Intelligent Systems with Applications 26 (2025) 200517

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Improving speaker-independent visual language identification using deep neural networks with training batch augmentation

Jacob L. Newman

School of Computing Sciences, University of East Anglia, University Drive, Norwich, NR4 7TJ, Norfolk, England, United Kingdom

ARTICLE INFO

Keywords:
Language identification
Lip reading
Neural networks
Time series classification

ABSTRACT

Visual Language Identification (VLID) is concerned with using the appearance and movement of the mouth to determine the identity of spoken language. VLID has applications where conventional audio based approaches are ineffective due to acoustic noise, or where an audio signal is unavailable, such as remote surveillance. The main challenge associated with VLID is the speaker-dependency of image based visual recognition features, which bear little meaningful correspondence between speakers.

In this work, we examine a novel VLID task using video of 53 individuals reciting the Universal Declaration of Human Rights in their native languages of Arabic, English or Mandarin. We describe a speaker-independent, five fold cross validation experiment, where the task is to discriminate the language spoken in 10 s videos of the mouth. We use the YOLO object detection algorithm to track the mouth through time, and we employ an ensemble of 3D Convolutional and Recurrent Neural Networks for this classification task. We describe a novel approach to the construction of training batches, in which samples are duplicated, then reversed in time to form a *distractor* class. This method encourages the neural networks to learn the discriminative temporal features of language rather than the identity of individual speakers.

The maximum accuracy obtained across all three language experiments was 84.64%, demonstrating that the system can distinguish languages to a good degree, from just 10 s of visual speech. A 7.77% improvement on classification accuracy was obtained using our distractor class approach compared to normal batch selection. The use of ensemble classification consistently outperformed the results of individual networks, increasing accuracies by up to 7.27%. In a two language experiment intended to provide a comparison with our previous work, we observed an absolute improvement in classification accuracy of 3.6% (90.01% compared to 83.57%).

1. Introduction

Automatic identification of spoken language relates to classifying *which* language is being spoken, rather than understanding *what* is being said (Van Segbroeck, Travadi, & Narayanan, 2015). Language identification using audio speech is a process that usually precedes the selection of a language-dependent subsystem, such as a speech recogniser, call router, or public information terminal. Visual Language Identification (VLID) relies on the appearance and motion of the mouth to determine language (Newman & Cox, 2009). This has potential applications in noisy environments, where conventional approaches that rely on acoustic data are largely ineffective. For example, a system that can lip read language could be used to automate the selection of the appropriate language on a public information terminal. It could also have security applications, such as for long distance surveillance or for the analysis of CCTV footage with application to forensics (Preethi et al., 2023; Rothkrantz, 2017). The author is particularly interested in this topic in light of recent work by others (Afouras, Chung, & Zisserman, 2020; Cascone, Nappi, & Narducci, 2023), and having published the first work in VLID prior to 2011.

Lip reading is a challenging task for humans (Altieri, Pisoni, & Townsend, 2011). The positioning of the tongue, velum, vocal tract and vocal folds are crucial for determining the sounds produced during speech, and much of this information is hidden from view (Bernstein, Tucker, & Demorest, 2000). Specifically, only the shape of the mouth and the front most speech articulators are visible, limiting our ability as humans to determine which sounds and words have been spoken (Bernstein, Jordan, Auer, & Eberhardt, 2022). Finding the identity of a spoken language using audio speech is a generally easier task than identifying individual spoken words. Languages differ in terms of their vocabulary, phonemes, and critically, their phonotactics: the order of permitted phonemes (Jannah, Mashalani, Lubis, & Amaro, 2023). Exploiting phonotactic differences, computers can achieve language discrimination accuracies of over 98% from short extracts of audio

E-mail address: jacob.newman@uea.ac.uk.

https://doi.org/10.1016/j.iswa.2025.200517
Received 8 January 2025; Received in revised form 11 February 2025; Accepted 8 April 2025
Available online 22 April 2025

Table 1
A comparison between this study and all other known works related to VLID.

Study	Task	Method	Results
This study	3 languages (English, Arabic and Mandarin) from studio video data. 10 s utterances.	The mouth region is used as input to an ensemble of DNNs comprising 3D CNN and GRU layers. Network training includes a distractor class. Maximum likelihood classifies the speech.	84.6% speaker-independent classification accuracy.
(Cascone et al., 2023) (2023)	8 languages, 10 s utterances.	The mouth region is used as input to DNNs (ConvLSTMs with BLSTMs) and a separate SVM.	37.5% speaker-independent classification accuracy.
(Afouras et al., 2020) (2020)	14 languages from TEDx and YouTube videos. 10 s utterances.	Embeddings from ResNet18 as input features into DNNs comprising Time-Delay Neural Networks and bi-directional LSTMs (BLSTMs). Unknown if the whole face is used.	76.3% speaker-independent classification accuracy.
(Špetlík, Čech, Franc, & Matas, 2017) (2017)	2 languages (English and French) from YouTube videos. 20 s utterances.	Facial landmarks used in a soft-assignment variant of bag-of-words, followed by a linear classifier. Landmarks from the mouth region produce the best results.	73% speaker-independent classification accuracy.
(Chandrasekhar, Sargin, & Ross, 2011) (2011)	25 languages from YouTube music videos. Each song is an utterance.	Quantised features derived from a hue-saturation histogram, and motion characteristics from Motion Cuboids. Back end SVM classifier. Visual-only and audio-visual experiments are presented.	14.3% classification accuracy. Unknown if the experiment is speaker-independent.
(Newman & Cox, 2012) (2011)	2 languages (English and Arabic) from UN2. 7 s utterances.	Active Appearance Model features used in hidden Markov models of triphones. Language models of trigram sequences are built. Back end SVM classifier.	86.4% speaker-independent classification accuracy.

speech (Biswas, Rahaman, Ahmadian, Subari, & Singh, 2023; Heracleous, Takai, Yasuda, Mohammad, & Yoneyama, 2018). With just 6 s of representative speech, humans too can use audio speech to accurately discriminate 10 languages with 69.4% accuracy (Komatsu, 2007). It has also been shown that babies can tell languages apart (Zacharaki & Sebastian-Galles, 2021). However, the performance of language identification using lip reading is harder for both humans (Soto-Faraco et al., 2007) and computers.

There are two significant challenges associated with computer lip reading: as described above, a lack of speech information visible on the mouth, and secondly, the speaker-dependency of visual recognition features. As with human lip reading, the lack of information presented on the face limits the capability of computer lip reading to distinguish spoken phonemes. It is generally accepted that there is a many-to-one mapping between phonemes and their visual equivalent, *visemes*. However, visemes are neither well defined, nor consistent (Bear & Harvey, 2019; Taylor, Theobald, & Matthews, 2015), and they are strongly affected by speech context (Taylor, Theobald, & Matthews, 2014). For any given lip shape, a multitude of different sounds could be expressed, greatly complicating the task of discriminating spoken language from visual information.

Compared to language identification using acoustic features, VLID remains an immature field of research with very few studies focussing on this task. However, of the limited number of previous studies involving computer lip reading, most have focused on developing multispeaker speech recognition systems. A multispeaker system uses models that are trained on the same set of speakers as exists in the testing dataset (Cox, Harvey, Lan, Newman, & Theobald, 2008). The multispeaker testing framework masks the speaker-dependency issue affecting visual recognition features. In Cox et al. (2008), we showed that image based recognition features are unique to each speaker, with little observable relationship or meaningful overlap in the feature space. Thus, if a system is tested on different speakers than those presented during training, recognition accuracy is poor. The issue of speaker-dependency remains significant as, firstly, speech has not evolved to be discriminative from a visual perspective, and because computer lip reading is a far less mature research domain compared to audio speech recognition.

