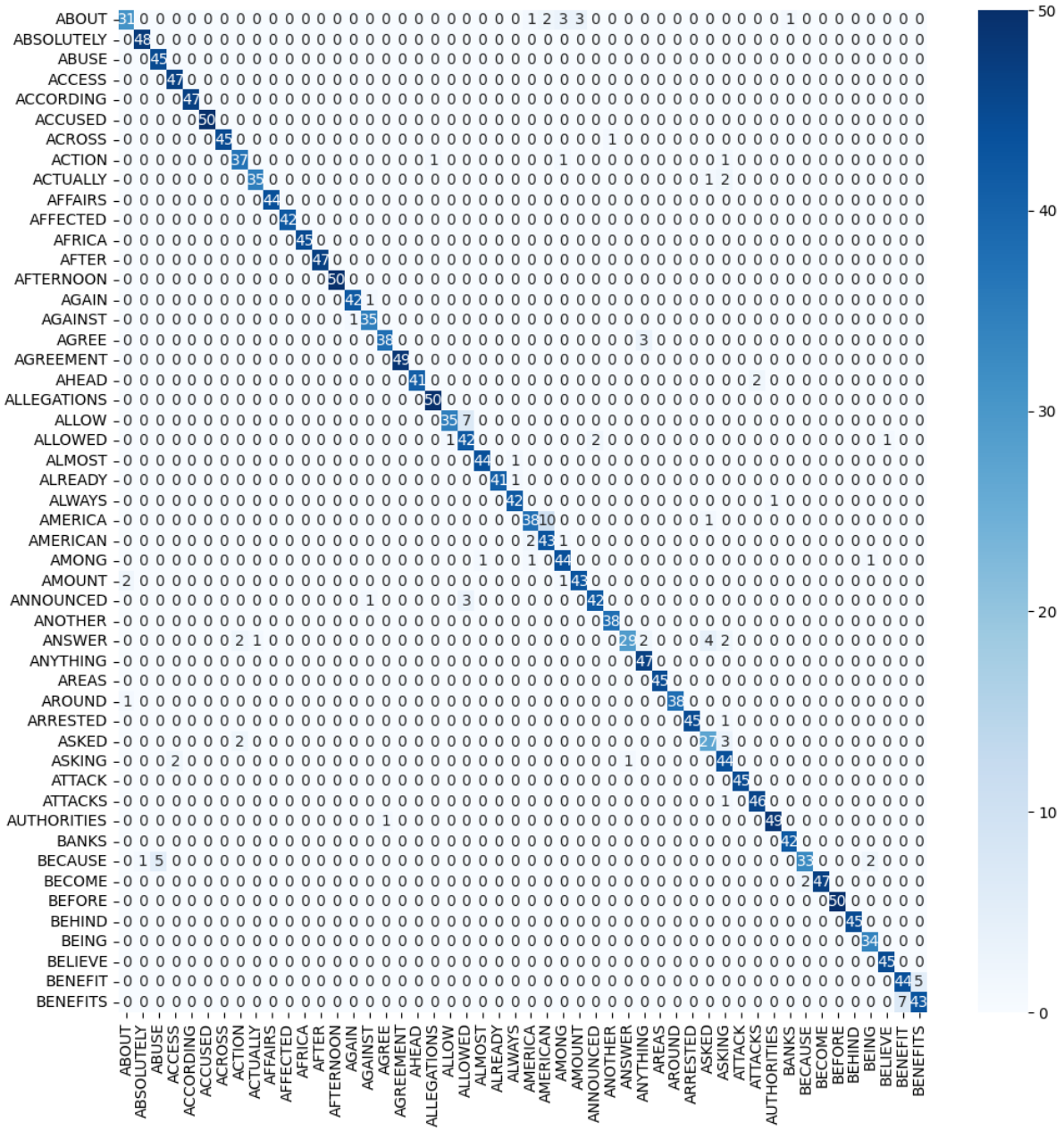


Figure 31. a.) Confusion matrix first 50 words. Commonly confused word pairs: [('ALLOW', 'ALLOWED'), ('BENEFIT', 'BENEFITS'), ('ANSWER', 'ASKED')]