All known previously published works relating to VLID are summarised in Table 1. We will describe these studies in turn. In our own previous VLID work, we applied conventional machine learning techniques to both speaker-dependent language identification involving multilingual speakers (Newman & Cox, 2009), and a two language speaker-independent problem (Newman & Cox, 2010). We adapted approaches for audio language identification that exploit the phonotactic differences between languages (Mohapatra, Dash, & Majhi, 2016; Ulbrich, Alday, Knaus, Orzechowska, & Wiese, 2016). In our speaker-independent study, language models of viseme sequences were built from training data, with test data assigned to the class whose language model produced the highest likelihood. We achieved an error rate of 4.6% for a two language problem, with 30 s of representative test speech data. Using 7 s of test data, the accuracy was much lower, with 7 of the speakers producing an error rate of between 20% and 45%, and a mean error rate of 16.43%.

In Špetlík et al. (2017), researchers trained classifiers to discriminate between the English and French languages using a histogram of lip shapes detected during speech—this was based on the idea that there are key mouth shapes that are unique to particular languages. While this work is a novel and intuitive approach to identifying language, it ignores the temporal features of speech and considers only shape-based features. They achieved a comparatively low accuracy of 73%, which was close to the results they obtained performing the same experiment with human participants. In a related language task, In Chandrasekhar et al. (2011), the language of singing in music videos from YouTube was classified using image and motion based features, and SVMs. Classification accuracies of 14.3% were obtained for a 25 language task.

More recently, Deep Neural Networks (DNNs) have been widely applied to the field of acoustic speech recognition (Fenghour, Chen, Guo, Li, & Xiao, 2021; Mehrish, Majumder, Bharadwaj, Mihalcea, & Poria, 2023), and also to VLID (Afouras et al., 2020; Cascone et al., 2023). In Afouras et al. (2020), DNNs were used to discriminate between 14 different languages using computer lip reading. The system was trained using over 1700 h of TEDx talk videos, and achieved an accuracy of 76.3% with 10 s test data segments. The quantity of data used to achieve these results suggests that an increased diversity of speakers may have provided a better coverage of the feature space, helping to overcome the speaker-dependency of image based recognition features. However, Afouras et al. (2020) also reported a control experiment in which a network pretrained for face recognition, ResNet50, was fine-tuned to the VLID task. They achieved above chance language discrimination (40.8%), suggesting that a feature of the YouTube videos was providing an unintended indication of the geography of the recording or some other biasing factor.

In Cascone et al. (2023), an eight language task was considered as part of a wider system to perform speaker identification. A custom

dataset of YouTube videos from 256 individuals was constructed, comprising 1280 10 s speech samples. They explored the use of machine learning and deep learning for a speaker-independent classification task, achieving a peak mean accuracy of 37.5% using Support Vector Machines. As far as the authors are aware, there are no other studies specifically addressing the challenge of VLID. Therefore, there is an opportunity to revisit this classification problem, applying contemporary machine learning methods to a simpler task in order to focus on improving speaker-independent classification performance.

This article extends our previous work involving a two language classification task, by applying DNNs to a three language task. In this work, we conduct a speaker-independent, five fold cross validation experiment to classify the identity of Arabic, English and Mandarin speech, from 10 s test segments of video. The novelty of the work presented here firstly lies in the classification problem itself. We perform a constrained classification task in which three very different languages are considered, using video captured in studio conditions. Despite this, the problem remains challenging. We use an unconventional approach to mouth-tracking using a near real time, object detection algorithm, which is shown to work remarkably well across multiple speakers. We use ensemble classifiers with a 3D Convolutional Neural Network baseline architecture and extend the baseline by incorporating different types of Recurrent Neural Networks.

The significant novel contributions of this work are as follows:

- We use a novel neural network configuration involving GRU layers, which has not previously been applied to language discrimination through lip reading.
- We introduce a simple, yet novel and effective method of combining the classification output from an ensemble of neural networks.
- The most significant contribution of this work lies in the training of the neural networks. We significantly improve the classification accuracy of our models by supplementing each training batch with *distractor* samples, designed to discourage the networks from relying on the identity of each speaker to determine their spoken language. Effectively, this improves the speaker-independent performance of the system.

The remainder of this article is organised as follows: In Section 2 we describe the dataset used in these experiments. In Section 3, we outline the classification problem and experimental framework. Our language identification system is presented in Section 4, including the recognition features used (Section 4.1), the neural network architectures explored (Section 4.2) and our novel method for constructing training batches (Section 4.3). Results, Discussion and Conclusion are presented in Sections 5, 6 and 7, respectively.

2. Dataset

In this section, we describe details of the dataset used in this work. The data capture process, the nature and content of the data, and the size of the dataset are provided.

The dataset used in these experiments is the United Nations 2 (UN2) dataset introduced in our previous work (Newman & Cox, 2010), plus an additional twenty-three Mandarin speakers (Table 2). This dataset comprises individuals reciting the first sixteen articles of the United Nations Declaration of Human Rights in one of three languages: Arabic, English and Mandarin. The speakers have native language proficiency. The distribution of sex per language is also presented in the table, showing that there are more male speakers overall.

This data was recorded using a Sanyo Xacti VPC-FH1 video camera, captured at 1920 by 1080 pixels, at 60 frames per second. The video was recorded in a studio environment and includes the bottom half of the face only, maximising the resolution of the mouth region (Fig. 1).

Table 2
Details of the UN2 dataset. The dataset contains videos of speakers reading the first sixteen articles of the UN Declaration of Human Rights.

Speaker IDs	Language	Male/Female	Total Duration (hh:mm:ss)
1, 2, 3, 4, 5, 6, 7, 8, 12, 14, 15, 16, 17, 18, 20, 21, 23, 24, 26, 29	English	16/4	4:08:50
32, 33, 34, 35, 36, 37, 38, 39, 40, 41	Arabic	8/2	2:13:00
42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64	Mandarin	13/10	4:45:10

Fig. 1. Example framing of the face during recordings of the UN2 dataset.

3. Experimental setup

This section provides a broad overview of the approach adopted for the development and evaluation of our language identification system. We explain the specific classification task undertaken, and give details of the separate experimental configurations explored.

We designed our experimental setup to evaluate language identification in a speaker-independent mode. The speakers used to test the system were not included in the training set and we used five fold cross validation, in which the total pool of speakers was divided into five testing folds. Each speaker appears in only one test fold. For each testing fold, the remaining speakers were used as training and validation data, with the latter formed by randomly selecting 3 speakers from the training data. In this way, for each fold, the speakers in the test data were not present in the training or validation data. The full details of this experimental setup are presented in Table 3.

Video data from each speaker was processed to give 10 s, non-overlapping speech segments (See Section 4.1 for more information).

We undertook four main experiments. Firstly, starting from a baseline neural network, we compared three configurations of neural network layers. These networks are described in Section 4.2. Next, using the best performing architecture from the first experiments, we examined the effect on VLID performance of using a novel approach to neural network training in which speech samples are modified to create a distractor class (Section 4.3). Finally, we explored the use of an ensemble approach to classification, comparing results from individual networks to those produced by an ensemble of networks. Finally, we used the same system to conduct an experiment using 7 s test utterances and a two language (English and Arabic) classification task, to provide a comparison with our previous work.

4. Language identification system

This section illustrates the language identification system developed for this work. We start by providing the overall system design, and then subsequent subsections clarify each component of the system in turn: the feature extraction process is discussed in Section 4.1, the neural

Table 3
Details of the speakers included in each of the folds of the five fold cross validation experiments.

Fold	Training Speaker IDs	Validation Speaker IDs	Testing Speaker IDs
1	**Eng:** 1, 4, 6, 7, 12, 14, 16, 18, 20, 21, 23, 24, 29 **Ara:** 33, 34, 35, 37, 38, 39, 40, 41 **Man:** 42, 43, 44, 45, 47, 48, 49, 50, 51, 53, 55, 56, 58, 60, 61, 62, 63, 64	**Eng:** 5 **Ara:** **Man:** 52, 59	**Eng:** 2, 3, 8, 15, 17, 26 **Ara:** 36 **Man:** 46, 54, 57
2	**Eng:** 2, 3, 4, 5, 6, 7, 8, 12, 15, 16, 17, 18, 20, 21, 24, 26, 29 **Ara:** 34, 36, 38, 39, 40, 41 **Man:** 44, 45, 46, 47, 50, 52, 53, 55, 56, 57, 58, 59, 61, 62, 63	**Eng:** **Ara:** 33 **Man:** 42, 54	**Eng:** 1, 14, 23 **Ara:** 35, 37 **Man:** 43, 48, 49, 51, 60, 64
3	**Eng:** 1, 2, 3, 4, 5, 7, 8, 14, 15, 16, 17, 18, 21, 23, 24, 26, 29 **Ara:** 33, 34, 35, 36, 37, 38, 39, 40 **Man:** 43, 44, 46, 48, 51, 54, 55, 58, 60, 61, 62, 63, 64	**Eng:** **Ara:** **Man:** 49. 52, 57	**Eng:** 6, 12, 20 **Ara:** 41 **Man:** 42, 45, 47, 50, 53, 56, 59
4	**Eng:** 1, 2, 3, 6, 7, 12, 14, 15, 18, 20, 23, 24, 26, 29 **Ara:** 35, 36, 37, 39, 40, 41 **Man:** 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 57, 59, 60, 61, 62, 63, 64	**Eng:** 8, 17 **Ara:** **Man:** 56	**Eng:** 4, 5, 16, 21 **Ara:** 33, 34, 38 **Man:** 44, 55, 58
5	**Eng:** 1, 2, 3, 4, 5, 6, 8, 12, 14, 15, 16, 20, 21, 23, 26 **Ara:** 33, 34, 35, 36, 37, 41 **Man:** 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 57, 58, 59, 60, 64	**Eng:** 17 **Ara:** 38 **Man:** 56	**Eng:** 7, 18, 24, 29 **Ara:** 39, 40 **Man:** 52, 61, 62, 63

network architecture used in Section 4.2, and our novel approach to augmentation of training batches is explained in Section 4.3.

The language identification system we evaluated in our experiments is shown in Fig. 2. The system shown was configured as an ensemble classifier, but we also determined the classification performance of each network individually. At the front end, there is a feature extraction module which extracts the mouth region through time, constructs 10 s sequences of speech and normalises the data. This subsystem is described fully in Section 4.1. Next, five end-to-end neural networks take the 600 frame video samples as input. Their outputs are combined and input to an ensemble classifier, and the final output is either 3 or 4 classes, depending on whether the experiment performed includes a distractor class (Section 4.3).

When used as an ensemble classifier, the system combines the output of five networks to predict the identity of the spoken language in each 10 s speech utterance. In this approach, the outputs from the networks are combined by summing the confidence scores for each sample and then selecting the class with the highest resulting score. For experiments including a distractor class, only the confidences for the genuine language classes are used in this calculation.

The formal description of this process is as follows: let C represent the 3 language classes ($C = \{C_1, C_2, C_3\}$), and N the number of networks trained ($N = 5$). For each network $i = \{1, 2, 3, 4, 5\}$, let $S_{i,j}$ be the confidence score assigned by network i to class C_j.

The confidence score CS_j for class C_j is therefore given as:

$$CS_j = \sum_{i=1}^{N} S_{i,j}$$

The predicted class $\hat{y}$ for a given utterance is determined by the class with the maximum confidence score:

$$\hat{y} = \arg\max_{C_j \in C} CS_j$$

4.1. Recognition features

Extracting visual features for language identification requires the localisation and extraction of the mouth region in successive video frames. We initially considered two different approaches to tracking the mouth region in this work: Active Appearance Models (AAMs) (Matthews & Baker, 2004) and a Viola Jones cascade classifier (Wang, 2014). As well as providing a method to track the contour of the lips, AAMs provide shape and appearance-based recognition features (Lan, Harvey, & Theobald, 2012) which can be used for computer lip reading. However, AAMs used for tracking are subject-dependent, meaning that a unique model must be trained for each speaker, which is both time consuming and impractical for a generalisable lip reading system. Cascade classifiers are more generally applicable and can be used to extract a bounding box containing the mouth region. However, they rely on first detecting the whole face, which is not possible in our dataset, as the video only contains the lower portion of the face. Instead, we chose to use the YOLOv8 object detection algorithm to extract a bounding box for each frame of video (Al-obidi & Kacmaz, 2023).

We hand-labelled the mouth region in 2713 random video frames from across all speakers in our dataset. See Fig. 3 for example labels. 2580 frames were used as training data to fine-tune a pre-trained YOLOv8x model, and data from three speakers, comprising 133 frames, were used as validation data. The trained model detected mouths regions in the validation data with a Mean Average Precision (MAP) of 0.995, using an Intersection Over Union (IOU) threshold of 0.5. With an IOU of between 0.5 and 0.95, the MAP obtained was 0.931. After training, we applied this model to the complete video data from all 53 speakers. Visual inspection of the resulting classifications revealed impressive tracking accuracy across all speakers. Fig. 4 shows examples of mouth detections from three separate individuals.

The mouth regions detected in each image were cropped and resized to fit within a box of 150 pixels wide by 90 pixels high. The aspect ratios of the mouth regions were maintained during the resizing, meaning that they were usually a different overall shape to the target box. The resized regions were placed centrally in the target box and unoccupied pixels were assigned a value of 0. Each image was converted to grayscale, to mitigate skin tone (Fig. 5). Consecutive, non-overlapping sequences of 600 frames (or 10 s) were stacked into a four dimensional feature array. The resulting array for each speaker was of shape $(x, 600, 90, 150, 1)$, where x was the number of 10 s samples for that speaker. The pixel intensities within each 600 frame sample were normalised to have a maximum value of 1.

4.2. Neural networks

In this work, we use a baseline Neural Network architecture consisting of 3D CNNs, which are commonly applied to video classification problems (Wang, Pu, & Chen, 2022). They have also been used in computer lip reading (Exarchos et al., 2024; Margam et al., 2019). The network we use was developed using TensorFlow (Singh, Manure, Singh, & Manure, 2020) and is illustrated in Fig. 6. The hyperparameters used in this network were determined through manual optimisation during early experiments using the validation data.

The input layer in this network performed a further rescaling of the pixel intensities within each batch, ensuring the samples were in the range of −1 to 1. The first two hidden layers, which can be considered as feature extraction layers, are 3D convolutional layers, each containing eight kernels. The convolutional kernels in the first of these layers operate only on each 2D image, as shown by the kernel size of $(1, 3, 3)$. The kernels in the second layer apply a convolution through time and are of size $(3, 1, 1)$. Separating a 3D convolutional operation into two separate 2D convolutions (known as *2D + 1*) has