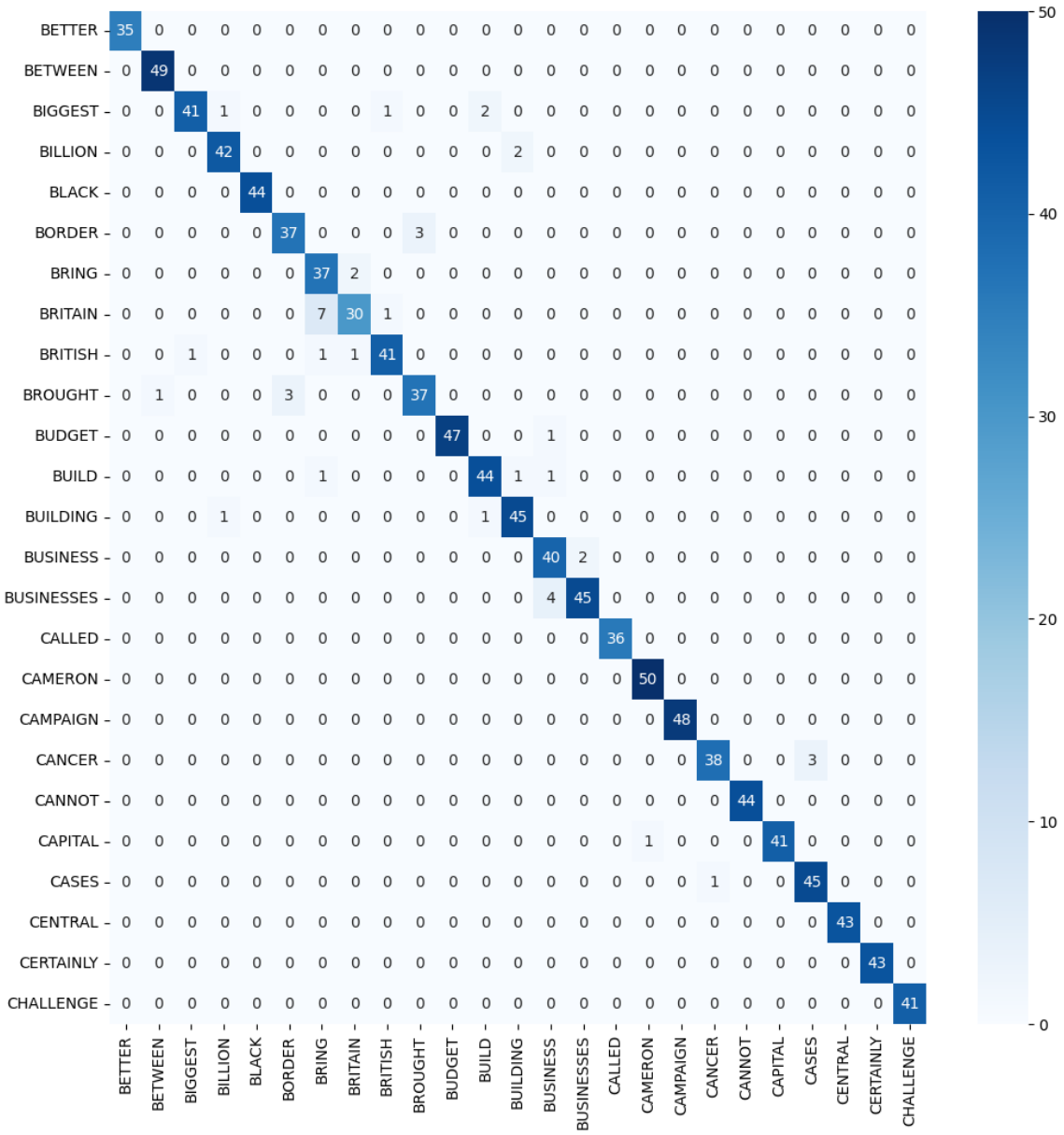


Figure 31. b.) Confusion matrix, words 50-75

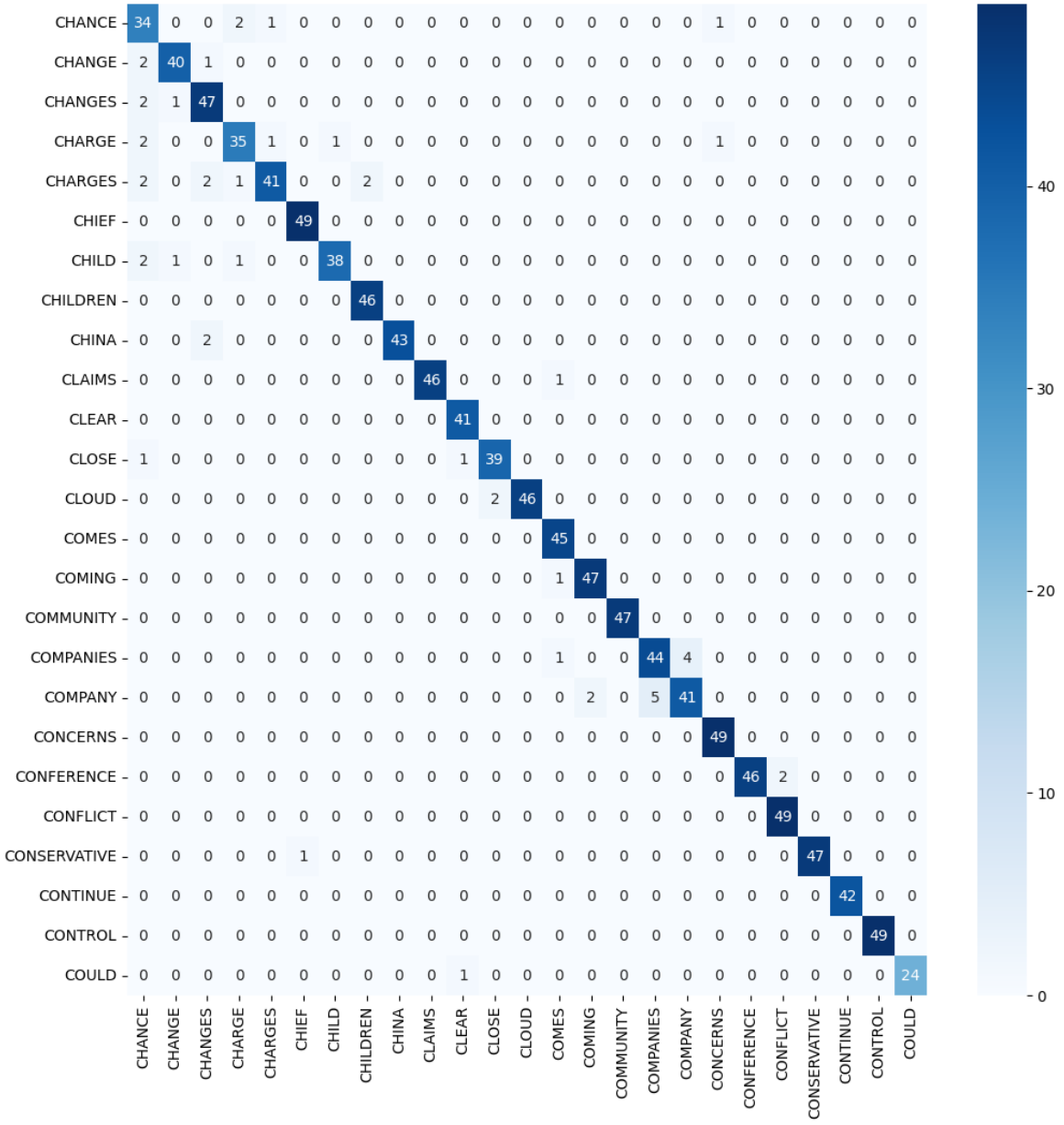


Figure 31. c.) Confusion matrix, words 75-100

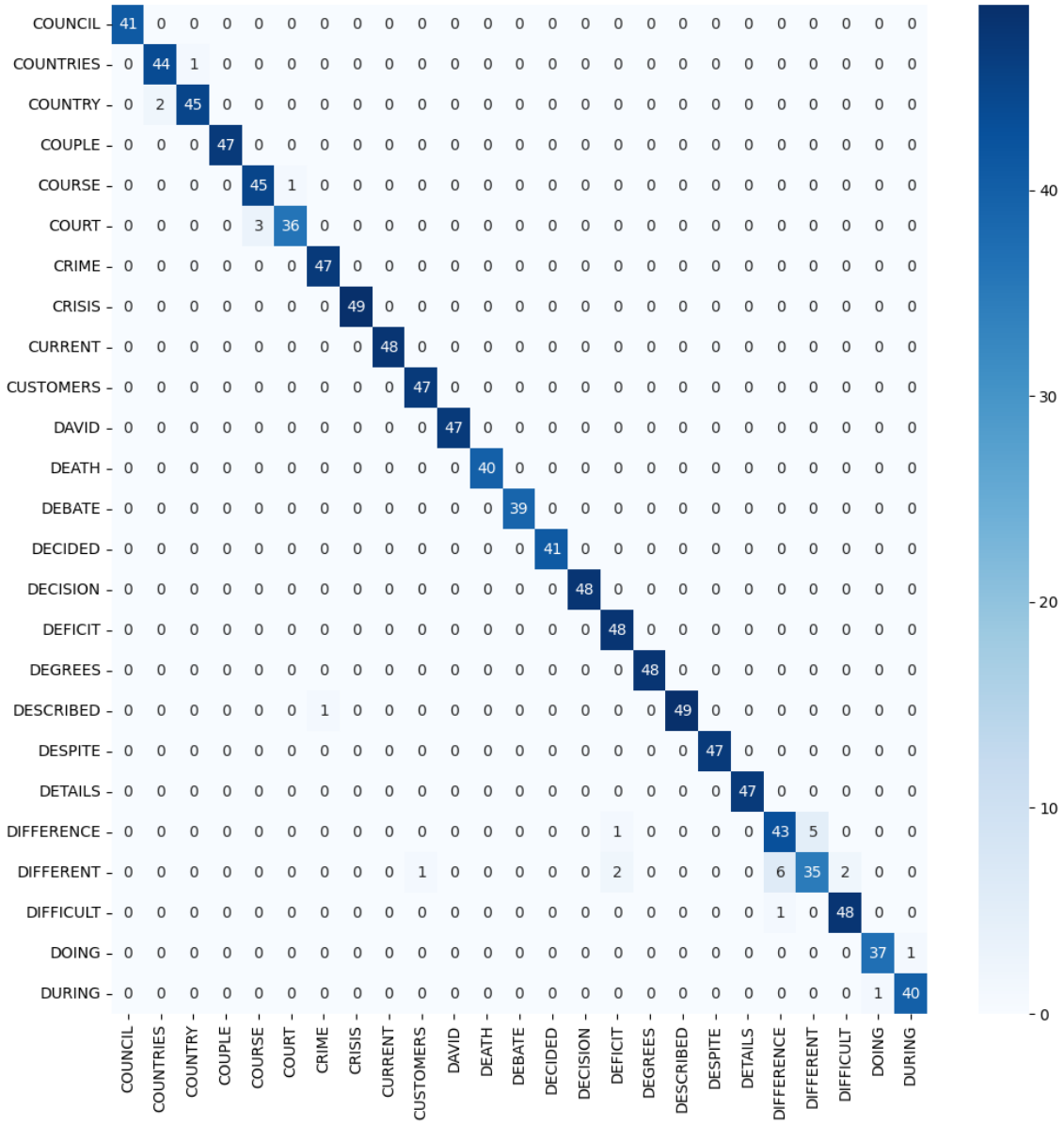


Figure 31. d.) Confusion matrix, words 100-125

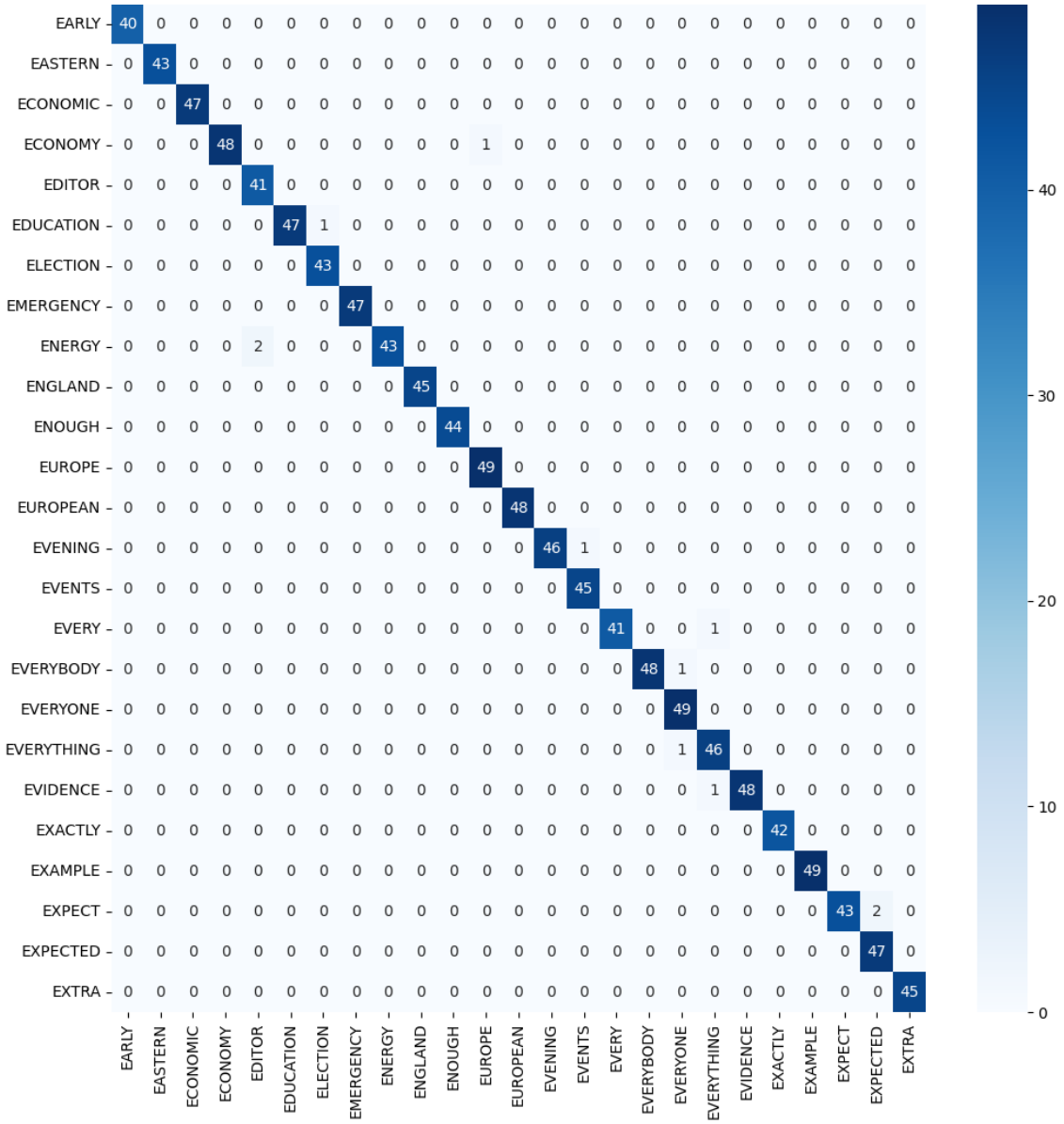


Figure 31. e.) Confusion matrix, words 125-150

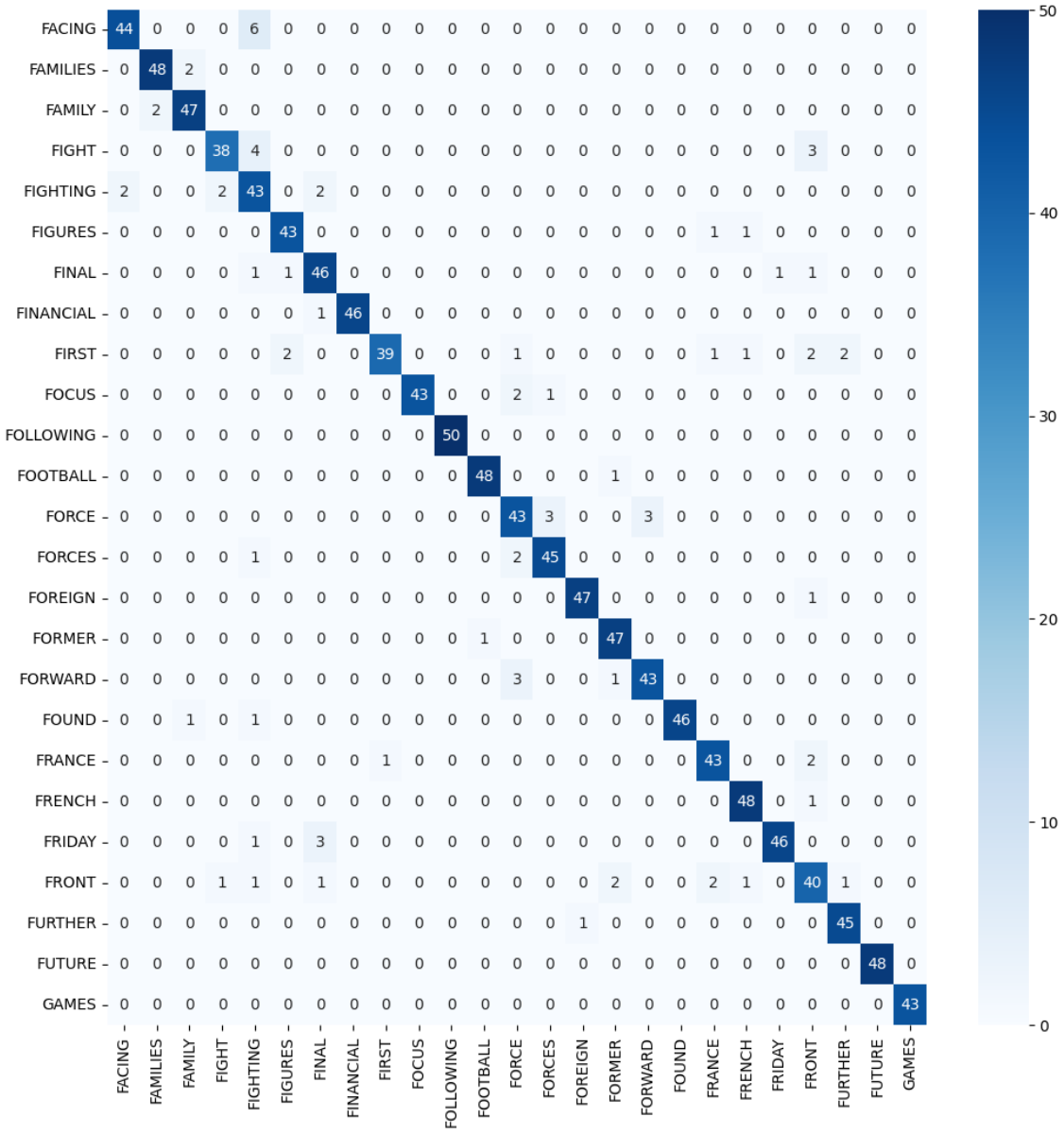


Figure 31. f.) Confusion matrix, words 150-175

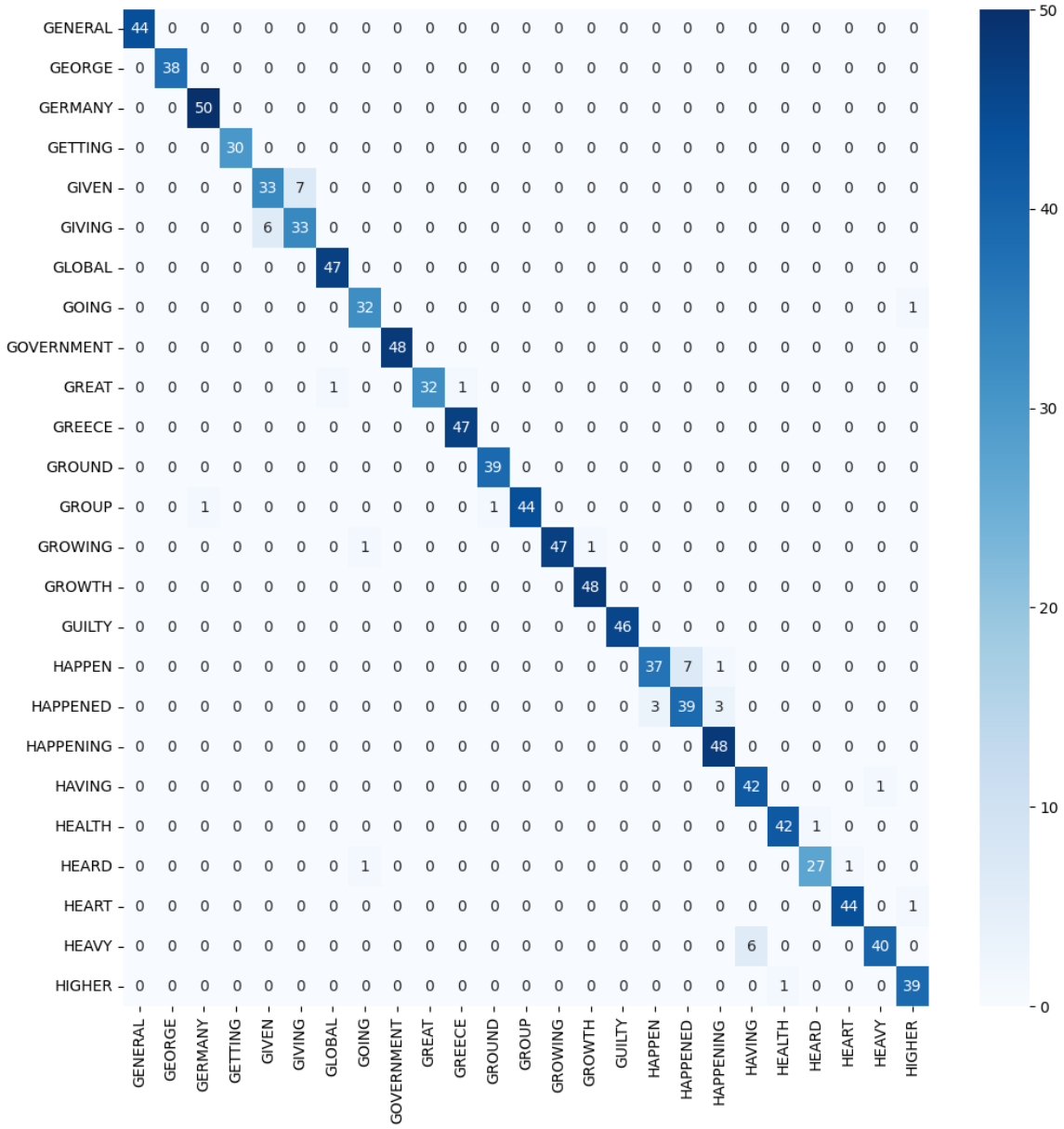


Figure 31. g.) Confusion matrix, words 175-200

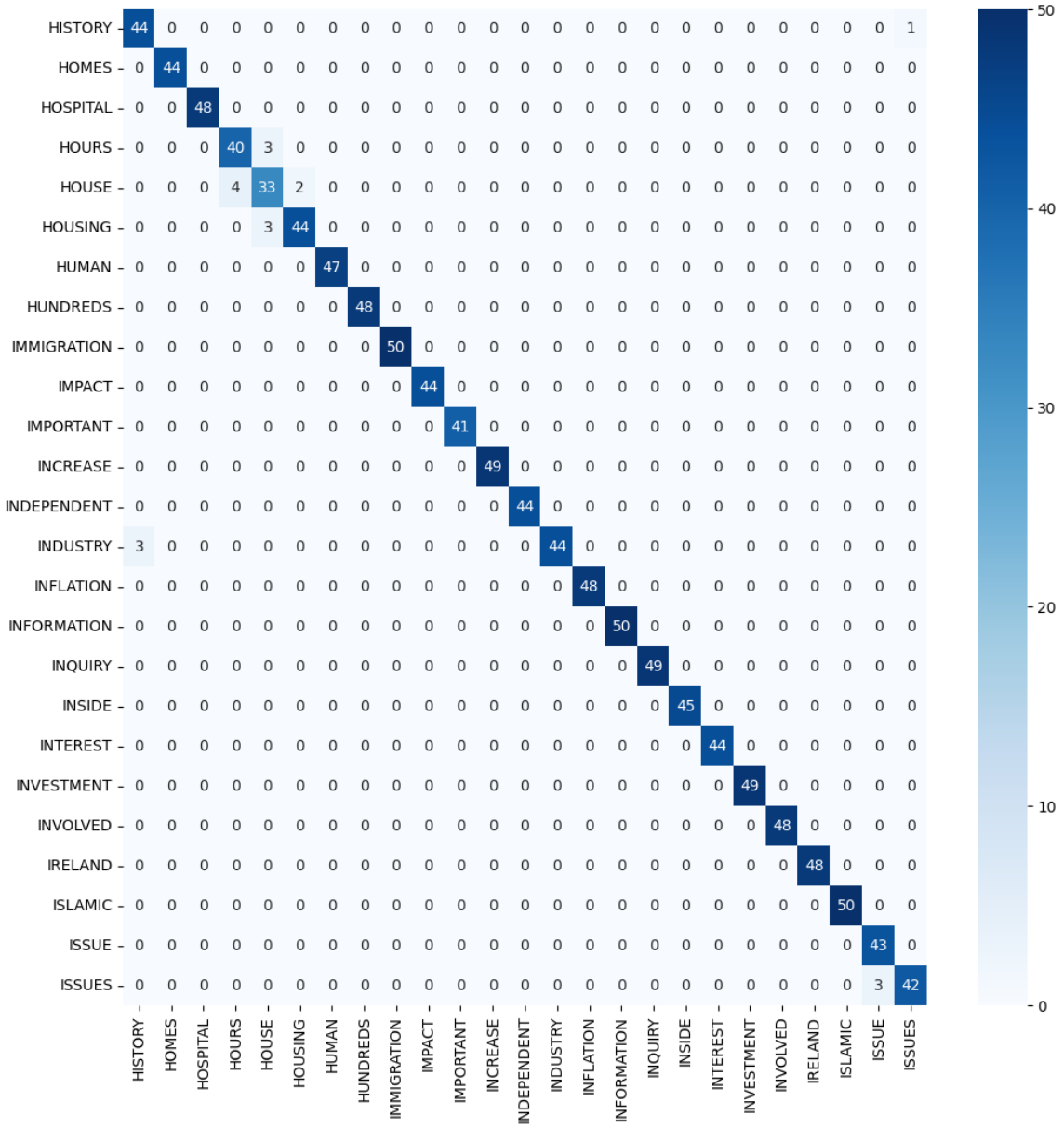


Figure 31. h.) Confusion matrix, words 200-225

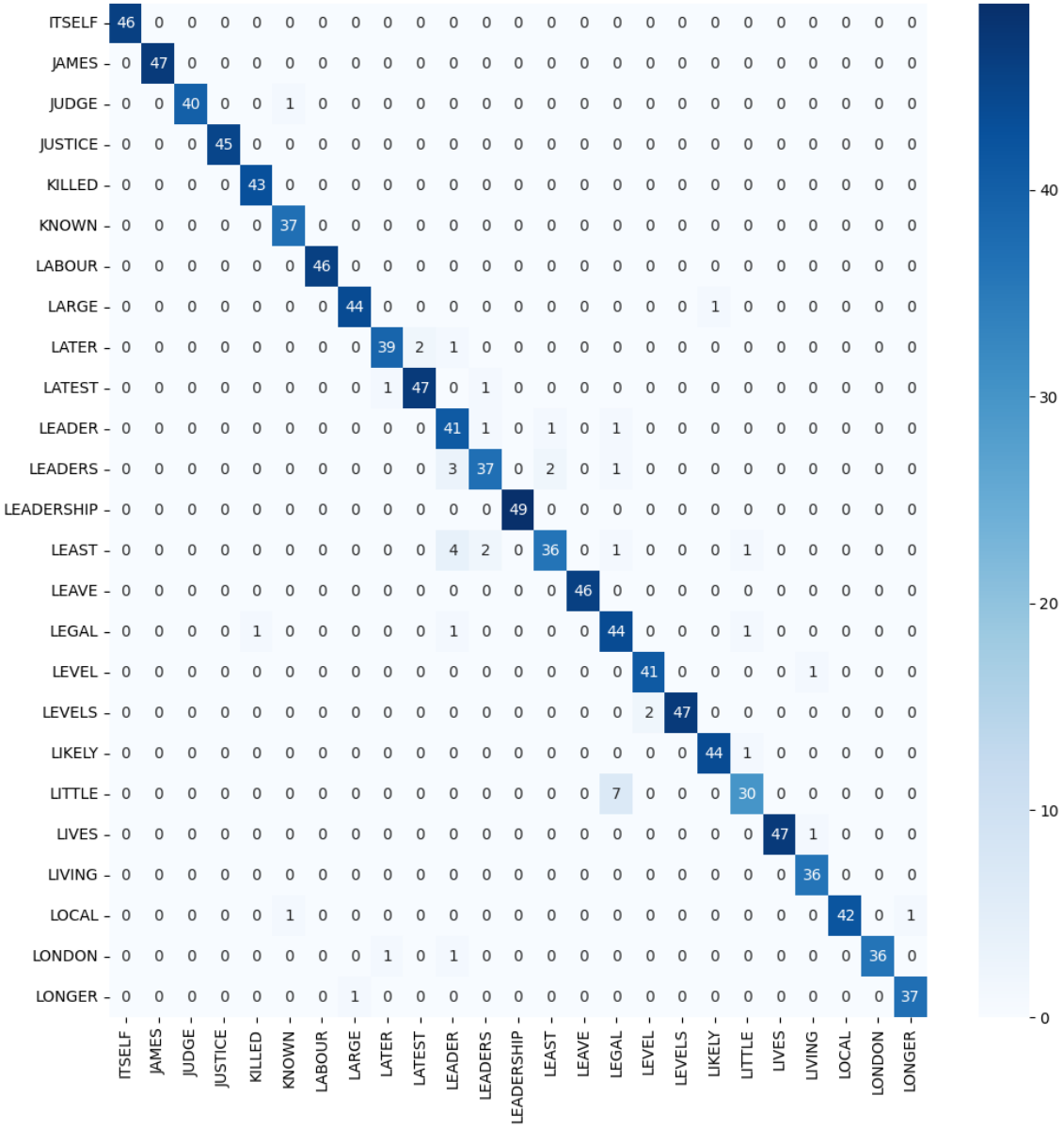


Figure 31. i.) Confusion matrix, words 225-250