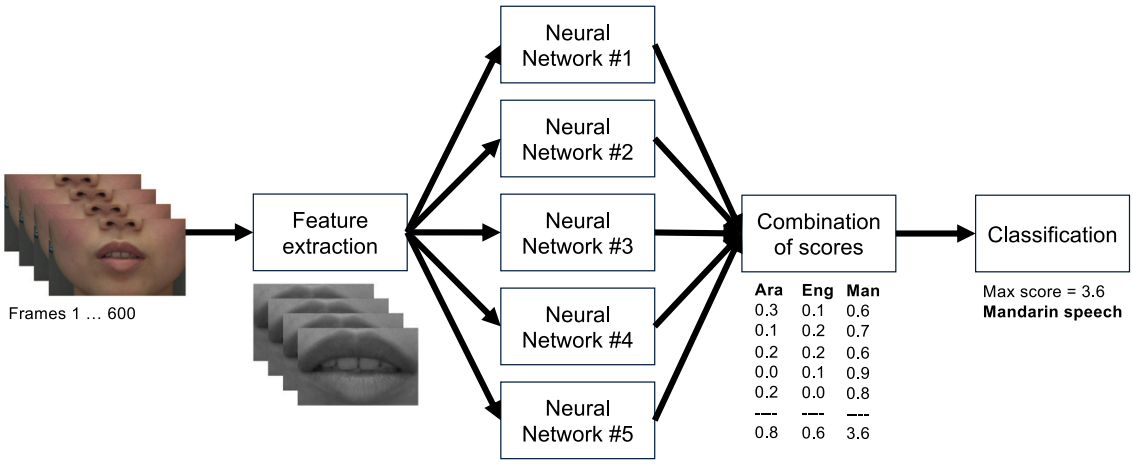

Fig. 2. Language identification system in an ensemble configuration. The mouth region is extracted through time, then five neural networks operate in parallel, followed by a module to combine the confidence scores from the networks to produce a final classification. Ara, Eng and Man refer to Arabic, English and Mandarin speech, respectively.

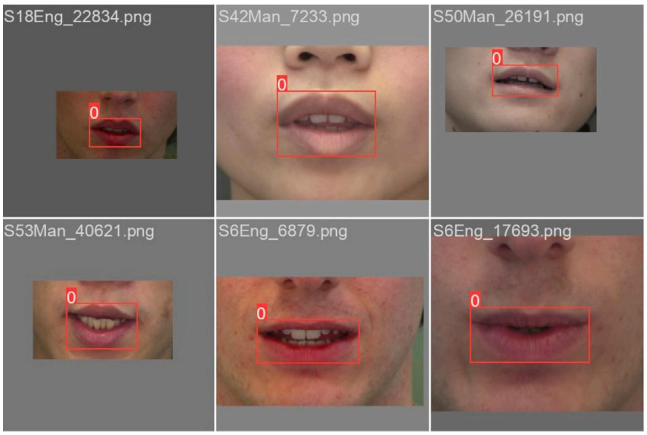

Fig. 3. Examples of hand labelled mouth regions (red boxes) in a selection of frames used for training a YOLOv8 object detection model. Note that the YOLO algorithm augments training samples by adjusting the size, hue, saturation and value of the training images, which is why the appearance of these images varies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Example mouth region classifications on test data. Red bounding boxes indicate the detected mouth regions, alongside network confidence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. Examples of cropping and resizing mouth regions to a fixed image size of 150 pixels wide by 90 pixels high. The resizing operation maintains the aspect ratio of the original mouth shape and the resulting image is placed centrally in the new image.

been shown to be a more memory and computationally efficient way to perform 3D convolutions than using a single 3D convolution kernel of size $(3, 3, 3)$ (Kopuklu, Kose, Gunduz, & Rigoll, 2019). Three pairs of 3D convolutional layers are used, each followed by a max pooling layer, and a 10% dropout layer, to reduce overfitting.

Following the final 3D convolutional layer, we experimented with a selection of different layer types, including a fully connected layer, and two different types of RNNs: Long Short Term Memory (LSTM), and Gated Recurrent Units (GRU). Both LSTMs and GRU layers are commonly applied to time series classification problems (Nosouhian, Nosouhian, & Khoshouei, 2021). GRU layers have been applied to computer lip reading (Miled, Messaoud, & Bouzid, 2023), but not previously to VLID. The final layers of the network included two fully connected layers, the last of which had either three or four outputs, depending on the experiment performed. Categorical cross-entropy was the selected loss function and the Adam optimiser used. Our code is freely available so that all remaining hyperparameters can be identified by the reader.

All neural networks were trained using Graphical Processing Unit (GPU) resources on the University of East Anglia's High Performance Computing cluster. Mixed precision training was employed, in which all network layers, except the final layer, used 16-bit floating point precision (Micikevicius et al., 2017). This reduces the time to train networks, as modern GPUs perform 16-bit calculations very efficiently, and also reduces the required memory footprint, which is notably larger for 3D CNNs. Separate networks were trained for each fold of the cross validation experiment, to a maximum of 50 epochs, each taking up to 8 hours to train. The weights producing the lowest validation loss were retained. On occasion, a network would fail to train properly, which was evident by low training or validation classification accuracy. In

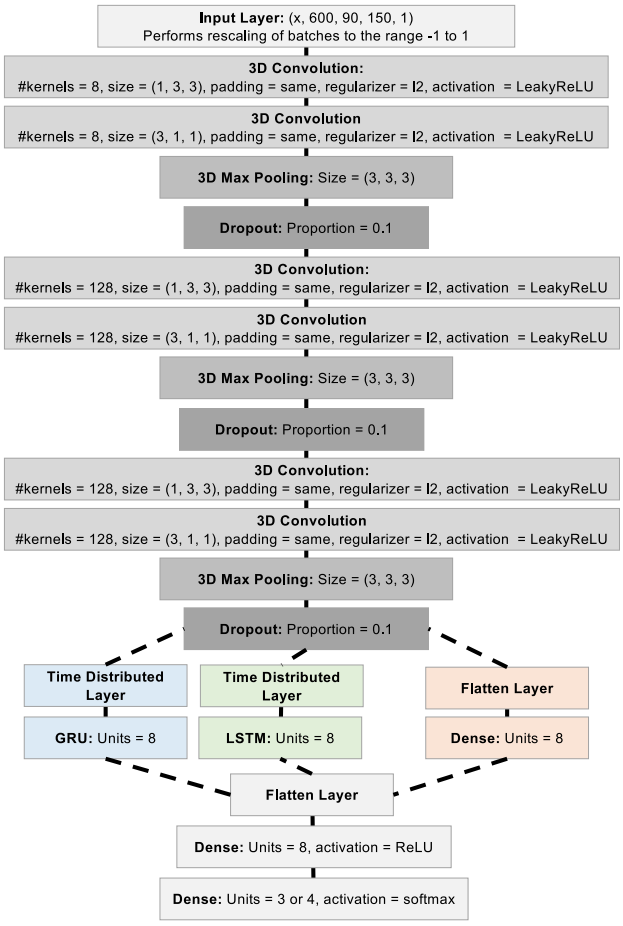

Fig. 6. Baseline neural network architecture and alternative configurations for later layers. The dotted lines highlight three separate architectures we compare in this work: A GRU layer (blue), an LSTM layer (green), and a fully connected (or Dense) layer (pink). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cases where the training or validation accuracy did not exceed 80% during training, the model weights were reinitialised and the training process was restarted.

4.3. Training batch augmentation

In this work, we explore two approaches to the generation of training batches when training our neural networks. The first is a standard approach, in which samples are randomly selected from the training set. For our experiments using normal batch selection, a batch size of eight was used. The second approach was to augment each batch of eight samples with eight *distractor* samples (Fig. 7), thus the batch size in these experiments was sixteen. Distractor samples were created as follows: Firstly, we duplicated all eight genuine samples. Then, the 600 video frames from each of these new samples were placed in reverse time order. These samples were labelled as a separate fourth class, which we term the distractor class. The distractor samples do not differ from genuine examples with respect to speaker identity, but they differ in that they contain reversed speech, which is not a genuine language. The motivation for using this technique is to encourage the network to identify the features of language, rather than using features specific to the appearance of the speaker. Further data augmentation was also used to reduce overfitting. Half of the sixteen samples within each batch were selected randomly and mirrored along the horizontal axis of each image (Shijie, Ping, Peiyi, & Siping, 2017).

Table 4
Classification accuracies (%) for experiments testing different neural networks, and experiments using the distractor class. Results of a two language experiment is also presented. Results are shown for each cross validation fold, and for 5 separate models (Rep. 1 to 5). Bold type face is used to highlight the best repetition accuracy and mean accuracy for each testing fold. This is only shown for experiments 1 to 4, as experiment 5 is not directly comparable.

Fold	Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Mean
Exp. 1: Neural Network with Dense Layer						
1	41.18	39.64	68.08	52.90	58.94	52.15
2	71.62	73.89	58.32	71.98	64.67	68.10
3	64.63	71.70	87.29	89.09	87.05	79.95
4	64.91	60.54	54.88	63.11	70.95	62.88
5	78.53	73.01	79.43	63.24	67.99	72.44
Exp. 2: Neural Network with LSTM Layer						
1	72.97	62.81	59.59	55.98	59.59	62.19
2	**86.23**	71.74	72.81	81.08	78.44	**78.06**
3	74.46	79.86	86.57	62.95	69.54	74.68
4	60.41	66.20	64.01	50.77	58.61	60.00
5	51.03	71.21	74.94	79.43	76.61	70.64
Exp. 3: Neural Network with GRU Layer						
1	67.44	56.76	60.88	63.96	72.07	64.22
2	76.17	70.78	65.15	79.52	70.90	72.50
3	81.06	84.77	87.17	71.94	**93.29**	83.65
4	60.28	62.34	77.12	64.65	55.53	63.98
5	60.28	72.62	73.01	84.19	75.19	73.06
Exp. 4: Neural Network with GRU Layer and Distractor Class						
1	78.25	80.82	**86.74**	85.20	76.96	**81.60**
2	61.92	75.93	75.69	74.25	73.29	72.22
3	92.33	88.25	86.93	85.49	86.45	**87.89**
4	**83.42**	72.88	74.16	64.14	81.88	**75.30**
5	72.88	**85.22**	84.45	80.33	73.39	**79.25**
Exp. 5: As in Exp. 4 but Testing Two Languages with 7 sec Utterances						
1	58.89	66.82	89.39	68.46	71.60	71.03
2	95.68	96.58	98.51	98.07	97.47	97.26
3	51.65	94.26	59.30	66.26	69.04	68.10
4	99.40	98.06	99.70	99.70	98.06	98.99
5	76.23	95.39	94.35	90.04	99.11	91.03

5. Results

This section presents the results of experiments undertaken in order to evaluate the language identification system developed. We first present the results of experiments using different network architectures for later layers in our baseline network. Specifically, we examine GRU, LSTM and Fully Connected (or Dense) layers. Results from five repeat experiments are presented, for each of the five cross validation folds (Table 4). We also present average classification performance for these and later experiments in Fig. 8. These first experiments used normal training batch selection (i.e. They did not include a distractor class). The architecture including the GRU layer produced the highest mean test accuracy of 71.48%, followed by the LSTM layer (69.11%) and finally the Dense layer (67.10%). According to the error bars in Fig. 8, the differences between these results are not statistically significant.

Following these results, we sought to examine the effect of using a distractor class with the best performing network. The results for Exp. 4 in Table 4 relate to the network containing a GRU layer, and using a distractor class, as described in Section 4.3. As before, results are shown for 5 repeat experiments, for each of the 5 testing folds. The average accuracy of these experiments increased by 7.77% to 79.25%. A paired samples t-test was used to compare the 25 results from these two experiments, and the increase was found to be statistically significant ($t(24) = 3.54, p = 0.0017$). Compared to the experiment using normal batch selection, the average results for each fold were higher for all folds except fold 2, which was only 0.28% lower. The greatest improvement was observed for fold 1, where the mean accuracy increased from 64.22% to 81.60% by using the distractor class. The lowest accuracy obtained across all repeats was 6.39% higher for the experiments using

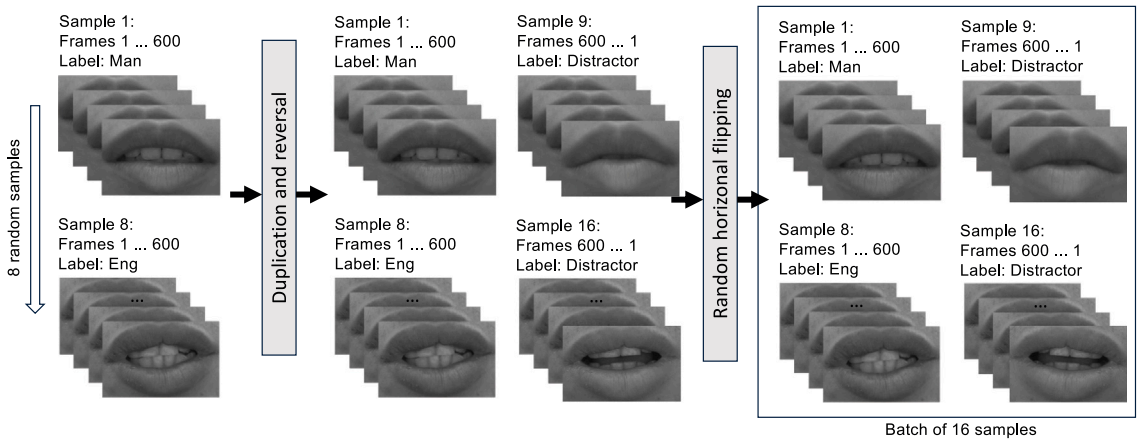

Fig. 7. Graphical illustration of the process for constructing training batches by augmenting with a distractor class. Eight genuine samples are selected and then duplicated. The new samples are reversed along the axis representing time. The final batch size is sixteen samples. Half of the samples are selected randomly and flipped (mirrored) along the longest image axis. Eng and Man refer to English and Mandarin speech, respectively.

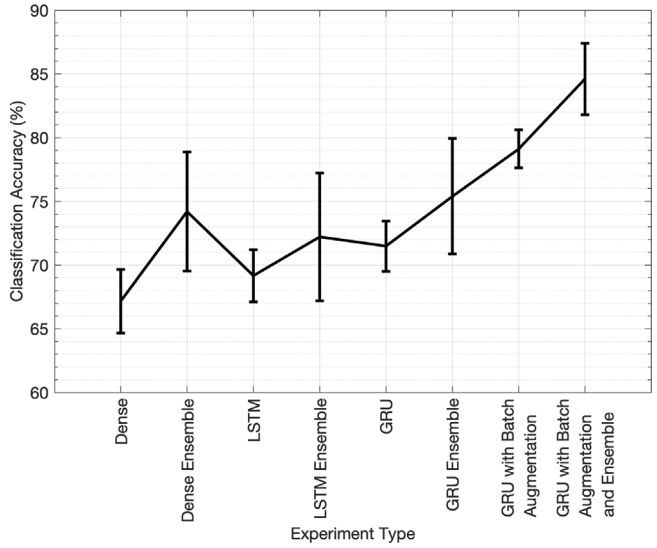

Fig. 8. Mean accuracies for each experiment undertaken. For each network, a corresponding ensemble result is also presented. Error bars indicate standard error of the mean. LSTM refers to a Long Short Term Memory network architecture, and GRU refers to a network containing Gated Recurrent Units.

Table 5
Classification accuracies for ensemble classifiers. Exp. 1 is the model with Dense layer. Exp. 2 uses an LSTM layer. Exp. 3 uses a GRU layer. Exp. 4 uses a GRU layer but is trained using the distractor class. Exp. 5 is the same as 4 except for a two language (English and Arabic) classification task using 7 s speech utterances. Bold type face is used to highlight the best results for each testing fold. This is only shown for experiments 1 to 4, as experiment 5 is not directly comparable.