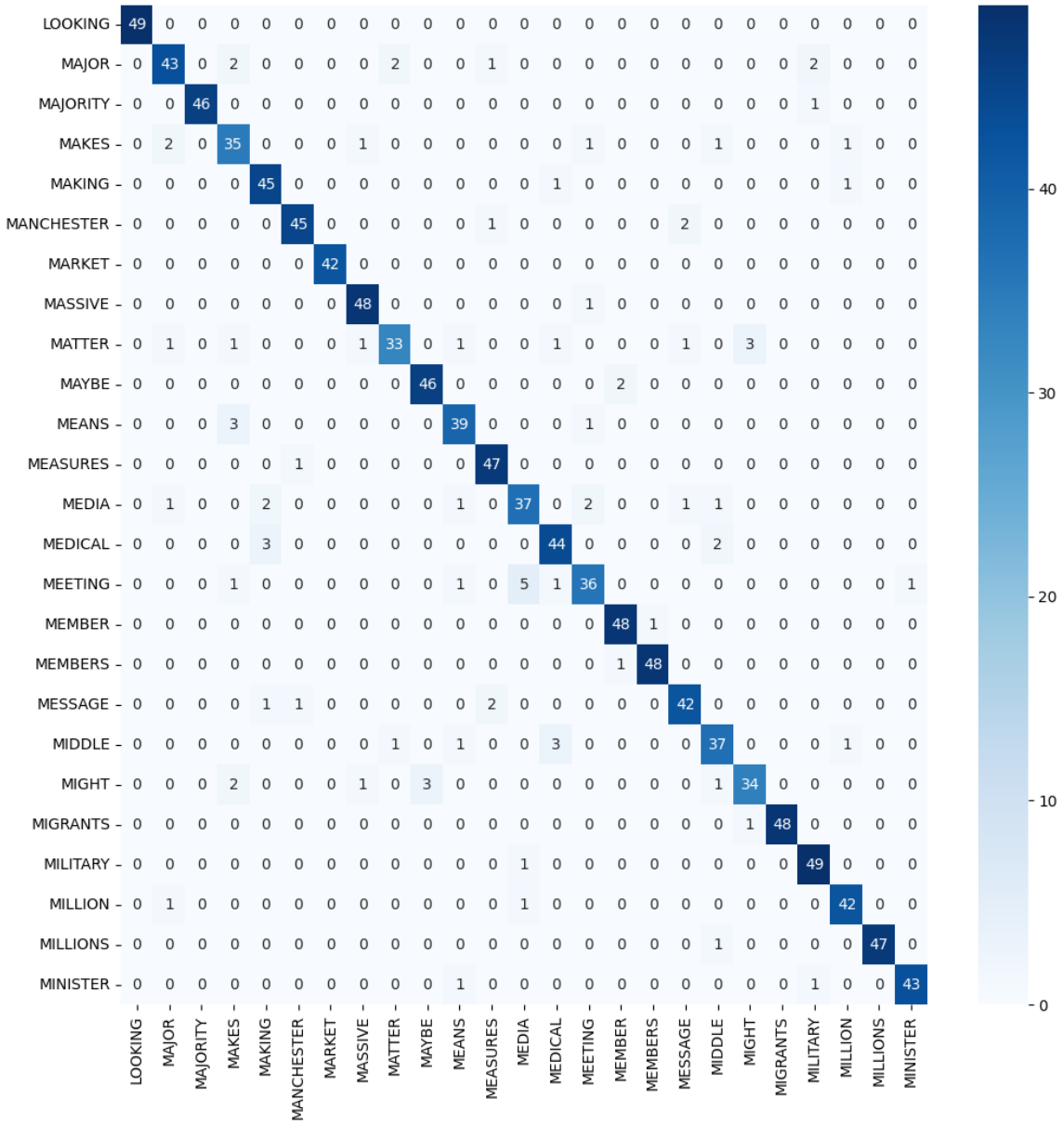


Figure 31. j.) Confusion matrix, words 250-275

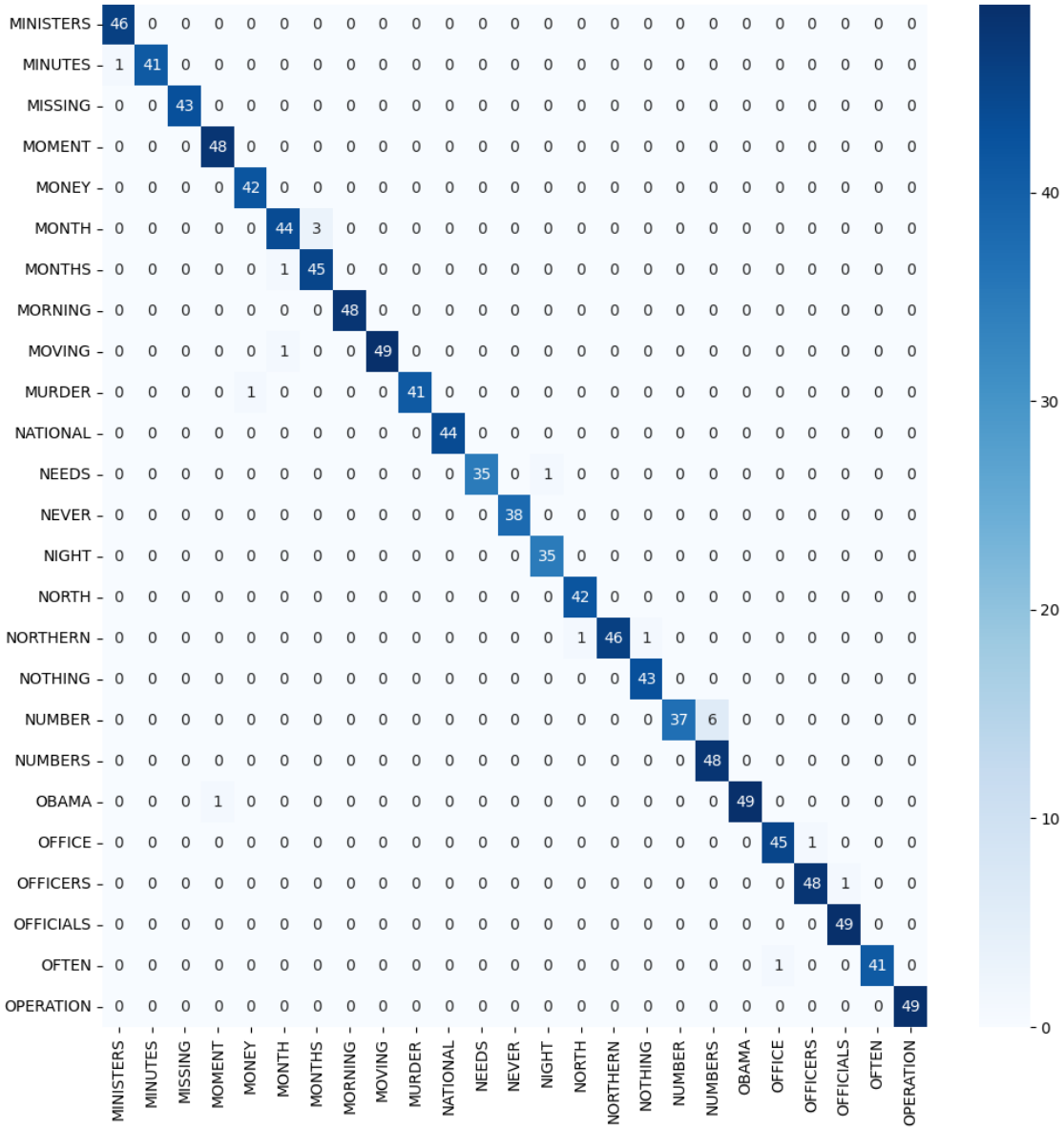


Figure 31. k.) Confusion matrix, words 275-300

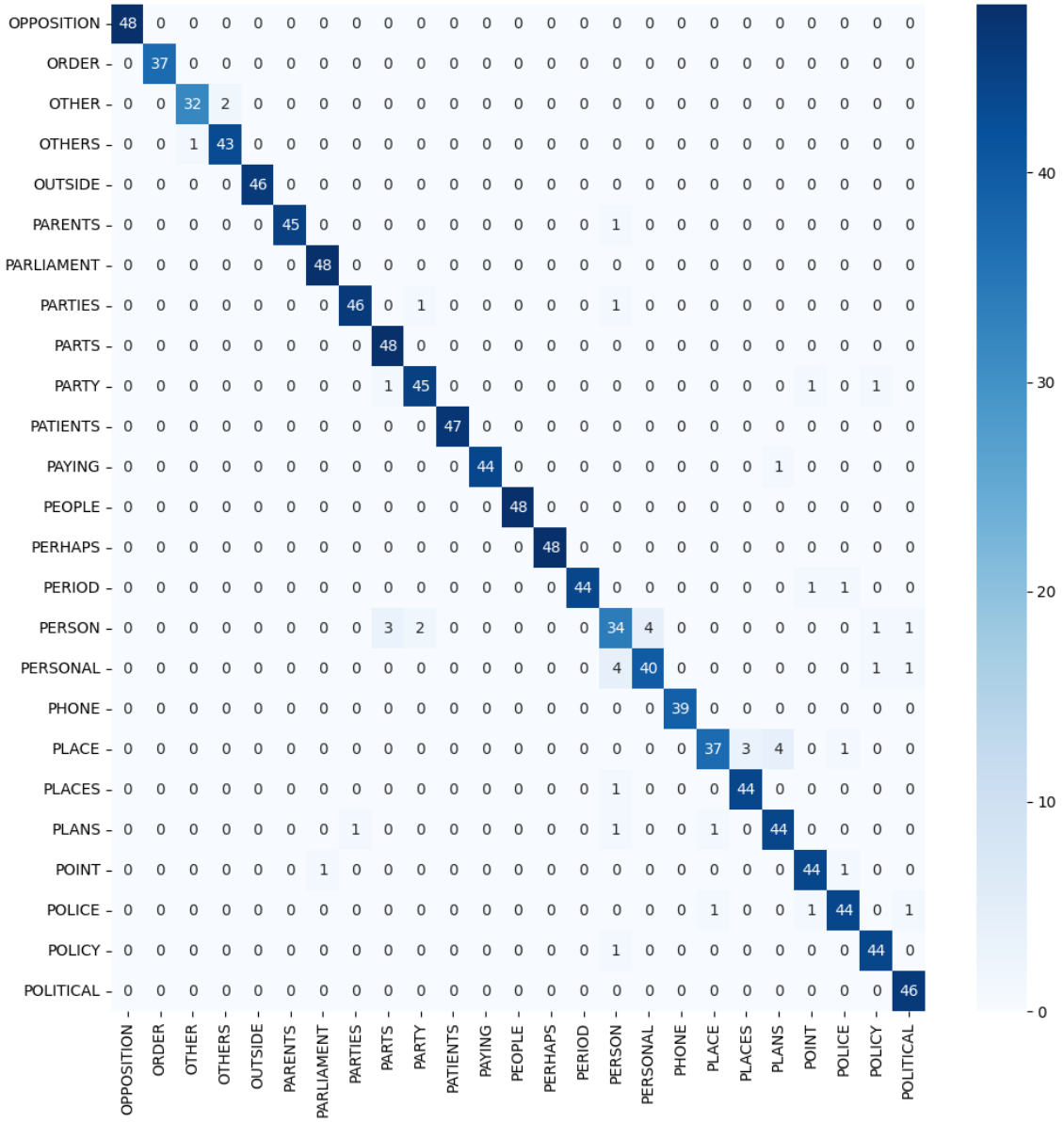


Figure 31. l.) Confusion matrix, words 300-325

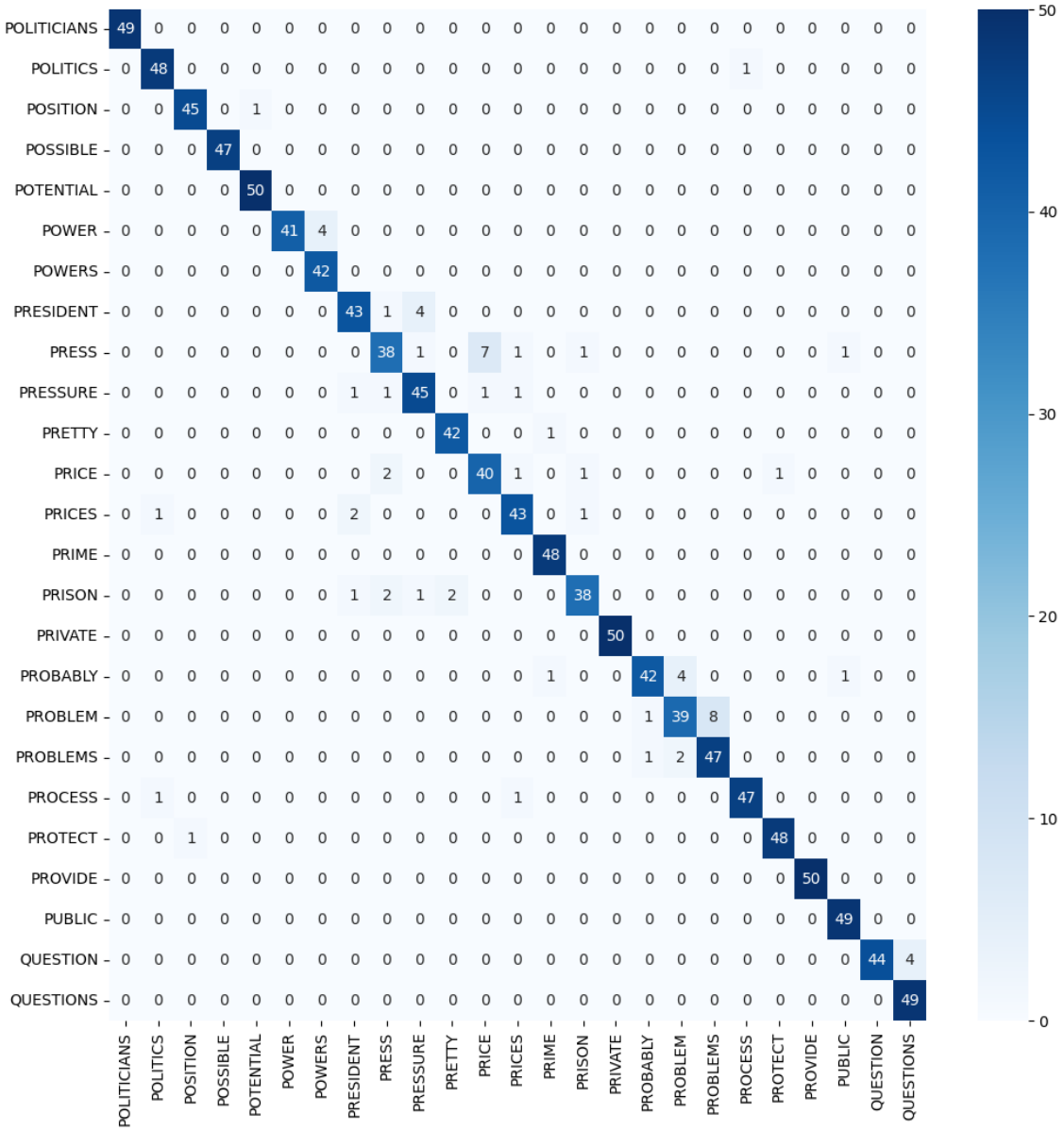


Figure 31. m.) Confusion matrix, words 325-350

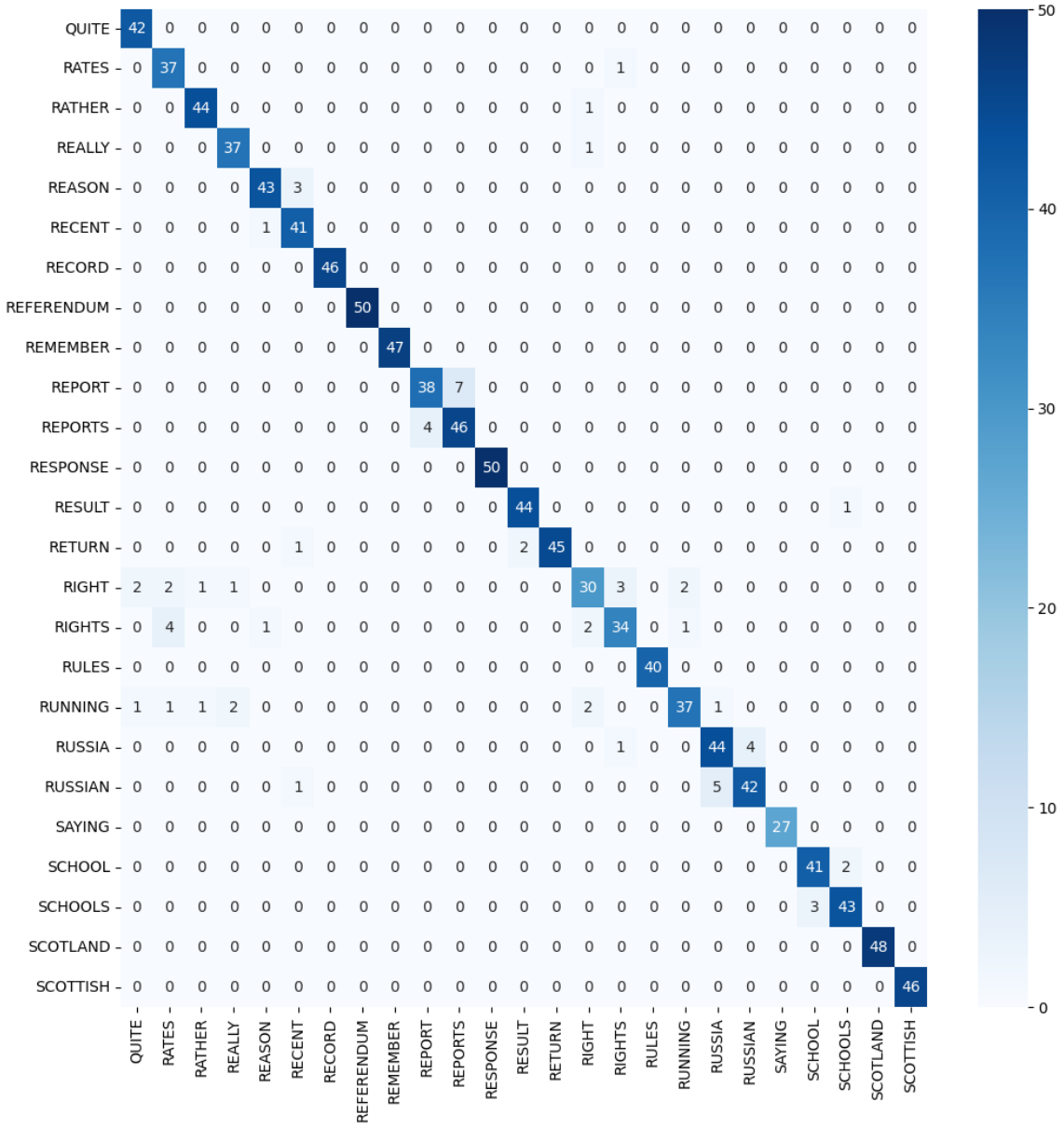


Figure 31. n.) Confusion matrix, words 350-375

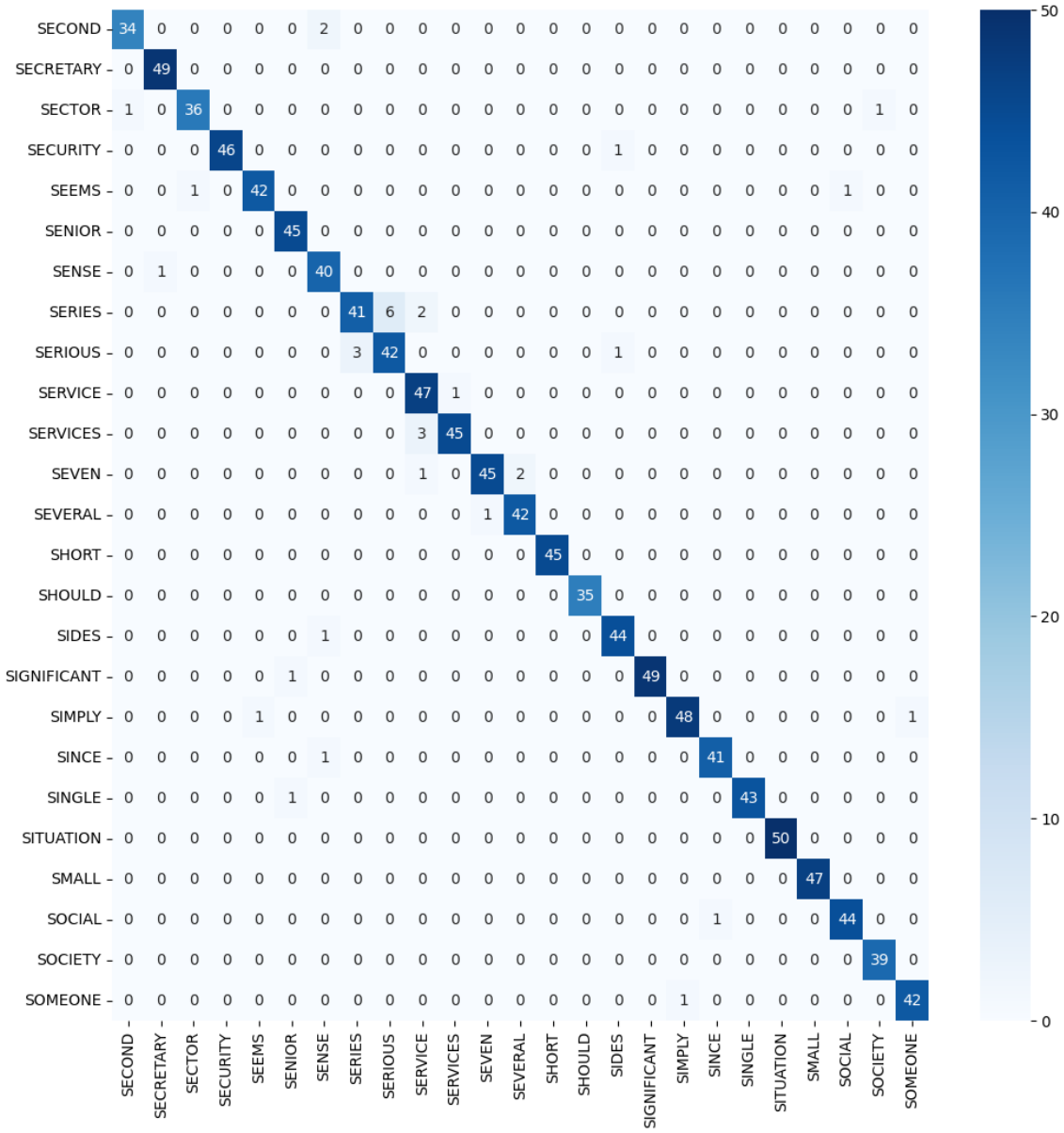


Figure 31. o.) Confusion matrix, words 375-400

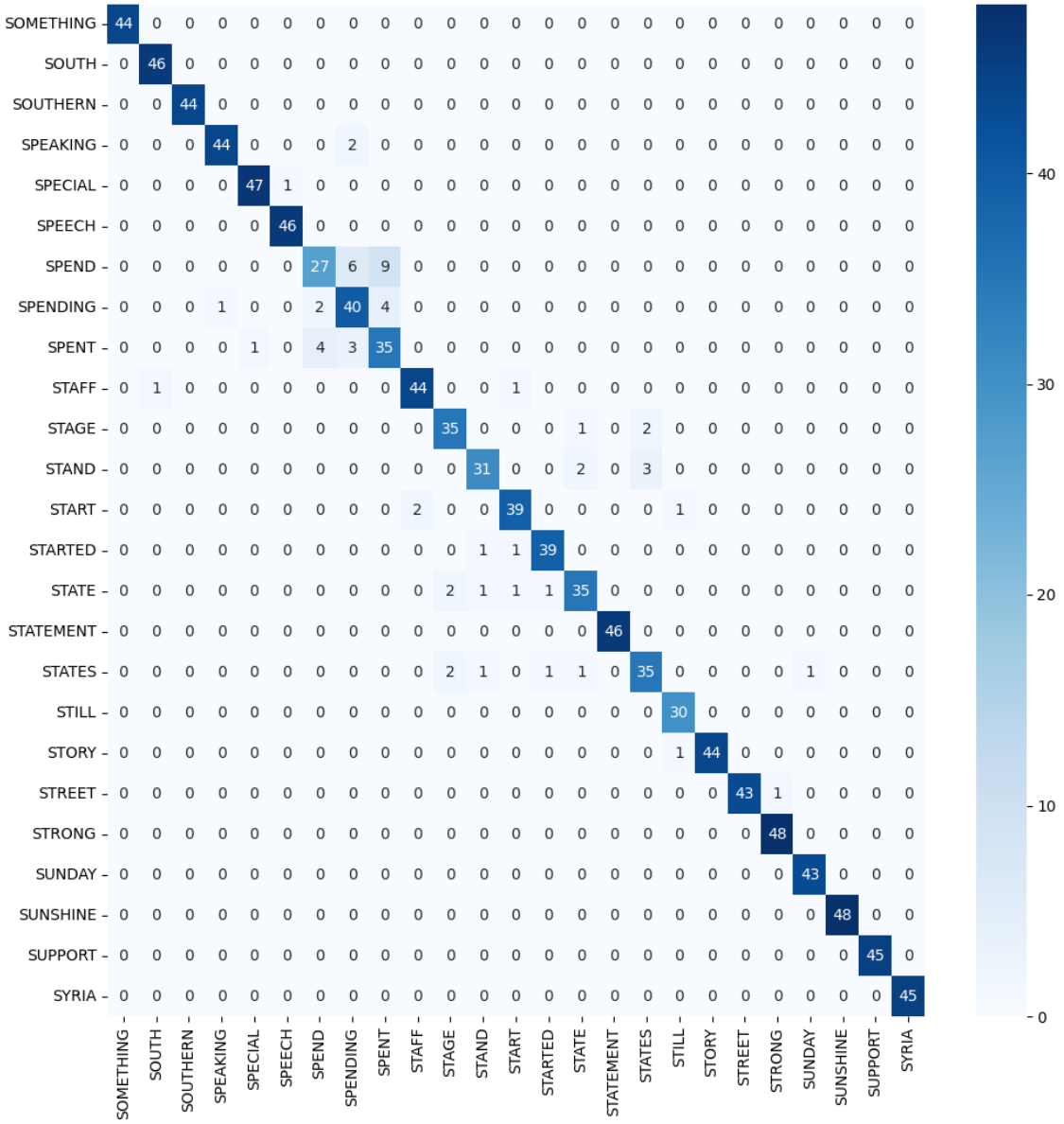


Figure 31. p.) Confusion matrix, words 400-425

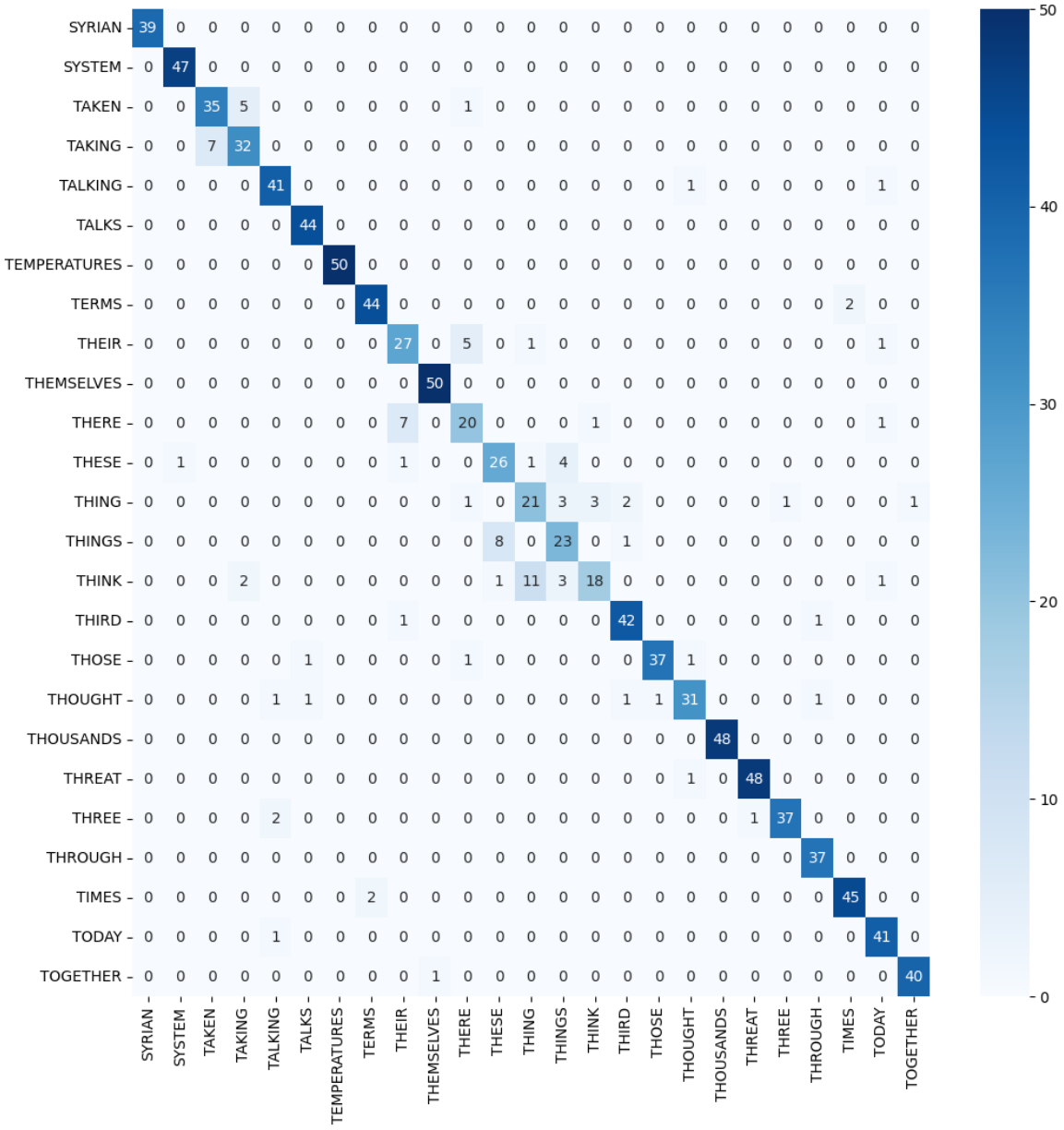


Figure 31. q.) Confusion matrix, words 425-450

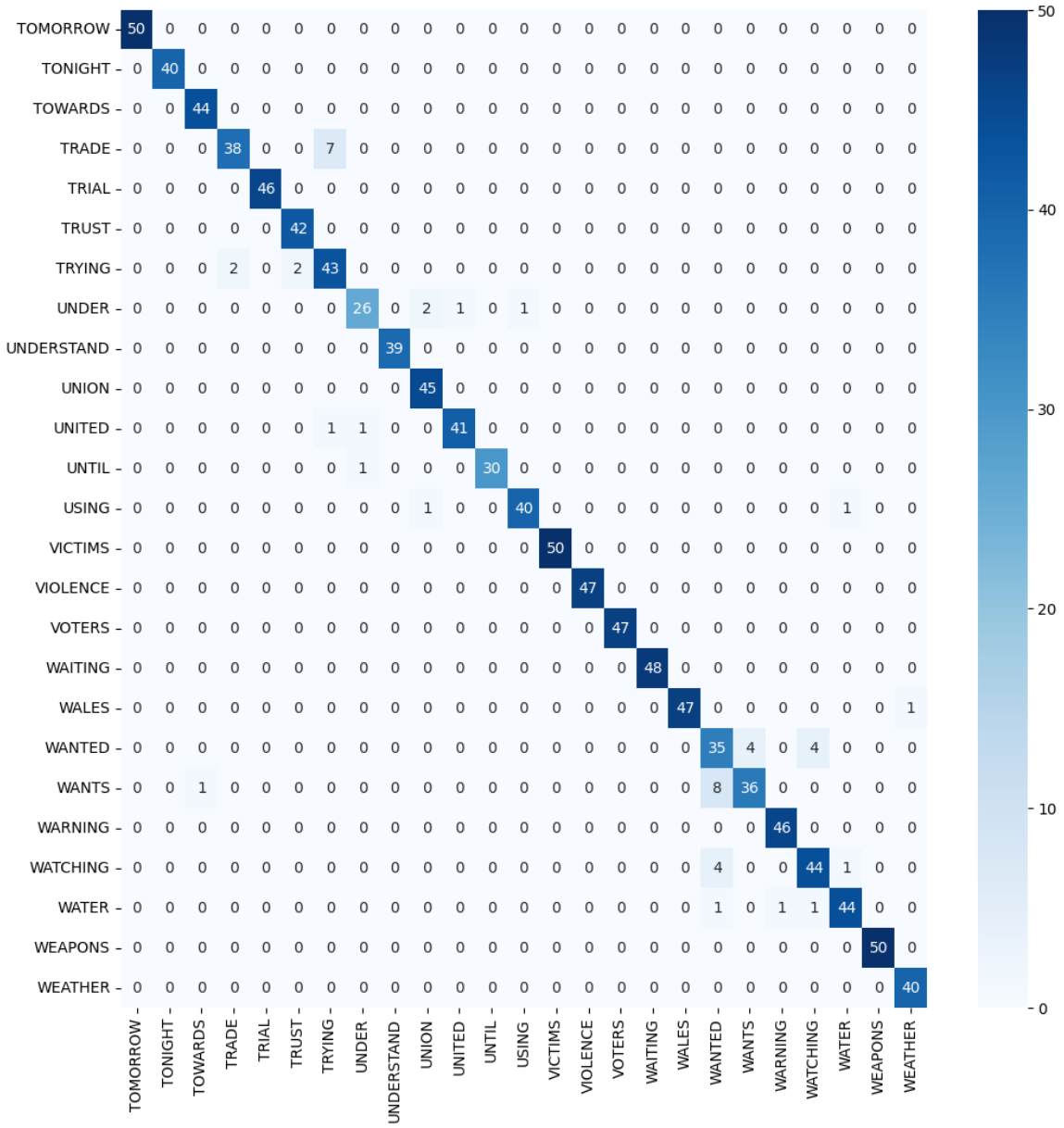