Fold	Exp. 1 % acc.	Exp. 2 % acc.	Exp. 3 % acc.	Exp. 4 % acc.	Exp. 5 % acc.
1	60.75	61.26	67.70	**91.76**	72.20
2	73.41	**81.56**	76.17	77.01	98.66
3	87.17	80.34	90.29	**90.41**	81.04
4	68.38	59.13	63.62	**81.23**	100.00
5	82.13	78.92	78.66	**82.78**	98.51
Mean % acc.	74.37	72.24	75.29	**84.64**	90.01

the distractor class than for those without (61.92% versus 55.53%). Therefore, we can say with confidence that the speaker-independent classification performance improved by inclusion of the distractor class during training.

While training these networks, we noted a different pattern of learning between networks using the distractor class and those that did not use it. In Fig. 9, we show examples of the accuracy recorded across 50 training epochs for two separate networks. Fig. 9(a) relates to a network containing a GRU layer but not using the distractor class. This plot corresponds to repeat 5 from fold 1, which was the best performing repeat of this experiment. The figure shows that the training accuracy rises to over 90% after a single epoch. By the third epoch, the training accuracy is close to 100%. By contrast, the validation accuracy starts at close to 100% and then appears to fluctuate considerably over time. These results demonstrate that the network is overfitting to the training data, likely because the network quickly learns to discriminate the target classes by relying on speaker identity.

Fig. 9(b) shows data for the same network architecture and same testing fold, but this time trained using the distractor class. This data comes from repeat 3 of fold 1, which was the best performing repeat of this experiment. In (b), it takes until epoch 20 to reach a high level of accuracy on the training data. The network's accuracy on the training data is initially stable at around 50%, which represents the approximate probability of the network guessing whether a speech sample belongs to the correct language class or the distractor class. The validation accuracy increases along a similar trajectory, stabilising close to 100% after around 25 epochs and remaining stable for the final epochs.

The more gradual learning trajectory for the training data suggests that custom training batches may be steering the loss minimisation process towards an alternative minima, better suited to the task of discriminating language rather than speaker identity. Despite evidence of overfitting in both (a) and (b), (b) does show reduced overfitting as the validation data accuracy is very close to the training accuracy. While these plots are representative of the model behaviour across different repeats, sometimes the networks could achieve a similarly high level of validation accuracy without the use of a distractor class. This behaviour was unpredictable and inconsistent compared to the network training using the distractor class.

Next, we provide the results of using an ensemble of five networks including either a Dense layer, LSTM layer, or a GRU layer trained either with or without a distractor class (Table 5). We also present the result of a benchmark experiment, providing a comparison with our previous work. For every network configuration, the use of an ensemble classifier increased the classification performance (Fig. 8), with improvements from ranging from 3% to 7%. The same degree of improvement was observed for the two language (English and Arabic), benchmark experiment (Exp. 5). For the network including a GRU layer and normal batch selection, the use of an ensemble classifier increased

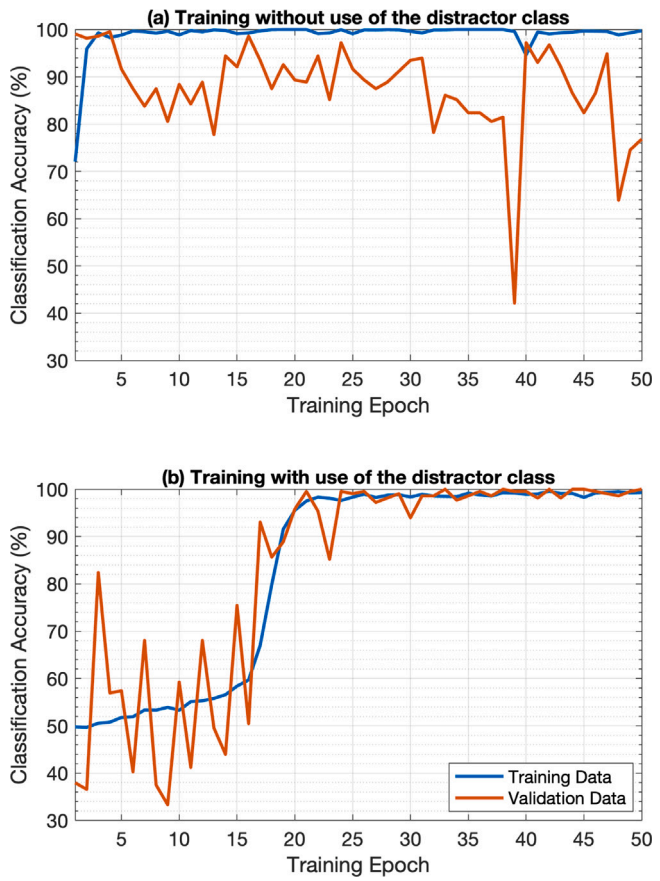

Fig. 9. Representative training and validation data accuracy plots for two experiments; one without the distractor class (a), and the other including it (b).

the mean classification accuracy by 3.81% to 75.29% (Exp. 3 in Table 5 compared to Table 4). For the GRU network using the distractor class, the ensemble classifier increased the accuracy from 79.25% to 84.64%. This three language experiment also produced the cross validation fold with the highest classification accuracy (Fold 1 at 91.76%).

The two language benchmark experiment produced a mean classification accuracy of 90.01%, which is 6.44% higher in absolute terms than our previously published results for an equivalent experiment (Newman & Cox, 2012).

In Fig. 10, we present a confusion matrix for the best three language ensemble classification experiment, using GRU layers and the distractor class. The results shown represent the classifications of each 10 s speech sample from across the five testing folds. For analysis purposes, classifications for all four classes are shown, even though the ensemble classifier only classifies utterances as one of the three genuine language classes. The matrix shows a good degree of discrimination is achieved for all languages. Mandarin and English speech samples achieve discrimination at close to 90% accuracy, with lower but still impressive performance observed for Arabic speech, at close to 66%. Some confusions are displayed for all language pairs, but the most prominent are for Mandarin and English, and Arabic and Mandarin. Few samples are misclassified as belonging to the distractor class, however Arabic has 38 more misclassifications than the next highest, English, at 5 misclassifications.

Next, we visualised a selection of the misclassifications made by the ensemble of networks containing GRU layers. Specifically, we examined the classifications generated for the best performing test fold, which was fold 1, in which ten speakers were tested. The results revealed that the spoken language of four speakers were discriminated with 100% accuracy. These four speakers comprised three English speakers and one Mandarin speaker. A further five speakers each produced a small number of misclassification, and one subject produced a comparatively poor classification accuracy.

Of the five speakers giving a small number of misclassifications, three produced a single misclassified sample (See Videos 1, 2 and 3). Two of these samples (Videos 1 and 2) contained a period in which the tracking of the mouth was shown to have failed momentarily. Video 1 showed a short period of the mouth not moving followed by brief conversational English speech. This sample occurred at the end of the recording and was misclassified as Mandarin speech. Video 2 shows Mandarin speech, which was misclassified as English speech. The third sample (Video 3), which also occurred at the end of a recording, displays almost no movement of the mouth throughout the sample, apart from a single word acknowledgement spoken in English. This sample was classified as Mandarin speech.

One speaker, reading in Arabic, was mostly classified correctly except for two samples, which were classified as Mandarin speech (Videos 4 and 5). Observing these samples provides no clear or consistent reason for these misclassifications, except that both contain a brief pursing of the lips. Similarly, one English speaker achieved an accuracy of 88.16%, with no observable issues in the misclassified samples, except that the motions of the mouth were not visibly distinctive (See Video 6 for a representative sample). A single Mandarin speaker produced a comparatively poor classification accuracy of 52.83%. Observation of one of these misclassifications revealed that the video was not in focus for extended periods of the recording (See Video 7).

6. Discussion

In our first set of experiments we found that networks incorporating GRU and LSTM layers outperformed alternative networks using fully connected layers. RNNs have previously shown to be successful when applied to computer lip reading, most likely because they explicitly capture temporal information, and the discriminative features of language are primarily temporal in nature. Our work is the first to use GRU layers for VLID, although they have previously been applied to a general computer lip reading task (Miled et al., 2023). The networks using GRU layers outperformed the corresponding networks using LSTM layers. GRU layers are not as effective at learning long-term dependencies as LSTMs (Liu, Lin, & Feng, 2021), and they are less likely to overfit the training data. Therefore, it is possible that GRUs are more effective at modelling the short term, phonotactic variations in language, which were historically the focus of traditional machine learning approaches to language identification.

In our work, we observed most confusions occurring between Arabic and Mandarin speech. This might be because Arabic is under represented in the UN2 dataset, limiting the extent to which the networks can model the temporal characteristics of Arabic speech. We also confirmed that the use of an ensemble classifier outperformed individual networks, because this approach combines the strengths of multiple classifiers, in much the same way as human experts might improve their capabilities by combining their opinions (Song, Jiao, Yang, Zhang, & Shang, 2013).

The best mean accuracy we obtained across all experiments was 84.64%, which is higher than the 83.57% accuracy we obtained for 7 s utterances in our previous work, a simpler, two-class problem (Newman & Cox, 2010). This result is also better than the 37.5% classification accuracy obtained for the 8 language task in Cascone et al. (2023), and the 73% accuracy for the 2 language task in Špetlík et al. (2017). It is also higher than the 76.3% accuracy reported by Afouras et al. (2020) for 10 s utterances, although that result is not directly comparable as it relates to a harder, 14-class problem. Similarly, the 14.3% accuracy obtained in Chandrasekhar et al. (2011) relates to a 25 language task, where each test sample consists of an entire music video.

Analysis of the misclassifications in our best performing system suggested several plausible explanations for the errors observed. Firstly,

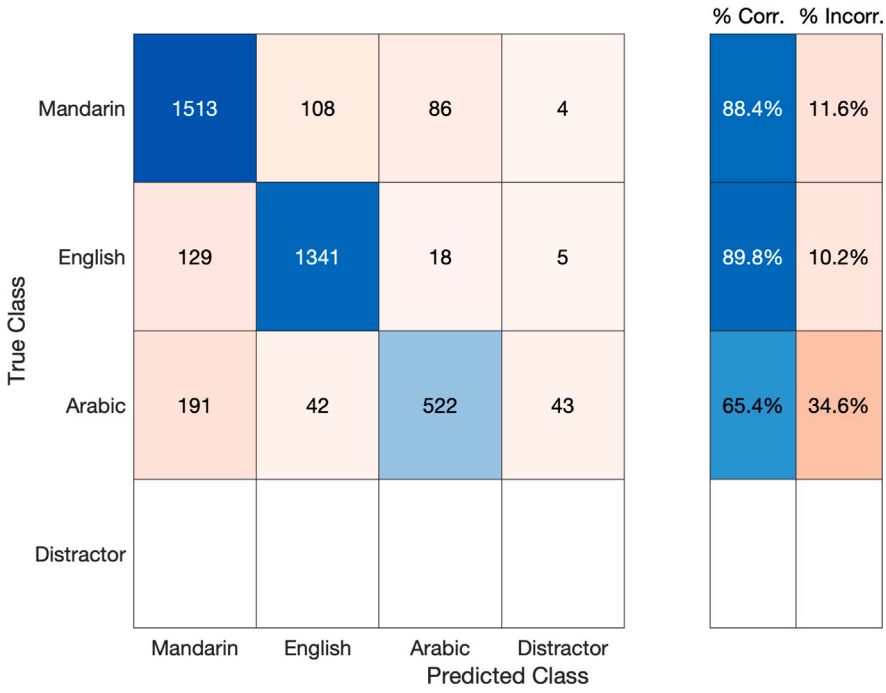

Fig. 10. Confusion matrix for the neural network incorporating a GRU layer and the distractor class during training.

classification performance is sensitive to mouth tracking errors, even if the error is only momentary. Poor camera focus was also shown to impact classification accuracy. It is reassuring that for one of the apparent misclassifications (Video 3), the sample actually contained almost no speech at all. This result confirms that the dynamics of speech are integral to the classification decisions made by our system. However, this result also motivates a need to discard periods of visual 'silence' to avoid such misclassifications. Finally, some speakers were shown to perform less well than others, with little obvious explanation for this difference. It is possible that this difference reflects the physiological variability between individuals, and that a greater number of training subjects might help to overcome this limitation.

Training networks without using the distractor class resulted in neural networks that quickly overfitted to the training data. This was clear from the training data accuracy, which typically exceeded 90% after just a couple of training epochs. We suggest that this overfitting was the result of the neural network quickly learning to discriminate the languages based on the identity of each speaker, which would be correlated with language in the training dataset but is not a general feature of language. Whilst this occasionally produced models with some discriminative capabilities, as shown in our results, in most cases these models gave comparatively poor results when applied to unseen data.

By contrast, using our distractor class approach, a different trajectory of learning was observed. Initially, training accuracy was stable, around the probability of guessing whether each sample belonged to a true language or the distractor class. Then, several epochs later, the training and validation accuracies would start to increase, before both settling close to 100%. Crucially, this was reflected in statistically significant improvements in the classification accuracy of the test data. By the nature of this speaker-independent testing framework, this means that the training batch augmentation led to improved speaker-independent performance.

7. Conclusion

In this article, we have presented a novel lip reading system, using a unique combination of 3D Convolutional Neural Network layers and Gated Recurrent Units, operating as an ensemble classifier. On a constrained language recognition task, we achieved a classification accuracy of 84.64% using 10 s test segments. We showed that the use of an ensemble classifier improved the results of all systems, including the best network, by a minimum of 3%. To assist the neural network in learning to discriminate languages rather than speaker identity, we developed a custom approach to the creation of batches used during the neural network training process. We augmented each training batch with a copy of each sample, reversing the speech in the new examples and labelling them as a *distractor* class. This approach improved the classification accuracy in comparable networks by more than 7%.

Another way to overcome the speaker-dependency of the recognition features might be to include a broader range of speakers in the training dataset, in order to provide better coverage of the feature space, as in Afouras et al. (2020), whose dataset contains over 1000 h of speech from several thousand speakers. The size of the UN2 dataset is small in the context of training a DNN, and although it was useful to consider a simpler task in order to focus on the issue of speaker-dependency, the size of the dataset used here may have limited the accuracies obtained. Having established a baseline accuracy on a smaller task, next we will explore how this approach extends to a more complex classification task, involving a larger volume of training data and variety of speakers.

It is promising that the YOLOv8 object detection algorithm was capable of tracking the mouth region with an excellent degree of accuracy. It would be interesting to undertake a more detailed comparison between the performance of this approach and that of a cascade classifier, especially with regard to diversity of individuals and variations in presentation (e.g. image cropping, image size, and perspective of the face). Another object detection algorithm, such as RetinaNet (Cheng et al., 2020), could also be considered. Given the apparent degradation of lip reading performance when mouth tracking fails, further work could focus on making the system robust to such errors.

Recent work into VLID has abandoned the previous phonotactic based approaches which persisted for many years (Zissman, 1996). We would like to explore whether viseme tokenisation followed by language modelling could be used to provide language discrimination capabilities beyond those reported here. A deep learning approach to

speech recognition, such as wav2vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020), could be used to tokenise speech. In wav2vec 2.0, sub units of audio speech are learnt via a contrastive learning approach, in which short durations of speech are hidden and the network learns to guess the missing speech. Such an approach could be applied to visual speech and combined with language modelling for VLID.

We used a simple but effective approach to ensemble classification, but it might be more effective to explore the use of an integrated neural network, in which the outputs from parallel networks are combined into an output network, and all networks are trained concurrently. Also, instead of combining five networks with the same architectures, a combination of disparate networks could be used, each offering different strengths.