Figure 31. r.) Confusion matrix, words 450-475

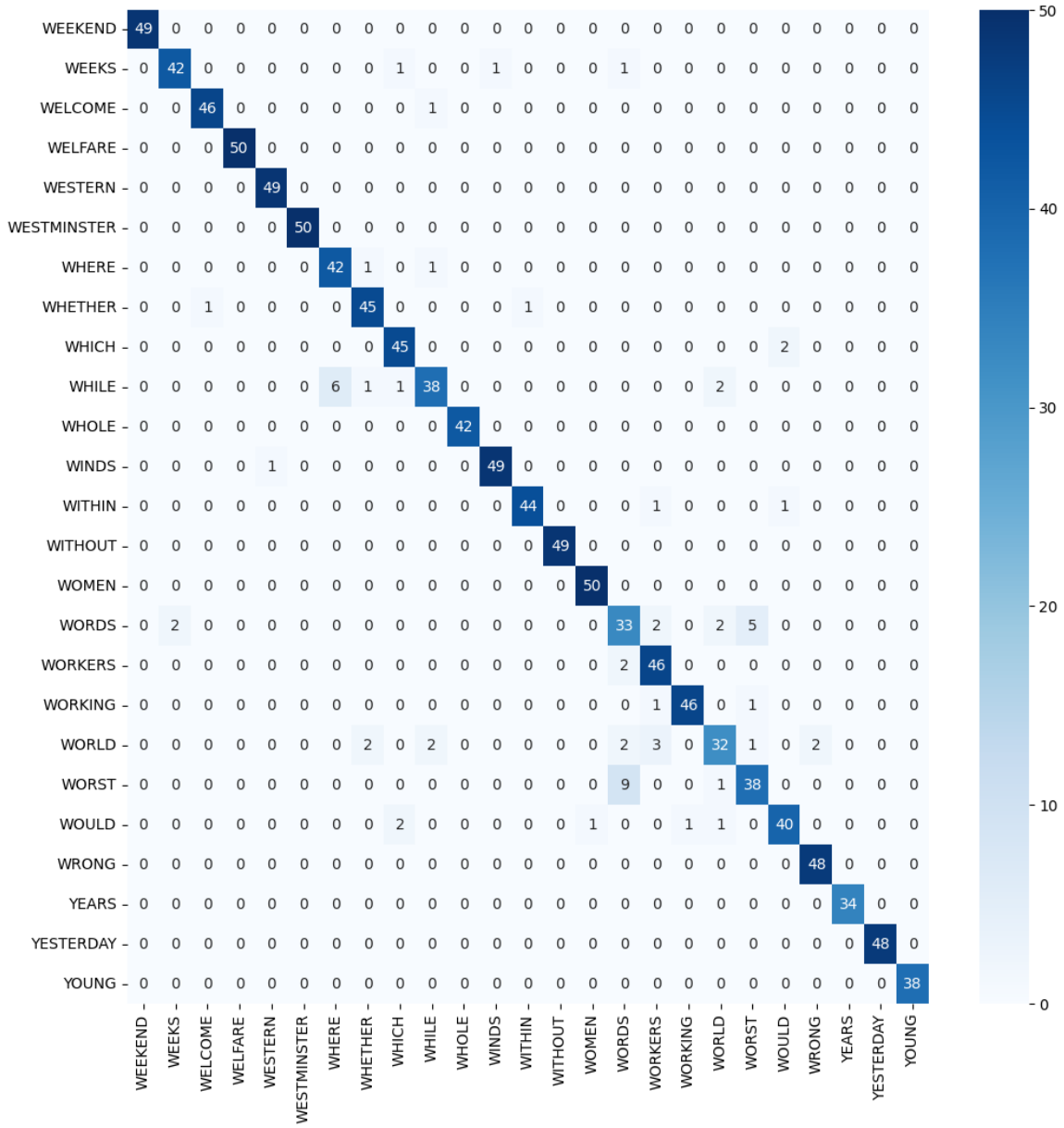


Figure 31. s.) Confusion matrix, words 475-500

C. Top-5 predictions per label

Since the diagrams (Figure 31.a. - 31.s.) only depict a narrow window of size 25 for visibility purposes (e.g. a full 500 x 500 matrix would be too large to include in this document), Table 10 lists the top 5 predictions for each of the 500 words in the labels. Note: some words have less than 5 predictions e.g. word 2: 'ABSOLUTELY', while some have more than 5 predictions which are not included in the table. Each class/word contains 50 samples in the test set.
e.g. for the word 'ABSOLUTELY',

The total count of the 3 predictions = 48 + 1 + 1 = 50.

Words that are confused more than 30 times i.e. 60% of the time are highlighted in red.
Words with prediction accuracy (WAR) of 90% or more are highlighted in green.

Table 10: *Top-5 predictions per label*

Word	**(Prediction, Count)**				
	Prediction1	**Prediction2**	**Prediction3**	**Prediction4**	**Prediction5**
ABOUT	('ABOUT', 31)	('AMONG', 3)	('AMOUNT', 3)	('AMERICAN', 2)	('DEBATE', 2)
ABSOLUTELY	('ABSOLUTELY', 48)	('HAPPEN', 1)	('TEMPERATURES', 1)		
ABUSE	('ABUSE', 45)	('COMMUNITY', 3)	('BUILD', 1)	('REPORTS', 1)	
ACCESS	('ACCESS', 47)	('CASES', 3)			
ACCORDING	('ACCORDING', 47)	('COURT', 1)	('INCREASE', 1)	('WHOLE', 1)	
ACCUSED	('ACCUSED', 50)				
ACROSS	('ACROSS', 45)	('ANOTHER', 1)	('GROWTH', 1)	('IRELAND', 1)	('LOCAL', 1)
ACTION	('ACTION', 37)	('COUNCIL', 2)	('NATIONAL', 2)	('ALLEGATIONS', 1)	('AMONG', 1)
ACTUALLY	('ACTUALLY', 35)	('GETTING', 3)	('ASKING', 2)	('ASKED', 1)	('CHANGES', 1)

AFFAIRS	('AFFAIRS', 44)	('FIGHT', 3)	('FACING', 1)	('FIGURES', 1)	('RIGHTS', 1)
AFFECTED	('AFFECTED', 42)	('EVENTS', 3)	('FIGHTING', 2)	('FIGHT', 1)	('FOUND', 1)
AFRICA	('AFRICA', 45)	('EVERYTHING', 2)	('CONFERENCE', 1)	('EVERY', 1)	('LIVING', 1)
AFTER	('AFTER', 47)	('EVERY', 1)	('GIVEN', 1)	('POSSIBLE', 1)	
AFTERNOON	('AFTERNOON', 50)				
AGAIN	('AGAIN', 42)	('AGAINST', 1)	('CLEAR', 1)	('EARLY', 1)	('EXACTLY', 1)
AGAINST	('AGAINST', 35)	('CANCER', 3)	('LATER', 2)	('AGAIN', 1)	('CASES', 1)
AGREE	('AGREE', 38)	('ANYTHING', 3)	('GREAT', 2)	('GREECE', 2)	('THREE', 2)
AGREEMENT	('AGREEMENT', 49)	('WOMEN', 1)			
AHEAD	('AHEAD', 41)	('ATTACKS', 2)	('ENGLAND', 1)	('GETTING', 1)	('GOING', 1)
ALLEGATIONS	('ALLEGATIONS', 50)				
ALLOW	('ALLOW', 35)	('ALLOWED', 7)	('CONTINUE', 2)	('DETAILS', 1)	('SHOULD', 1)
ALLOWED	('ALLOWED', 42)	('ANNOUNCED', 2)	('HOUSE', 2)	('ALLOW', 1)	('BELIEVE', 1)
ALMOST	('ALMOST', 44)	('COMES', 2)	('ALWAYS', 1)	('FORWARD', 1)	('WANTS', 1)
ALREADY	('ALREADY', 41)	('ALWAYS', 1)	('DEGREES', 1)	('GREAT', 1)	('LOOKING', 1)
ALWAYS	('ALWAYS', 42)	('AUTHORITIES', 1)	('COMING', 1)	('COURSE', 1)	('FORWARD', 1)
AMERICA	('AMERICA', 38)	('AMERICAN', 10)	('ASKED', 1)	('RETURN', 1)	

AMERICAN	('AMERICAN', 43)	('AMERICA', 2)	('AMONG', 1)	('BUILDING', 1)	('MAKING', 1)
AMONG	('AMONG', 44)	('ALMOST', 1)	('AMERICA', 1)	('BEING', 1)	('CAMPAIGN', 1)
AMOUNT	('AMOUNT', 43)	('ABOUT', 2)	('MATTER', 2)	('AMONG', 1)	('MAKES', 1)
ANNOUNCED	('ANNOUNCED', 42)	('ALLOWED', 3)	('COUNCIL', 3)	('AGAINST', 1)	('KNOWN', 1)
ANOTHER	('ANOTHER', 38)	('SOUTHERN', 3)	('BUILDING', 1)	('ENGLAND', 1)	('ENOUGH', 1)
ANSWER	('ANSWER', 29)	('ASKED', 4)	('ACTION', 2)	('ANYTHING', 2)	('ASKING', 2)
ANYTHING	('ANYTHING', 47)	('CONTINUE', 1)	('LIVING', 1)	('OTHER', 1)	
AREAS	('AREAS', 45)	('HOURS', 2)	('LEVELS', 2)	('SERIES', 1)	
AROUND	('AROUND', 38)	('GROUND', 4)	('ABOUT', 1)	('GROWTH', 1)	('HIGHER', 1)
ARRESTED	('ARRESTED', 45)	('ASKING', 1)	('CRISIS', 1)	('REASON', 1)	('RECENT', 1)
ASKED	('ASKED', 27)	('CANCER', 5)	('ASKING', 3)	('HEART', 3)	('ACTION', 2)
ASKING	('ASKING', 44)	('ACCESS', 2)	('ANSWER', 1)	('GETTING', 1)	('PATIENTS', 1)
ATTACK	('ATTACK', 45)	('EXACTLY', 2)	('SAYING', 1)	('STAND', 1)	('TODAY', 1)
ATTACKS	('ATTACKS', 46)	('ASKING', 1)	('DECIDED', 1)	('STATES', 1)	('VIOLENCE', 1)
AUTHORITIES	('AUTHORITIES', 49)	('AGREE', 1)			
BANKS	('BANKS', 42)	('PLANS', 3)	('PLACE', 2)	('DEBATE', 1)	('MAKES', 1)
BECAUSE	('BECAUSE', 33)	('ABUSE', 5)	('BEING', 2)	('MEANS', 2)	('ABSOLUTELY', 1)
BECOME	('BECOME', 47)	('BECAUSE', 2)	('PARLIAMENT', 1)		

BEFORE	('BEFORE', 50)				
BEHIND	('BEHIND', 45)	('PERSON', 2)	('NIGHT', 1)	('PARTY', 1)	('PAYING', 1)
BEING	('BEING', 34)	('PAYING', 4)	('BIGGEST', 3)	('MIDDLE', 2)	('BETTER', 1)
BELIEVE	('BELIEVE', 45)	('ENOUGH', 1)	('GIVEN', 1)	('PERIOD', 1)	('POLICE', 1)
BENEFIT	('BENEFIT', 44)	('BENEFITS', 5)	('IMPACT', 1)		
BENEFITS	('BENEFITS', 43)	('BENEFIT', 7)			
BETTER	('BETTER', 35)	('ABOUT', 2)	('MESSAGE', 2)	('BANKS', 1)	('BECOME', 1)
BETWEEN	('BETWEEN', 49)	('THREE', 1)			
BIGGEST	('BIGGEST', 41)	('BUILD', 2)	('BANKS', 1)	('BILLION', 1)	('BRITISH', 1)
BILLION	('BILLION', 42)	('MILLION', 4)	('BUILDING', 2)	('BEING', 1)	('PAYING', 1)
BLACK	('BLACK', 44)	('BANKS', 1)	('BEING', 1)	('IMPACT', 1)	('PLANS', 1)
BORDER	('BORDER', 37)	('IMPORTANT', 5)	('BROUGHT', 3)	('POINT', 2)	('POLICY', 1)
BRING	('BRING', 37)	('BEING', 3)	('BRITAIN', 2)	('EUROPEAN', 2)	('PRETTY', 2)
BRITAIN	('BRITAIN', 30)	('BRING', 7)	('PRISON', 6)	('BRITISH', 1)	('MEETING', 1)
BRITISH	('BRITISH', 41)	('BIGGEST', 1)	('BRING', 1)	('BRITAIN', 1)	('PRESIDENT', 1)
BROUGHT	('BROUGHT', 37)	('IMPORTANT', 4)	('BORDER', 3)	('POINT', 2)	('BETWEEN', 1)
BUDGET	('BUDGET', 47)	('BUSINESS', 1)	('PARTY', 1)	('PROCESS', 1)	
BUILD	('BUILD', 44)	('BEING', 1)	('BRING', 1)	('BUILDING', 1)	('BUSINESS', 1)
BUILDING	('BUILDING', 45)	('AMERICAN', 1)	('BECAUSE', 1)	('BILLION', 1)	('BUILD', 1)

BUSINESS	('BUSINESS', 40)	('BUSINESSES', 2)	('MESSAGE', 2)	('MINISTERS', 2)	('MIDDLE', 1)
BUSINESSES	('BUSINESSES', 45)	('BUSINESS', 4)	('DIFFERENCE', 1)		
CALLED	('CALLED', 36)	('COURT', 3)	('COURSE', 2)	('WHOLE', 2)	('ACCORDING', 1)
CAMERON	('CAMERON', 50)				
CAMPAIGN	('CAMPAIGN', 48)	('HAPPEN', 1)	('IMPACT', 1)		
CANCER	('CANCER', 38)	('ANSWER', 3)	('CASES', 3)	('ACTION', 2)	('COUNCIL', 2)
CANNOT	('CANNOT', 44)	('YOUNG', 2)	('ECONOMIC', 1)	('HEALTH', 1)	('IRELAND', 1)
CAPITAL	('CAPITAL', 41)	('COUPLE', 3)	('HAPPENING', 3)	('CAMERON', 1)	('HAPPEN', 1)
CASES	('CASES', 45)	('CANCER', 1)	('CHANGES', 1)	('LATEST', 1)	('PRICES', 1)
CENTRAL	('CENTRAL', 43)	('SECRETARY', 2)	('SINGLE', 2)	('CHANGES', 1)	('GENERAL', 1)
CERTAINLY	('CERTAINLY', 43)	('SUNDAY', 2)	('LIKELY', 1)	('SINCE', 1)	('TAKING', 1)
CHALLENGE	('CHALLENGE', 41)	('CHANGE', 8)	('SENSE', 1)		
CHANCE	('CHANCE', 34)	('CHARGE', 2)	('JUDGE', 2)	('SENSE', 2)	('ANSWER', 1)
CHANGE	('CHANGE', 40)	('JUDGE', 3)	('CHANCE', 2)	('CHANGES', 1)	('EDUCATION', 1)
CHANGES	('CHANGES', 47)	('CHANCE', 2)	('CHANGE', 1)		
CHARGE	('CHARGE', 35)	('START', 6)	('CHANCE', 2)	('JUDGE', 2)	('STARTED', 2)
CHARGES	('CHARGES', 41)	('CHANCE', 2)	('CHANGES', 2)	('CHILDREN', 2)	('CHARGE', 1)
CHIEF	('CHIEF', 49)	('GIVEN', 1)			

CHILD	('CHILD', 38)	('CHANCE', 2)	('TRADE', 2)	('CHALLENGE', 1)	('CHANGE', 1)
CHILDREN	('CHILDREN', 46)	('CHALLENGE', 1)	('SECURITY', 1)	('SHORT', 1)	('SINGLE', 1)
CHINA	('CHINA', 43)	('CHANGES', 2)	('CERTAINLY', 1)	('CHALLENGE', 1)	('DECIDED', 1)
CLAIMS	('CLAIMS', 46)	('COMES', 1)	('GAMES', 1)	('JAMES', 1)	('PERHAPS', 1)
CLEAR	('CLEAR', 41)	('HIGHER', 2)	('AGAIN', 1)	('ENGLAND', 1)	('FIGURES', 1)
CLOSE	('CLOSE', 39)	('ALLOWED', 2)	('ACCUSED', 1)	('ASKED', 1)	('BECAUSE', 1)
CLOUD	('CLOUD', 46)	('CLOSE', 2)	('ALLOWED', 1)	('LABOUR', 1)	
COMES	('COMES', 45)	('GAMES', 1)	('HOMES', 1)	('SEEMS', 1)	('TERMS', 1)
COMING	('COMING', 47)	('COMES', 1)	('GLOBAL', 1)	('SOMEONE', 1)	
COMMUNITY	('COMMUNITY', 47)	('FORWARD', 1)	('SIMPLY', 1)	('WRONG', 1)	
COMPANIES	('COMPANIES', 44)	('COMPANY', 4)	('COMES', 1)	('NUMBER', 1)	
COMPANY	('COMPANY', 41)	('COMPANIES', 5)	('COMING', 2)	('COUPLE', 1)	('MEMBER', 1)
CONCERNS	('CONCERNS', 49)	('INTEREST', 1)			
CONFERENCE	('CONFERENCE', 46)	('CONFLICT', 2)	('AUTHORITIES', 1)	('OFTEN', 1)	
CONFLICT	('CONFLICT', 49)	('EVERYTHING', 1)			
CONSERVATIVE	('CONSERVATIVE', 47)	('CHIEF', 1)	('GERMANY', 1)	('TERMS', 1)	
CONTINUE	('CONTINUE', 42)	('EXACTLY', 1)	('ISSUE', 1)	('LARGE', 1)	('ORDER', 1)

CONTROL	('CONTROL', 49)	('ECONOMIC', 1)			
COULD	('COULD', 24)	('ECONOMIC', 3)	('SHOULD', 3)	('CALLED', 2)	('SECURITY', 2)
COUNCIL	('COUNCIL', 41)	('AFTER', 2)	('ANNOUNCED', 2)	('ALLOWED', 1)	('ANSWER', 1)
COUNTRIES	('COUNTRIES', 44)	('AGREE', 1)	('CANNOT', 1)	('CHANGES', 1)	('COMPANIES', 1)
COUNTRY	('COUNTRY', 45)	('COUNTRIES', 2)	('CENTRAL', 1)	('INDUSTRY', 1)	('LARGE', 1)
COUPLE	('COUPLE', 47)	('CAPITAL', 1)	('COMING', 1)	('ISLAMIC', 1)	
COURSE	('COURSE', 45)	('COURT', 1)	('FORCES', 1)	('QUITE', 1)	('TALKS', 1)
COURT	('COURT', 36)	('CALLED', 3)	('COURSE', 3)	('ORDER', 2)	('SCHOOLS', 2)
CRIME	('CRIME', 47)	('AGREEMENT', 1)	('GREAT', 1)	('RIGHT', 1)	
CRISIS	('CRISIS', 49)	('GREAT', 1)			
CURRENT	('CURRENT', 48)	('LEVELS', 1)	('SCOTLAND', 1)		
CUSTOMERS	('CUSTOMERS', 47)	('NEEDS', 2)	('SITUATION', 1)		
DAVID	('DAVID', 47)	('CONSERVATIVE', 1)	('GIVEN', 1)	('HAVING', 1)	
DEATH	('DEATH', 40)	('AGAINST', 1)	('ANOTHER', 1)	('ATTACK', 1)	('EXTRA', 1)
DEBATE	('DEBATE', 39)	('SPENT', 4)	('SPEND', 2)	('EXPECTED', 1)	('INDEPENDENT', 1)
DECIDED	('DECIDED', 41)	('STARTED', 5)	('BUSINESSES', 1)	('INSIDE', 1)	('SOCIETY', 1)
DECISION	('DECISION', 48)	('CENTRAL', 1)	('CONTINUE', 1)		
DEFICIT	('DEFICIT', 48)	('EVIDENCE', 1)	('NEVER', 1)		

DEGREES	('DEGREES', 48)	('AFTERNOON', 1)	('GREECE', 1)		
DESCRIBED	('DESCRIBED', 49)	('CRIME', 1)			
DESPITE	('DESPITE', 47)	('EXPECT', 2)	('SPEAKING', 1)		
DETAILS	('DETAILS', 47)	('CONCERNS', 1)	('SENIOR', 1)	('SENSE', 1)	
DIFFERENCE	('DIFFERENCE', 43)	('DIFFERENT', 5)	('DEFICIT', 1)	('FOCUS', 1)	
DIFFERENT	('DIFFERENT', 35)	('DIFFERENCE', 6)	('DEFICIT', 2)	('DIFFICULT', 2)	('SEVERAL', 2)
DIFFICULT	('DIFFICULT', 48)	('DIFFERENCE', 1)	('EXPECT', 1)		
DOING	('DOING', 37)	('SYRIA', 2)	('BETWEEN', 1)	('CONTINUE', 1)	('DURING', 1)
DURING	('DURING', 40)	('CONTROL', 1)	('DOING', 1)	('EASTERN', 1)	('GOING', 1)
EARLY	('EARLY', 40)	('CLEAR', 2)	('CASES', 1)	('HEARD', 1)	('HIGHER', 1)
EASTERN	('EASTERN', 43)	('YESTERDAY', 2)	('CENTRAL', 1)	('FIGURES', 1)	('LEADER', 1)
ECONOMIC	('ECONOMIC', 47)	('HUMAN', 1)	('ISLAMIC', 1)	('LEADER', 1)	
ECONOMY	('ECONOMY', 48)	('CANNOT', 1)	('EUROPE', 1)		
EDITOR	('EDITOR', 41)	('CANCER', 2)	('ARRESTED', 1)	('BANKS', 1)	('LATEST', 1)
EDUCATION	('EDUCATION', 47)	('ELECTION', 1)	('SECTOR', 1)	('UNITED', 1)	
ELECTION	('ELECTION', 43)	('ACTION', 4)	('ACTUALLY', 1)	('ALLEGATIONS', 1)	('RUSSIAN', 1)
EMERGENCY	('EMERGENCY', 47)	('ASKING', 1)	('COMMUNITY', 1)	('SPECIAL', 1)	
ENERGY	('ENERGY', 43)	('ANYTHING', 2)	('EDITOR', 2)	('ACTION', 1)	('JUSTICE', 1)

ENGLAND	('ENGLAND', 45)	('IRELAND', 1)	('LATEST', 1)	('LEADERS', 1)	('LONDON', 1)
ENOUGH	('ENOUGH', 44)	('BEING', 1)	('LEAVE', 1)	('LIVES', 1)	('NOTHING', 1)
EUROPE	('EUROPE', 49)	('PERIOD', 1)			
EUROPEAN	('EUROPEAN', 48)	('DEBATE', 1)	('REMEMBER', 1)		
EVENING	('EVENING', 46)	('EVENTS', 1)	('GIVING', 1)	('HEAVY', 1)	('LIVING', 1)
EVENTS	('EVENTS', 45)	('AFFECTED', 2)	('VIOLENCE', 2)	('FRENCH', 1)	
EVERY	('EVERY', 41)	('HAVING', 3)	('DAVID', 2)	('EVERYTHING', 1)	('HEAVY', 1)
EVERYBODY	('EVERYBODY', 48)	('EVERYONE', 1)	('FOOTBALL', 1)		
EVERYONE	('EVERYONE', 49)	('ALWAYS', 1)			
EVERYTHING	('EVERYTHING', 46)	('AFRICA', 1)	('EVERYONE', 1)	('LEVEL', 1)	('THINK', 1)
EVIDENCE	('EVIDENCE', 48)	('AREAS', 1)	('EVERYTHING', 1)		
EXACTLY	('EXACTLY', 42)	('ATTACK', 1)	('DEATH', 1)	('DECIDED', 1)	('GETTING', 1)
EXAMPLE	('EXAMPLE', 49)	('COUPLE', 1)			
EXPECT	('EXPECT', 43)	('EXPECTED', 2)	('INDEPENDENT', 2)	('AGAIN', 1)	('SPEND', 1)
EXPECTED	('EXPECTED', 47)	('INDEPENDENT', 1)	('SPEND', 1)	('SPENDING', 1)	
EXTRA	('EXTRA', 45)	('HISTORY', 2)	('COUNTRY', 1)	('NATIONAL', 1)	('SECTOR', 1)
FACING	('FACING', 44)	('FIGHTING', 6)			

FAMILIES	('FAMILIES', 48)	('FAMILY', 2)			
FAMILY	('FAMILY', 47)	('FAMILIES', 2)	('AMONG', 1)		
FIGHT	('FIGHT', 38)	('FIGHTING', 4)	('FRONT', 3)	('AFFECTED', 2)	('AFRICA', 1)
FIGHTING	('FIGHTING', 43)	('FACING', 2)	('FIGHT', 2)	('FINAL', 2)	('RUNNING', 1)
FIGURES	('FIGURES', 43)	('AFFAIRS', 3)	('FRANCE', 1)	('FRENCH', 1)	('INFLATION', 1)
FINAL	('FINAL', 46)	('FIGHTING', 1)	('FIGURES', 1)	('FRIDAY', 1)	('FRONT', 1)
FINANCIAL	('FINANCIAL', 46)	('NATIONAL', 2)	('FINAL', 1)	('VIOLENCE', 1)	
FIRST	('FIRST', 39)	('FIGURES', 2)	('FRONT', 2)	('FURTHER', 2)	('FORCE', 1)
FOCUS	('FOCUS', 43)	('FORCE', 2)	('VOTERS', 2)	('FORCES', 1)	('OFFICIALS', 1)
FOLLOWING	('FOLLOWING', 50)				
FOOTBALL	('FOOTBALL', 48)	('CAMPAIGN', 1)	('FORMER', 1)		
FORCE	('FORCE', 43)	('FORCES', 3)	('FORWARD', 3)	('RULES', 1)	
FORCES	('FORCES', 45)	('FORCE', 2)	('AUTHORITIES', 1)	('FIGHTING', 1)	('VOTERS', 1)
FOREIGN	('FOREIGN', 47)	('BEFORE', 1)	('FRONT', 1)	('VOTERS', 1)	
FORMER	('FORMER', 47)	('HUMAN', 2)	('FOOTBALL', 1)		
FORWARD	('FORWARD', 43)	('FORCE', 3)	('ALWAYS', 1)	('FORMER', 1)	('PHONE', 1)
FOUND	('FOUND', 46)	('FAMILY', 1)	('FIGHTING', 1)	('NOTHING', 1)	('PHONE', 1)
FRANCE	('FRANCE', 43)	('FRONT', 2)	('FIRST', 1)	('INFLATION', 1)	('SCOTLAND', 1)
FRENCH	('FRENCH', 48)	('FRONT', 1)	('RUSSIAN', 1)		

FRIDAY	('FRIDAY', 46)	('FINAL', 3)	('FIGHTING', 1)		
FRONT	('FRONT', 40)	('FORMER', 2)	('FRANCE', 2)	('DIFFERENT', 1)	('FIGHT', 1)
FURTHER	('FURTHER', 45)	('FOREIGN', 1)	('OFFICERS', 1)	('RATHER', 1)	('TRUST', 1)
FUTURE	('FUTURE', 48)	('TRIAL', 1)	('VOTERS', 1)		
GAMES	('GAMES', 43)	('TIMES', 3)	('CAPITAL', 1)	('COMES', 1)	('STATEMENT', 1)
GENERAL	('GENERAL', 44)	('CENTRAL', 2)	('SEVERAL', 2)	('CHILD', 1)	('SOMEONE', 1)
GEORGE	('GEORGE', 38)	('TALKS', 6)	('SHORT', 2)	('FORWARD', 1)	('SCHOOLS', 1)
GERMANY	('GERMANY', 50)				
GETTING	('GETTING', 30)	('TAKING', 2)	('AGAIN', 1)	('ASKING', 1)	('CERTAINLY', 1)
GIVEN	('GIVEN', 33)	('GIVING', 7)	('DIFFERENT', 2)	('LEVEL', 2)	('DAVID', 1)
GIVING	('GIVING', 33)	('GIVEN', 6)	('LIVING', 4)	('EVENING', 3)	('NEVER', 2)
GLOBAL	('GLOBAL', 47)	('COULD', 1)	('COUPLE', 1)	('REPORT', 1)	
GOING	('GOING', 32)	('LONGER', 2)	('ACCUSED', 1)	('AUTHORITIES', 1)	('CERTAINLY', 1)
GOVERNMENT	('GOVERNMENT', 48)	('COMPANY', 1)	('NEVER', 1)		
GREAT	('GREAT', 32)	('AGREE', 3)	('QUITE', 3)	('THREAT', 2)	('AGAIN', 1)
GREECE	('GREECE', 47)	('INCREASE', 3)			
GROUND	('GROUND', 39)	('AROUND', 8)	('FOUND', 1)	('ITSELF', 1)	('SOCIETY', 1)
GROUP	('GROUP', 44)	('EUROPE', 2)	('ECONOMIC', 1)	('FOOTBALL', 1)	('GERMANY', 1)

GROWING	('GROWING', 47)	('GOING', 1)	('GROWTH', 1)	('ORDER', 1)	
GROWTH	('GROWTH', 48)	('LATER', 1)	('NORTH', 1)		
GUILTY	('GUILTY', 46)	('ACCUSED', 1)	('CHILDREN', 1)	('HOUSING', 1)	('KILLED', 1)
HAPPEN	('HAPPEN', 37)	('HAPPENED', 7)	('CAPITAL', 1)	('COMING', 1)	('COUPLE', 1)
HAPPENED	('HAPPENED', 39)	('HAPPEN', 3)	('HAPPENING', 3)	('CAPITAL', 2)	('COMES', 1)
HAPPENING	('HAPPENING', 48)	('COMING', 1)	('SOMETHING', 1)		
HAVING	('HAVING', 42)	('AFTER', 1)	('AMERICAN', 1)	('COUNCIL', 1)	('CURRENT', 1)
HEALTH	('HEALTH', 42)	('ENGLAND', 2)	('ANNOUNCED', 1)	('COUNCIL', 1)	('CURRENT', 1)
HEARD	('HEARD', 27)	('EARLY', 2)	('FURTHER', 2)	('SENIOR', 2)	('ANOTHER', 1)
HEART	('HEART', 44)	('BECAUSE', 1)	('BEHIND', 1)	('HIGHER', 1)	('INVOLVED', 1)
HEAVY	('HEAVY', 40)	('HAVING', 6)	('AFTER', 1)	('EVERYTHING', 1)	('LEAVE', 1)
HIGHER	('HIGHER', 39)	('BEHIND', 2)	('AHEAD', 1)	('ATTACK', 1)	('CHINA', 1)
HISTORY	('HISTORY', 44)	('COUNTRY', 1)	('EXTRA', 1)	('FINANCIAL', 1)	('ISSUES', 1)
HOMES	('HOMES', 44)	('ALMOST', 2)	('TERMS', 2)	('COMES', 1)	('TIMES', 1)
HOSPITAL	('HOSPITAL', 48)	('GOING', 1)	('POSSIBLE', 1)		
HOURS	('HOURS', 40)	('HOUSE', 3)	('AGAINST', 1)	('AREAS', 1)	('COUNCIL', 1)
HOUSE	('HOUSE', 33)	('HOURS', 4)	('ANNOUNCED', 3)	('COUNCIL', 2)	('HOUSING', 2)
HOUSING	('HOUSING', 44)	('HOUSE', 3)	('ACTUALLY', 2)	('ANNOUNCED', 1)	

HUMAN	('HUMAN', 47)	('GLOBAL', 2)	('COMING', 1)		
HUNDREDS	('HUNDREDS', 48)	('EXTRA', 1)	('THOUSANDS', 1)		
IMMIGRATION	('IMMIGRATION', 50)				
IMPACT	('IMPACT', 44)	('AMOUNT', 1)	('BANKS', 1)	('CAMPAIGN', 1)	('PARTS', 1)
IMPORTANT	('IMPORTANT', 41)	('POINT', 3)	('BRING', 1)	('MORNING', 1)	('PARTS', 1)
INCREASE	('INCREASE', 49)	('SERIES', 1)			
INDEPENDENT	('INDEPENDENT', 44)	('SPENT', 2)	('BETTER', 1)	('DESPITE', 1)	('EXPECTED', 1)
INDUSTRY	('INDUSTRY', 44)	('HISTORY', 3)	('DOING', 1)	('LEADERSHIP', 1)	('SHOULD', 1)
INFLATION	('INFLATION', 48)	('EVENTS', 1)	('FRENCH', 1)		
INFORMATION	('INFORMATION', 50)				
INQUIRY	('INQUIRY', 49)	('WHERE', 1)			
INSIDE	('INSIDE', 45)	('CONCERNS', 2)	('DECIDED', 2)	('YESTERDAY', 1)	
INTEREST	('INTEREST', 44)	('ACROSS', 1)	('COUNTRIES', 1)	('SITUATION', 1)	('STREET', 1)
INVESTMENT	('INVESTMENT', 49)	('FACING', 1)			
INVOLVED	('INVOLVED', 48)	('DIFFICULT', 1)	('GROWTH', 1)		
IRELAND	('IRELAND', 48)	('EARLY', 1)	('VIOLENCE', 1)		
ISLAMIC	('ISLAMIC', 50)				

ISSUE	('ISSUE', 43)	('ACTION', 2)	('ACTUALLY', 1)	('COUNCIL', 1)	('DESCRIBED', 1)
ISSUES	('ISSUES', 42)	('ISSUE', 3)	('ACTION', 1)	('ELECTION', 1)	('RESULT', 1)
ITSELF	('ITSELF', 46)	('CONSERVATIVE', 1)	('ENOUGH', 1)	('SOUTH', 1)	('STAFF', 1)
JAMES	('JAMES', 47)	('CHANCE', 1)	('CONSERVATIVE', 1)	('GAMES', 1)	
JUDGE	('JUDGE', 40)	('CHANCE', 2)	('CHANGE', 2)	('CHALLENGE', 1)	('GEORGE', 1)
JUSTICE	('JUSTICE', 45)	('ANSWER', 1)	('CHANGES', 1)	('DECISION', 1)	('THINGS', 1)
KILLED	('KILLED', 43)	('ENGLAND', 2)	('HEALTH', 1)	('SINCE', 1)	('SINGLE', 1)
KNOWN	('KNOWN', 37)	('THOSE', 2)	('CONSERVATIVE', 1)	('CURRENT', 1)	('ENOUGH', 1)
LABOUR	('LABOUR', 46)	('COMING', 1)	('COUPLE', 1)	('FAMILY', 1)	('MAYBE', 1)
LARGE	('LARGE', 44)	('CHARGE', 1)	('CLOSE', 1)	('COUNCIL', 1)	('GIVING', 1)
LATER	('LATER', 39)	('LATEST', 2)	('ARRESTED', 1)	('CANCER', 1)	('CENTRAL', 1)
LATEST	('LATEST', 47)	('CASES', 1)	('LATER', 1)	('LEADERS', 1)	
LEADER	('LEADER', 41)	('CLEAR', 2)	('GETTING', 1)	('IRELAND', 1)	('LEADERS', 1)
LEADERS	('LEADERS', 37)	('LEADER', 3)	('LEAST', 2)	('STATES', 2)	('AREAS', 1)
LEADERSHIP	('LEADERSHIP', 49)	('ENERGY', 1)			
LEAST	('LEAST', 36)	('LEADER', 4)	('NEEDS', 3)	('LEADERS', 2)	('EASTERN', 1)
LEAVE	('LEAVE', 46)	('CLEAR', 1)	('GIVEN', 1)	('NEVER', 1)	('SIGNIFICANT', 1)

LEGAL	('LEGAL', 44)	('CLEAR', 1)	('FINANCIAL', 1)	('HEALTH', 1)	('KILLED', 1)
LEVEL	('LEVEL', 41)	('NEVER', 3)	('DAVID', 1)	('EVERYTHING', 1)	('GIVING', 1)
LEVELS	('LEVELS', 47)	('LEVEL', 2)	('HEAVY', 1)		
LIKELY	('LIKELY', 44)	('AGREE', 1)	('CERTAINLY', 1)	('HIGHER', 1)	('LITTLE', 1)
LITTLE	('LITTLE', 30)	('LEGAL', 7)	('CLEAR', 3)	('ALLOW', 1)	('GUILTY', 1)
LIVES	('LIVES', 47)	('ENOUGH', 1)	('HEALTH', 1)	('LIVING', 1)	
LIVING	('LIVING', 36)	('GIVING', 5)	('EVENING', 2)	('SIGNIFICANT', 2)	('ANYTHING', 1)
LOCAL	('LOCAL', 42)	('ACCUSED', 1)	('CERTAINLY', 1)	('CLOSE', 1)	('COURT', 1)
LONDON	('LONDON', 36)	('GETTING', 2)	('AGAINST', 1)	('ANOTHER', 1)	('CHINA', 1)
LONGER	('LONGER', 37)	('EUROPE', 2)	('CHILDREN', 1)	('COULD', 1)	('ECONOMY', 1)
LOOKING	('LOOKING', 49)	('BUILDING', 1)			
MAJOR	('MAJOR', 43)	('MAKES', 2)	('MATTER', 2)	('MILITARY', 2)	('MEASURES', 1)
MAJORITY	('MAJORITY', 46)	('CHANCE', 2)	('MILITARY', 1)	('STORY', 1)	
MAKES	('MAKES', 35)	('BETTER', 5)	('MAJOR', 2)	('AMOUNT', 1)	('BANKS', 1)
MAKING	('MAKING', 45)	('MEDICAL', 1)	('MILLION', 1)	('SPEAKING', 1)	('SPENDING', 1)
MANCHESTER	('MANCHESTER', 45)	('MESSAGE', 2)	('CONTINUE', 1)	('MEASURES', 1)	('OPPOSITION', 1)
MARKET	('MARKET', 42)	('BUILDING', 2)	('AMERICA', 1)	('AMERICAN', 1)	('BEHIND', 1)
MASSIVE	('MASSIVE', 48)	('MEETING', 1)	('PRESIDENT', 1)		

MATTER	('MATTER', 33)	('BETTER', 3)	('MIGHT', 3)	('AMOUNT', 2)	('ABOUT', 1)
MAYBE	('MAYBE', 46)	('MEMBER', 2)	('AMERICAN', 1)	('FAMILY', 1)	
MEANS	('MEANS', 39)	('MAKES', 3)	('MINUTES', 2)	('ALMOST', 1)	('BECAUSE', 1)
MEASURES	('MEASURES', 47)	('MANCHESTER', 1)	('PATIENTS', 1)	('SPECIAL', 1)	
MEDIA	('MEDIA', 37)	('MAKING', 2)	('MEETING', 2)	('BETTER', 1)	('MAJOR', 1)
MEDICAL	('MEDICAL', 44)	('MAKING', 3)	('MIDDLE', 2)	('BETTER', 1)	
MEETING	('MEETING', 36)	('MEDIA', 5)	('MISSING', 3)	('BRITAIN', 1)	('BUDGET', 1)
MEMBER	('MEMBER', 48)	('MEMBERS', 1)	('REMEMBER', 1)		
MEMBERS	('MEMBERS', 48)	('AMOUNT', 1)	('MEMBER', 1)		
MESSAGE	('MESSAGE', 42)	('BUSINESS', 3)	('MEASURES', 2)	('MAKING', 1)	('MANCHESTER', 1)
MIDDLE	('MIDDLE', 37)	('MEDICAL', 3)	('BUILD', 2)	('AMERICAN', 1)	('BECAUSE', 1)
MIGHT	('MIGHT', 34)	('AMONG', 3)	('MAYBE', 3)	('MAKES', 2)	('ALMOST', 1)
MIGRANTS	('MIGRANTS', 48)	('AMERICAN', 1)	('MIGHT', 1)		
MILITARY	('MILITARY', 49)	('MEDIA', 1)			
MILLION	('MILLION', 42)	('BILLION', 4)	('AMERICA', 1)	('COMING', 1)	('MAJOR', 1)
MILLIONS	('MILLIONS', 47)	('MIDDLE', 1)	('MINUTES', 1)	('PLACE', 1)	
MINISTER	('MINISTER', 43)	('SPEECH', 2)	('LEADER', 1)	('MEANS', 1)	('MILITARY', 1)
MINISTERS	('MINISTERS', 46)	('MEANS', 1)	('MEASURES', 1)	('MEDICAL', 1)	('MINISTER', 1)

MINUTES	('MINUTES', 41)	('MAKES', 4)	('EVENTS', 1)	('MEANS', 1)	('MILLIONS', 1)
MISSING	('MISSING', 43)	('MEETING', 3)	('MILLION', 2)	('BETTER', 1)	('MEDIA', 1)
MOMENT	('MOMENT', 48)	('MAYBE', 2)			
MONEY	('MONEY', 42)	('MAKING', 2)	('AMERICA', 1)	('BRITISH', 1)	('BUILDING', 1)
MONTH	('MONTH', 44)	('MONTHS', 3)	('FAMILIES', 1)	('MAKING', 1)	('MIGHT', 1)
MONTHS	('MONTHS', 45)	('BETTER', 1)	('MAKES', 1)	('MEANS', 1)	('MONTH', 1)
MORNING	('MORNING', 48)	('AMERICAN', 1)	('POINT', 1)		
MOVING	('MOVING', 49)	('MONTH', 1)			
MURDER	('MURDER', 41)	('RIGHT', 2)	('BUILD', 1)	('EMERGENCY', 1)	('MARKET', 1)
NATIONAL	('NATIONAL', 44)	('ELECTION', 3)	('COUNTRY', 1)	('DEATH', 1)	('LITTLE', 1)
NEEDS	('NEEDS', 35)	('CANCER', 2)	('THESE', 2)	('YEARS', 2)	('ALLEGATIONS', 1)
NEVER	('NEVER', 38)	('SEVEN', 3)	('AFTER', 1)	('ENOUGH', 1)	('FOUND', 1)
NIGHT	('NIGHT', 35)	('TONIGHT', 3)	('UNITED', 2)	('CERTAINLY', 1)	('EARLY', 1)
NORTH	('NORTH', 42)	('COURT', 2)	('THOUGHT', 2)	('ACROSS', 1)	('HEARD', 1)
NORTHERN	('NORTHERN', 46)	('NORTH', 1)	('NOTHING', 1)	('ORDER', 1)	('STORY', 1)
NOTHING	('NOTHING', 43)	('ANOTHER', 2)	('DIFFERENT', 1)	('EDUCATION', 1)	('LIKELY', 1)
NUMBER	('NUMBER', 37)	('NUMBERS', 6)	('COMPANY', 3)	('COMES', 1)	('COUPLE', 1)
NUMBERS	('NUMBERS', 48)	('PERSON', 1)	('TERMS', 1)		

OBAMA	('OBAMA', 49)	('MOMENT', 1)			
OFFICE	('OFFICE', 45)	('CONFERENCE', 1)	('EVENTS', 1)	('OFFICERS', 1)	('SERVICE', 1)
OFFICERS	('OFFICERS', 48)	('OFFICIALS', 1)	('SERVICES', 1)		
OFFICIALS	('OFFICIALS', 49)	('CONTROL', 1)			
OFTEN	('OFTEN', 41)	('AFTER', 2)	('CONFERENCE', 2)	('ANSWER', 1)	('AUTHORITIES', 1)
OPERATION	('OPERATION', 49)	('PRESSURE', 1)			
OPPOSITION	('OPPOSITION', 48)	('POSITION', 2)			
ORDER	('ORDER', 37)	('AUTHORITIES', 2)	('INCREASE', 2)	('ACCORDING', 1)	('ACROSS', 1)
OTHER	('OTHER', 32)	('UNDER', 3)	('NOTHING', 2)	('OTHERS', 2)	('SOUTHERN', 2)
OTHERS	('OTHERS', 43)	('EVIDENCE', 2)	('AROUND', 1)	('HAVING', 1)	('NOTHING', 1)
OUTSIDE	('OUTSIDE', 46)	('EVERYTHING', 1)	('INSIDE', 1)	('SAYING', 1)	('STAFF', 1)
PARENTS	('PARENTS', 45)	('ABOUT', 1)	('AMOUNT', 1)	('BANKS', 1)	('HAVING', 1)
PARLIAMENT	('PARLIAMENT', 48)	('BEHIND', 1)	('POWERS', 1)		
PARTIES	('PARTIES', 46)	('BEHIND', 1)	('MARKET', 1)	('PARTY', 1)	('PERSON', 1)
PARTS	('PARTS', 48)	('BRITISH', 1)	('IMPACT', 1)		
PARTY	('PARTY', 45)	('COMMUNITY', 1)	('MARKET', 1)	('PARTS', 1)	('POINT', 1)

PATIENTS	('PATIENTS', 47)	('MEASURES', 2)	('MANCHESTER', 1)		
PAYING	('PAYING', 44)	('BEING', 3)	('IMPACT', 1)	('PLANS', 1)	('POWER', 1)
PEOPLE	('PEOPLE', 48)	('BUILDING', 1)	('REMEMBER', 1)		
PERHAPS	('PERHAPS', 48)	('MEMBERS', 1)	('PROBABLY', 1)		
PERIOD	('PERIOD', 44)	('ABUSE', 2)	('BUILDING', 1)	('COMMUNITY', 1)	('POINT', 1)
PERSON	('PERSON', 34)	('PERSONAL', 4)	('PARTS', 3)	('MINISTERS', 2)	('PARTY', 2)
PERSONAL	('PERSONAL', 40)	('PERSON', 4)	('PRESIDENT', 2)	('MEDICAL', 1)	('POLICY', 1)
PHONE	('PHONE', 39)	('FURTHER', 3)	('INVOLVED', 3)	('BUSINESS', 1)	('FOREIGN', 1)
PLACE	('PLACE', 37)	('PLANS', 4)	('PLACES', 3)	('BEHIND', 2)	('BETTER', 1)
PLACES	('PLACES', 44)	('MINISTERS', 2)	('BANKS', 1)	('BECAUSE', 1)	('PERSON', 1)
PLANS	('PLANS', 44)	('BANKS', 1)	('BEHIND', 1)	('BLACK', 1)	('PARTIES', 1)
POINT	('POINT', 44)	('AMONG', 1)	('BORDER', 1)	('MORNING', 1)	('PARLIAMENT', 1)
POLICE	('POLICE', 44)	('BRITISH', 1)	('BUILDING', 1)	('PLACE', 1)	('POINT', 1)
POLICY	('POLICY', 44)	('POLITICS', 4)	('MONTHS', 1)	('PERSON', 1)	
POLITICAL	('POLITICAL', 46)	('BILLION', 2)	('BLACK', 1)	('POLITICIANS', 1)	
POLITICIANS	('POLITICIANS', 49)	('MESSAGE', 1)			
POLITICS	('POLITICS', 48)	('PARTIES', 1)	('PROCESS', 1)		
POSITION	('POSITION', 45)	('OPPOSITION',	('BUSINESS', 1)	('POTENTIAL', 1)	

		3)			
POSSIBLE	('POSSIBLE', 47)	('PLACE', 1)	('SMALL', 1)	('WELCOME', 1)	
POTENTIAL	('POTENTIAL', 50)				
POWER	('POWER', 41)	('POWERS', 4)	('ABOUT', 3)	('AMOUNT', 1)	('MARKET', 1)
POWERS	('POWERS', 42)	('PARENTS', 4)	('AMOUNT', 1)	('PARTS', 1)	('PERHAPS', 1)
PRESIDENT	('PRESIDENT', 43)	('PRESSURE', 4)	('BUSINESS', 1)	('PRESS', 1)	('WATCHING', 1)
PRESS	('PRESS', 38)	('PRICE', 7)	('BRING', 1)	('PRESSURE', 1)	('PRICES', 1)
PRESSURE	('PRESSURE', 45)	('BRITISH', 1)	('PRESIDENT', 1)	('PRESS', 1)	('PRICE', 1)
PRETTY	('PRETTY', 42)	('BRING', 4)	('BRITISH', 1)	('MORNING', 1)	('PRIME', 1)
PRICE	('PRICE', 40)	('PLACE', 3)	('PRESS', 2)	('BECAUSE', 1)	('PERSON', 1)
PRICES	('PRICES', 43)	('PRESIDENT', 2)	('BANKS', 1)	('PARTIES', 1)	('PARTY', 1)
PRIME	('PRIME', 48)	('BEING', 1)	('OBAMA', 1)		
PRISON	('PRISON', 38)	('BRITAIN', 4)	('PRESS', 2)	('PRETTY', 2)	('BRITISH', 1)
PRIVATE	('PRIVATE', 50)				
PROBABLY	('PROBABLY', 42)	('PROBLEM', 4)	('FINAL', 1)	('MAYBE', 1)	('PRIME', 1)
PROBLEM	('PROBLEM', 39)	('PROBLEMS', 8)	('OBAMA', 1)	('PROBABLY', 1)	('SOMEONE', 1)
PROBLEMS	('PROBLEMS', 47)	('PROBLEM', 2)	('PROBABLY', 1)		
PROCESS	('PROCESS', 47)	('BROUGHT', 1)	('POLITICS', 1)	('PRICES', 1)	
PROTECT	('PROTECT', 48)	('POLICE', 1)	('POSITION', 1)		

PROVIDE	('PROVIDE', 50)				
PUBLIC	('PUBLIC', 49)	('MAYBE', 1)			
QUESTION	('QUESTION', 44)	('QUESTIONS', 4)	('RUSSIAN', 1)	('WHICH', 1)	
QUESTIONS	('QUESTIONS', 49)	('ARRESTED', 1)			
QUITE	('QUITE', 42)	('BRING', 1)	('GROUND', 1)	('PERIOD', 1)	('POINT', 1)
RATES	('RATES', 37)	('GREECE', 2)	('CRISIS', 1)	('EVENTS', 1)	('FRANCE', 1)
RATHER	('RATHER', 44)	('ANOTHER', 1)	('AROUND', 1)	('GROWTH', 1)	('RIGHT', 1)
REALLY	('REALLY', 37)	('EVERYTHING', 2)	('WITHIN', 2)	('ACTUALLY', 1)	('BRING', 1)
REASON	('REASON', 43)	('RECENT', 3)	('AREAS', 1)	('GREECE', 1)	('POLICY', 1)
RECENT	('RECENT', 41)	('WESTERN', 2)	('COMMUNITY', 1)	('CRISIS', 1)	('GREECE', 1)
RECORD	('RECORD', 46)	('GROUND', 1)	('GROWTH', 1)	('INCREASE', 1)	('WALES', 1)
REFERENDUM	('REFERENDUM', 50)				
REMEMBER	('REMEMBER', 47)	('MAYBE', 1)	('MEMBER', 1)	('PUBLIC', 1)	
REPORT	('REPORT', 38)	('REPORTS', 7)	('SUPPORT', 2)	('BORDER', 1)	('BROUGHT', 1)
REPORTS	('REPORTS', 46)	('REPORT', 4)			
RESPONSE	('RESPONSE', 50)				
RESULT	('RESULT', 44)	('UNTIL', 2)	('CERTAINLY', 1)	('SCHOOLS', 1)	('STRONG', 1)
RETURN	('RETURN', 45)	('RESULT', 2)	('BETTER', 1)	('RECENT', 1)	('TRIAL', 1)

RIGHT	('RIGHT', 30)	('RIGHTS', 3)	('QUITE', 2)	('RATES', 2)	('RUNNING', 2)
RIGHTS	('RIGHTS', 34)	('RATES', 4)	('RIGHT', 2)	('ALREADY', 1)	('CRISIS', 1)
RULES	('RULES', 40)	('FORWARD', 2)	('COURT', 1)	('FUTURE', 1)	('GEORGE', 1)
RUNNING	('RUNNING', 37)	('REALLY', 2)	('RIGHT', 2)	('ALREADY', 1)	('MONEY', 1)
RUSSIA	('RUSSIA', 44)	('RUSSIAN', 4)	('RIGHTS', 1)	('WRONG', 1)	
RUSSIAN	('RUSSIAN', 42)	('RUSSIA', 5)	('RECENT', 1)	('TRUST', 1)	('WATCHING', 1)
SAYING	('SAYING', 27)	('ATTACK', 2)	('DEATH', 2)	('STAND', 2)	('STATE', 2)
SCHOOL	('SCHOOL', 41)	('SCHOOLS', 2)	('TALKING', 2)	('TALKS', 2)	('KNOWN', 1)
SCHOOLS	('SCHOOLS', 43)	('SCHOOL', 3)	('CALLED', 1)	('SHORT', 1)	('TALKS', 1)
SCOTLAND	('SCOTLAND', 48)	('CERTAINLY', 1)	('GOING', 1)		
SCOTTISH	('SCOTTISH', 46)	('CERTAINLY', 1)	('DECISION', 1)	('STARTED', 1)	('THIRD', 1)
SECOND	('SECOND', 34)	('TAKEN', 3)	('EXACTLY', 2)	('SENSE', 2)	('TAKING', 2)
SECRETARY	('SECRETARY', 49)	('NOTHING', 1)			
SECTOR	('SECTOR', 36)	('STATES', 5)	('STAND', 2)	('ATTACKS', 1)	('CERTAINLY', 1)
SECURITY	('SECURITY', 46)	('ALWAYS', 1)	('SIDES', 1)	('STORY', 1)	('YEARS', 1)
SEEMS	('SEEMS', 42)	('STATEMENT', 3)	('COMES', 1)	('COMING', 1)	('SECTOR', 1)
SENIOR	('SENIOR', 45)	('CERTAINLY', 1)	('DETAILS', 1)	('POSITION', 1)	('STATE', 1)
SENSE	('SENSE', 40)	('STATES', 3)	('EXACTLY', 2)	('BUSINESSES', 1)	('GETTING', 1)

SERIES	('SERIES', 41)	('SERIOUS', 6)	('SERVICE', 2)	('CONSERVATIVE', 1)	
SERIOUS	('SERIOUS', 42)	('SERIES', 3)	('KNOWN', 1)	('SIDES', 1)	('SYRIA', 1)
SERVICE	('SERVICE', 47)	('EVIDENCE', 1)	('SERVICES', 1)	('STAFF', 1)	
SERVICES	('SERVICES', 45)	('SERVICE', 3)	('AUTHORITIES', 1)	('OFFICERS', 1)	
SEVEN	('SEVEN', 45)	('SEVERAL', 2)	('DIFFERENT', 1)	('NEVER', 1)	('SERVICE', 1)
SEVERAL	('SEVERAL', 42)	('LEVEL', 2)	('CENTRAL', 1)	('DAVID', 1)	('DIFFERENT', 1)
SHORT	('SHORT', 45)	('STRONG', 2)	('CONTROL', 1)	('EUROPE', 1)	('ISSUES', 1)
SHOULD	('SHOULD', 35)	('ACTUALLY', 2)	('COULD', 2)	('GEORGE', 2)	('ISSUE', 2)
SIDES	('SIDES', 44)	('ATTACK', 1)	('CHARGE', 1)	('FRANCE', 1)	('GROWTH', 1)
SIGNIFICANT	('SIGNIFICANT', 49)	('SENIOR', 1)			
SIMPLY	('SIMPLY', 48)	('SEEMS', 1)	('SOMEONE', 1)		
SINCE	('SINCE', 41)	('THESE', 2)	('GREECE', 1)	('INSIDE', 1)	('JUDGE', 1)
SINGLE	('SINGLE', 43)	('CENTRAL', 1)	('CERTAINLY', 1)	('ITSELF', 1)	('KILLED', 1)
SITUATION	('SITUATION', 50)				
SMALL	('SMALL', 47)	('REPORT', 1)	('SUPPORT', 1)	('TOWARDS', 1)	
SOCIAL	('SOCIAL', 44)	('CENTRAL', 1)	('GEORGE', 1)	('SCHOOLS', 1)	('SCOTTISH', 1)
SOCIETY	('SOCIETY', 39)	('DECIDED', 2)	('ATTACK', 1)	('ATTACKS', 1)	('CONCERNS', 1)
SOMEONE	('SOMEONE', 42)	('COMING', 2)	('SOMETHING', 2)	('DIFFERENT', 1)	('MISSING', 1)

SOMETHING	('SOMETHING', 44)	('SOMEONE', 2)	('COMING', 1)	('DIFFERENCE', 1)	('SIMPLY', 1)
SOUTH	('SOUTH', 46)	('DEATH', 2)	('SEVERAL', 1)	('THEMSELVES', 1)	
SOUTHERN	('SOUTHERN', 44)	('CERTAINLY', 2)	('SENSE', 2)	('DAVID', 1)	('SINCE', 1)
SPEAKING	('SPEAKING', 44)	('SPENDING', 2)	('BEING', 1)	('BUILDING', 1)	('COULD', 1)
SPECIAL	('SPECIAL', 47)	('BUILD', 1)	('MONTHS', 1)	('SPEECH', 1)	
SPEECH	('SPEECH', 46)	('BIGGEST', 1)	('CUSTOMERS', 1)	('INFORMATION', 1)	('MEANS', 1)
SPEND	('SPEND', 27)	('SPENT', 9)	('SPENDING', 6)	('DESPITE', 2)	('EXPECTED', 2)
SPENDING	('SPENDING', 40)	('SPENT', 4)	('SPEND', 2)	('INDEPENDENT', 1)	('MAKING', 1)
SPENT	('SPENT', 35)	('SPEND', 4)	('INDEPENDENT', 3)	('SPENDING', 3)	('EXPECTED', 2)
STAFF	('STAFF', 44)	('CHARGE', 1)	('ENOUGH', 1)	('ITSELF', 1)	('SERVICE', 1)
STAGE	('STAGE', 35)	('NEEDS', 2)	('SIDES', 2)	('STATES', 2)	('ALLOW', 1)
STAND	('STAND', 31)	('EXACTLY', 3)	('STATES', 3)	('ATTACKS', 2)	('SAYING', 2)
START	('START', 39)	('CONCERNS', 2)	('DECIDED', 2)	('STAFF', 2)	('CANNOT', 1)
STARTED	('STARTED', 39)	('CERTAINLY', 2)	('DECIDED', 2)	('ASKING', 1)	('NIGHT', 1)
STATE	('STATE', 35)	('UNDERSTAND', 3)	('STAGE', 2)	('DECIDED', 1)	('ITSELF', 1)
STATEMENT	('STATEMENT', 46)	('SEEMS', 4)			
STATES	('STATES', 35)	('SECTOR', 2)	('STAGE', 2)	('CASES', 1)	('EDUCATION', 1)

STILL	('STILL', 30)	('ITSELF', 5)	('UNTIL', 4)	('SINGLE', 2)	('ASKING', 1)
STORY	('STORY', 44)	('DURING', 1)	('FORWARD', 1)	('SCOTLAND', 1)	('STILL', 1)
STREET	('STREET', 43)	('INDUSTRY', 2)	('DURING', 1)	('STRONG', 1)	('THIRD', 1)
STRONG	('STRONG', 48)	('CONTROL', 1)	('WRONG', 1)		
SUNDAY	('SUNDAY', 43)	('CERTAINLY', 3)	('SENSE', 3)	('USING', 1)	
SUNSHINE	('SUNSHINE', 48)	('COUNTRY', 1)	('VOTERS', 1)		
SUPPORT	('SUPPORT', 45)	('SMALL', 2)	('BECAUSE', 1)	('IMPORTANT', 1)	('REPORT', 1)
SYRIA	('SYRIA', 45)	('SYRIAN', 4)	('SERIOUS', 1)		
SYRIAN	('SYRIAN', 39)	('SYRIA', 5)	('CONCERNS', 1)	('DURING', 1)	('INSIDE', 1)
SYSTEM	('SYSTEM', 47)	('CONTINUE', 1)	('LITTLE', 1)	('SHOULD', 1)	
TAKEN	('TAKEN', 35)	('TAKING', 5)	('SECOND', 3)	('EXACTLY', 2)	('GETTING', 1)
TAKING	('TAKING', 32)	('TAKEN', 7)	('CLEAR', 2)	('SAYING', 2)	('CERTAINLY', 1)
TALKING	('TALKING', 41)	('DOING', 2)	('BETWEEN', 1)	('DECIDED', 1)	('LOOKING', 1)
TALKS	('TALKS', 44)	('COURSE', 1)	('COURT', 1)	('EUROPE', 1)	('GEORGE', 1)
TEMPERATURES	('TEMPERATURES', 50)				
TERMS	('TERMS', 44)	('TIMES', 2)	('EXAMPLE', 1)	('GROUP', 1)	('NUMBER', 1)
THEIR	('THEIR', 27)	('THERE', 5)	('AGAIN', 2)	('ASKING', 2)	('STAND', 2)
THEMSELVES	('THEMSELVES', 50)				

THERE	('THERE', 20)	('THEIR', 7)	('SERIOUS', 2)	('AFRICA', 1)	('AGREE', 1)
THESE	('THESE', 26)	('NEEDS', 4)	('THINGS', 4)	('AGAINST', 1)	('BUSINESS', 1)
THING	('THING', 21)	('THINGS', 3)	('THINK', 3)	('NOTHING', 2)	('THIRD', 2)
THINGS	('THINGS', 23)	('THESE', 8)	('SENSE', 3)	('STATES', 3)	('SINCE', 2)
THINK	('THINK', 18)	('THING', 11)	('THINGS', 3)	('TAKING', 2)	('ACTUALLY', 1)
THIRD	('THIRD', 42)	('ANOTHER', 1)	('CANNOT', 1)	('FIGURES', 1)	('FOCUS', 1)
THOSE	('THOSE', 37)	('CLOSE', 2)	('COURSE', 2)	('COULD', 1)	('FRANCE', 1)
THOUGHT	('THOUGHT', 31)	('COURT', 2)	('GOING', 2)	('CALLED', 1)	('CONCERNS', 1)
THOUSANDS	('THOUSANDS', 48)	('AGAINST', 1)	('FIRST', 1)		
THREAT	('THREAT', 48)	('RATES', 1)	('THOUGHT', 1)		
THREE	('THREE', 37)	('AGREE', 3)	('TALKING', 2)	('CONFLICT', 1)	('DIFFERENT', 1)
THROUGH	('THROUGH', 37)	('HISTORY', 2)	('SCHOOL', 2)	('ACROSS', 1)	('AFTERNOON', 1)
TIMES	('TIMES', 45)	('GAMES', 2)	('TERMS', 2)	('COMES', 1)	
TODAY	('TODAY', 41)	('UNDERSTAND', 2)	('ATTACKS', 1)	('INSIDE', 1)	('PROTECT', 1)
TOGETHER	('TOGETHER', 40)	('ANOTHER', 2)	('NEVER', 2)	('DEATH', 1)	('GETTING', 1)
TOMORROW	('TOMORROW', 50)				
TONIGHT	('TONIGHT', 40)	('NIGHT', 4)	('ATTACKS', 1)	('INSIDE', 1)	('LEADERSHIP', 1)
TOWARDS	('TOWARDS', 44)	('COURSE', 1)	('ORDER', 1)	('SCHOOLS', 1)	('TALKS', 1)

TRADE	('TRADE', 38)	('TRYING', 7)	('GREAT', 2)	('CONTINUE', 1)	('RECORD', 1)
TRIAL	('TRIAL', 46)	('CHILD', 1)	('STAFF', 1)	('THREE', 1)	('WHERE', 1)
TRUST	('TRUST', 42)	('AGAINST', 1)	('CHANCE', 1)	('CHARGES', 1)	('COURSE', 1)
TRYING	('TRYING', 43)	('TRADE', 2)	('TRUST', 2)	('CHILD', 1)	('GREAT', 1)
UNDER	('UNDER', 26)	('LONDON', 2)	('STARTED', 2)	('UNION', 2)	('ACTION', 1)
UNDERSTAND	('UNDERSTAND', 39)	('STAND', 3)	('ACTION', 1)	('AFFAIRS', 1)	('INSIDE', 1)
UNION	('UNION', 45)	('DOING', 2)	('LONGER', 1)	('LOOKING', 1)	('SOCIETY', 1)
UNITED	('UNITED', 41)	('ASKING', 1)	('CANNOT', 1)	('DOING', 1)	('HEART', 1)
UNTIL	('UNTIL', 30)	('STILL', 3)	('ANYTHING', 1)	('ASKING', 1)	('COMES', 1)
USING	('USING', 40)	('LOOKING', 2)	('COUNCIL', 1)	('ENOUGH', 1)	('INCREASE', 1)
VICTIMS	('VICTIMS', 50)				
VIOLENCE	('VIOLENCE', 47)	('FINAL', 2)	('YEARS', 1)		
VOTERS	('VOTERS', 47)	('FOCUS', 2)	('FORCES', 1)		
WAITING	('WAITING', 48)	('QUESTION', 1)	('WESTERN', 1)		
WALES	('WALES', 47)	('RATES', 1)	('WEATHER', 1)	('WORLD', 1)	
WANTED	('WANTED', 35)	('WANTS', 4)	('WATCHING', 4)	('PERSON', 1)	('POLITICS', 1)
WANTS	('WANTS', 36)	('WANTED', 8)	('MONTHS', 1)	('TOWARDS', 1)	('WESTERN', 1)
WARNING	('WARNING', 46)	('MORNING', 4)			
WATCHING	('WATCHING', 44)	('WANTED', 4)	('PROCESS', 1)	('WATER', 1)	

WATER	('WATER', 44)	('BORDER', 1)	('ORDER', 1)	('WANTED', 1)	('WARNING', 1)
WEAPONS	('WEAPONS', 50)				
WEATHER	('WEATHER', 40)	('WHETHER', 5)	('ABOUT', 1)	('ANYTHING', 1)	('RATHER', 1)
WEEKEND	('WEEKEND', 49)	('AHEAD', 1)			
WEEKS	('WEEKS', 42)	('REASON', 2)	('FRONT', 1)	('MINUTES', 1)	('NEEDS', 1)
WELCOME	('WELCOME', 46)	('QUESTION', 2)	('PARLIAMENT', 1)	('WHILE', 1)	
WELFARE	('WELFARE', 50)				
WESTERN	('WESTERN', 49)	('QUESTION', 1)			
WESTMINSTER	('WESTMINSTER', 50)				
WHERE	('WHERE', 42)	('INQUIRY', 1)	('POWERS', 1)	('PRESS', 1)	('QUITE', 1)
WHETHER	('WHETHER', 45)	('RATHER', 1)	('WANTS', 1)	('WEATHER', 1)	('WELCOME', 1)
WHICH	('WHICH', 45)	('WOULD', 2)	('BENEFITS', 1)	('QUESTION', 1)	('QUESTIONS', 1)
WHILE	('WHILE', 38)	('WHERE', 6)	('WORLD', 2)	('MEANS', 1)	('RATHER', 1)
WHOLE	('WHOLE', 42)	('KNOWN', 2)	('FORWARD', 1)	('HOURS', 1)	('INCREASE', 1)
WINDS	('WINDS', 49)	('WESTERN', 1)			
WITHIN	('WITHIN', 44)	('REALLY', 2)	('EVERYTHING', 1)	('THESE', 1)	('WORKERS', 1)
WITHOUT	('WITHOUT', 49)	('ANNOUNCED', 1)			
WOMEN	('WOMEN', 50)				

WORDS	('WORDS', 33)	('WORST', 5)	('WANTS', 2)	('WEEKS', 2)	('WORKERS', 2)
WORKERS	('WORKERS', 46)	('WARNING', 2)	('WORDS', 2)		
WORKING	('WORKING', 46)	('MAKING', 1)	('MORNING', 1)	('WORKERS', 1)	('WORST', 1)
WORLD	('WORLD', 32)	('WORKERS', 3)	('WALES', 2)	('WHETHER', 2)	('WHILE', 2)
WORST	('WORST', 38)	('WORDS', 9)	('POINT', 1)	('WANTED', 1)	('WORLD', 1)
WOULD	('WOULD', 40)	('STATEMENT', 2)	('WHICH', 2)	('GLOBAL', 1)	('PARLIAMENT', 1)
WRONG	('WRONG', 48)	('FOREIGN', 1)	('RIGHT', 1)		
YEARS	('YEARS', 34)	('NEEDS', 3)	('THESE', 3)	('HEARD', 2)	('AGAINST', 1)
YESTERDAY	('YESTERDAY', 48)	('GETTING', 1)	('SINCE', 1)		
YOUNG	('YOUNG', 38)	('CHILD', 2)	('BECOME', 1)	('CERTAINLY', 1)	('CHINA', 1)

D. Code

HOMOPHEME GENERATOR

The configuration files, pickle files, pickling process, csv files and GUI related code is not included. Only the code related to the logic of homopheme generation is shown in the following blocks. The generator has been made available for public use at: https://lsbu-analytics.org/deeplip/playground/similarLookingSentences/fromSaved

```
def save_csv(data, file):
  file = os.path.abspath(os.path.join(tmp_folder, file +'.csv'))
  with open(file, 'w', encoding='utf-8') as f:
    w = csv.DictWriter(f, data.keys())
```

```
    w.writeheader()
    w.writerow(data)
```

```
def word_to_visemes(word, show_phonemes=False, save_csv_file=True):
  # get the cmu phones for the word
  ph = []
  word_vis = []
  with open(lrs_cmu_phones_pickle_file+'_all.pickle', 'rb') as f:
    all_phones = pickle.load(f)
  try:
    ph = all_phones[word]
    # remove the syllable stress markers '0, 1, 2, 3' from ph
    for i, p in enumerate(ph):
      for n in '0123':
        if n in p:
          ph[i] = p.strip(n)
    # map ph to visemes
    for p in ph:
      for vi, phon in viseme_phoneme.items():
        if p in phon:
          word_vis.append(vi)
    if show_phonemes == True:
      print('Phonemes for', word, ':',ph)
  except KeyError:
    word_vis = ['N/A']
  word_visemes = dict(enumerate(word_vis))
  if save_csv_file:
    save_csv(word_visemes, 'word_visemes')
  return word_vis
```

```
def find_homophemes(word, save_csv_file=True, include_message=False):
 visemes = word_to_visemes(word, show_phonemes=False, save_csv_file=False)
 possible_sounds = []
 for v in visemes:
   for v1, s in viseme_sound.items():
     if v == v1:
       possible_sounds.append(s)
 possible_word_sounds = list(itertools.product(*possible_sounds))
 possible_words = []
 for k, v in cmudict.dict().items():
   for w in possible_word_sounds:
     if list(w) in v:
       possible_words.append(k)
str(len(possible_words)) + '/' + str(len(possible_word_sounds))
 homophemes = dict(enumerate(possible_words))
 if save_csv_file:
   save_csv(homophemes, 'homophemes')
 else:
```

```
   return possible_words
```

```
def similar_looking_sentences(sentence, show_sentences=0 ):
  words = sentence.lower().split()
  homopheme_map = {}
  # find similar looking words for each word
  for word in words:
    homophemes = find_homophemes(word, save_csv_file=False,include_message=True)
    homopheme_map[word] = homophemes
  homopheme_list = []
  messages = []
  for k,v in homopheme_map.items():
    homopheme_list.append(v['possible_words'])
    messages.append(v['message'])
  # create all possible sentences
  sentences = list(itertools.product(*homopheme_list))
  print('Possible sentences: ', len(sentences))
  if show_sentences <= 100:
    out = {}
    # show a shuffled sample of the given % of sentences
    num_sentences = int((show_sentences/100) * len(sentences))
    random_sentences = random.sample(sentences, num_sentences)
    out['sentence'] = sentence
    out['messages'] = messages
    out['num_sentences'] = len(sentences)
    out['sentences'] = random_sentences
    save_csv(out, 'similar_looking_sentences')
    return out
  else:
    save_csv(homopheme_map, 'similar_looking_sentences')
    return homopheme_map
```

FACE ROTATOR

```
detector = dlib.get_frontal_face_detector()
predictor = dlib.shape_predictor(shape_predictor_file_68)

def rotate_face(show_drawing, show_window, gray):
    # angle calculated below based on mouth corners
    def rotate_image(image, angle):
        image_center = tuple(np.array(image.shape[1::-1]) / 2)
        rot_mat = cv.getRotationMatrix2D(image_center, angle, 1.0)
        result = cv.warpAffine(image, rot_mat, image.shape[1::-1],
flags=cv.INTER_LINEAR)
        return result
    # currently only works for single face per video
    crop_files = glob.glob(os.path.join(rotated_face_crops_dir, "*.png"))
    # remove existing faces if any
    for f in crop_files:
        os.remove(f)
```

```
    img_files = os.listdir(face_crops_dir)
    img_files = [os.path.abspath(os.path.join(face_crops_dir, x)) for x in
img_files]
    for index, f in enumerate(tqdm(img_files)):
        img = cv.imread(f)
        if gray:
            img = cv.cvtColor(img, cv.COLOR_BGR2GRAY)
            dets = detector(img, 1)
        angle = 0
        try:
            d = dets[0]  # the only face in the video
        except IndexError:  # no face detected
            continue
        # Get the landmarks/parts for the face in box d.
        shape = predictor(img, d)
        left_mouth_corner = shape.part(48)
        right_mouth_corner = shape.part(54)
        left_axes = left_mouth_corner.x, left_mouth_corner.y
        right_axes = right_mouth_corner.x, right_mouth_corner.y
        # calculate angle
        base, perpendicular = tuple(np.subtract(right_axes, left_axes))
        angle = atan(perpendicular/base) * 180 / pi
        # print('angle of rotation: ', angle, 'degrees')

        if show_drawing:
            # draw on image
            cv.line(img, left_axes, right_axes, (255,0,0), 2)
            corrected_right = right_axes[0], left_axes[1]
            cv.line(img, left_axes, corrected_right, (255,0,0), 2)
            cv.circle(img, left_axes, 15, (255,0,0), 1)
            cv.circle(img, right_axes, 5, (255,0,0), 1)
            cv.circle(img, corrected_right, 5, (255,0,0), 1)
            cv.putText(img, str(int(angle))+' degrees', (left_axes[0],
left_axes[1]-3), cv.FONT_HERSHEY_SIMPLEX, 0.7, (0,255,0), 1, cv.LINE_AA)
            # rotate image
            rotated = rotate_image(img, angle)
            # rotated_crops.append(rotated)
            cv.imwrite(rotated_face_crops_dir + '/cropped_' +
str(index).zfill(5)+ '.png', rotated)
        else:
            # rotate image
            rotated = rotate_image(img, angle)
            cv.imwrite(rotated_face_crops_dir + '/cropped_' +
str(index).zfill(5) + '.png', rotated)

    if show_window:
        while True:
            cv.imshow('Original '+str(index)+' : ', img)
            cv.imshow('Rotated '+str(index)+' : ', rotated)
            if cv.waitKey(1) & 0xFF == ord('q'):
                break
```

```
    cv.destroyAllWindows()

rotate_face(show_drawing=False, show_window=False, gray=False)
```