In summary of the advantages and limitations of this work compared to other studies, the primary benefit of the system discussed here lies in the inclusion of the distractor class, which is an entirely novel approach not previously presented by other studies. Although the use of this approach during training did not entirely eliminate overfitting to the training data, reduced overfitting was evident through increased validation and test data accuracies.

The main limitation of this work lies in the size and constrained nature of the dataset used, which is considerably smaller than that used by some other studies. Thus, the narrower coverage of the feature space provides a plausible explanation for why a small number of speakers were very challenging to lip read, despite no obvious abnormalities with the data.

In conclusion, this study has confirmed findings from the literature, including our own work, that it is possible to distinguish spoken language using computer lip reading. These results were shown to exceed those of our previous work, and that of other studies.

Code and data availability

The data used in this article are available upon reasonable request. The code used and a description of how to use it can be found at https://github.com/JNewmanUEA/VLID.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work would not have been possible were it not for the University of East Anglia's High Performance Computing Cluster and its team of engineers. We would also like to acknowledge guidance from Professor Stephen Cox, and discussions with Harry Rogers and Professor Richard Harvey. The LILiR team were involved with design decisions regarding the video dataset used in these experiments. This work was originally funded by the EPSRC under EP/E028047/1.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.iswa.2025.200517.

References

Afouras, T., Chung, J. S., & Zisserman, A. (2020). Now you're speaking my language: Visual language identification. In *Proceedings of ISCA 2020*. (no. 2020), ISCA Archive.

Al-obidi, D., & Kacmaz, S. (2023). Facial features recognition based on their shape and color using YOLOv8. In *2023 7th International symposium on multidisciplinary studies and innovative technologies* (pp. 1–6). IEEE.

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills (L). *Journal of the Acoustical Society of America*, *130*(1), 1–4.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Bear, H. L., & Harvey, R. (2019). Alternative visual units for an optimized phoneme-based lipreading system. *Applied Sciences*, *9*(18), 3870.

Bernstein, L. E., Jordan, N., Auer, E. T., & Eberhardt, S. P. (2022). Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training. *American Journal of Audiology*, *31*(2), 453–469.

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252.

Biswas, M., Rahaman, S., Ahmadian, A., Subari, K., & Singh, P. K. (2023). Automatic spoken language identification using MFCC based time series features. *Multimedia Tools and Applications*, *82*(7), 9565–9595.

Cascone, L., Nappi, M., & Narducci, F. (2023). Language identification as improvement for lip-based biometric visual systems. In *2023 IEEE international conference on image processing* (pp. 1570–1574). IEEE.

Chandrasekhar, V., Sargin, M. E., & Ross, D. A. (2011). Automatic language identification in music videos with low level audio and visual features. In *2011 IEEE international conference on acoustics, speech and signal processing* (pp. 5724–5727). IEEE.

Cheng, M., Bai, J., Li, L., Chen, Q., Zhou, X., Zhang, H., et al. (2020). Tiny-RetinaNet: a one-stage detector for real-time object detection. In *Eleventh international conference on graphics and image processing*: *vol. 11373*, (pp. 195–202). SPIE.

Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B.-J. (2008). The challenge of multispeaker lip-reading.. In *AVSP* (pp. 179–184). Citeseer.

Exarchos, T., Dimitrakopoulos, G. N., Vrahatis, A. G., Chrysovitsiotis, G., Zachou, Z., & Kyrodimos, E. (2024). Lip-reading advancements: A 3D convolutional neural network/long short-term memory fusion for precise word recognition. *BioMedInformatics*, *4*(1), 410–422.

Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. (2021). Deep learning-based automated lip-reading: A survey. *IEEE Access*, *9*, 121184–121205.

Heracleous, P., Takai, K., Yasuda, K., Mohammad, Y., & Yoneyama, A. (2018). Comparative study on spoken language identification based on deep learning. In *2018 26th European signal processing conference* (pp. 2265–2269). IEEE.

Jannah, M., Mashalani, F., Lubis, Y., & Amaro, J. C. (2023). Phonology decoded: An exploration into the intricacies of language sound systems. *Socio-Economic and Humanistic Aspects for Township and Industry*, *1*(1), 127–131.

Komatsu, M. (2007). Reviewing human language identification. *Speaker Classification II: Selected Projects*, 206–228.

Kopuklu, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Lan, Y., Harvey, R., & Theobald, B.-J. (2012). Insights into machine lip reading. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 4825–4828). IEEE.

Liu, X., Lin, Z., & Feng, Z. (2021). Short-term offshore wind speed forecast by seasonal ARIMA-a comparison against GRU and LSTM. *Energy*, *227*, Article 120492.

Margam, D. K., Aralikatti, R., Sharma, T., Thanda, A., Roy, S., Venkatesan, S. M., et al. (2019). LipReading with 3D-2d-CNN BLSTM-HMM and word-CTC models. arXiv preprint arXiv:1906.12170.

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*, 135–164.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, Article 101869.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., et al. (2017). Mixed precision training. arXiv preprint arXiv:1710.03740.

Miled, M., Messaoud, M. A. B., & Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications*, *82*(1), 551–571.

Mohapatra, C., Dash, S., & Majhi, U. (2016). A comprehensive review of the speech dependent features and classification models used in identification of languages. *International Journal of Computer Applications*, *147*(5).

Newman, J. L., & Cox, S. J. (2009). Automatic visual-only language identification: A preliminary study. In *2009 IEEE international conference on acoustics, speech and signal processing* (pp. 4345–4348). IEEE.

Newman, J. L., & Cox, S. J. (2010). Speaker independent visual-only language identification. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 5026–5029). IEEE.

Newman, J. L., & Cox, S. J. (2012). Language identification using visual features. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(7), 1936–1947.

Nosouhian, S., Nosouhian, F., & Khoshouei, A. K. (2021). A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU.

Preethi, S., et al. (2023). Analyzing lower half facial gestures for lip reading applications: Survey on vision techniques. *Computer Vision and Image Understanding*, Article 103738.

Rothkrantz, L. (2017). Lip-reading by surveillance cameras. In *2017 Smart city symposium prague* (pp. 1–6). IEEE.

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). Research on data augmentation for image classification based on convolution neural networks. In *2017 Chinese automation congress* (pp. 4165–4170). IEEE.

Singh, P., Manure, A., Singh, P., & Manure, A. (2020). Introduction to tensorflow 2.0. In *Learn TensorFlow 2.0: Implement machine learning and deep learning models with python* (pp. 1–24). Springer.

Song, X., Jiao, L., Yang, S., Zhang, X., & Shang, F. (2013). Sparse coding and classifier ensemble based multi-instance learning for image categorization. *Signal Processing*, *93*(1), 1–11.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*(2), 218–231.

Špetlík, R., Čech, J., Franc, V., & Matas, J. (2017). Visual language identification from facial landmarks. In *Image analysis: 20th scandinavian conference, SCIA 2017, tromsø, Norway, June 12–14, 2017, proceedings, part II 20* (pp. 389–400). Springer.

Taylor, S., Theobald, B., & Matthews, I. (2014). The effect of speaking rate on audio and visual speech. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 3037–3041). IEEE.

Taylor, S., Theobald, B.-J., & Matthews, I. (2015). A mouth full of words: Visually consistent acoustic redubbing. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4904–4908). IEEE.

Ulbrich, C., Alday, P. M., Knaus, J., Orzechowska, P., & Wiese, R. (2016). The role of phonotactic principles in language processing. *Language, Cognition and Neuroscience*, *31*(5), 662–682.

Van Segbroeck, M., Travadi, R., & Narayanan, S. S. (2015). Rapid language identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(7), 1118–1129.

Wang, Y.-Q. (2014). An analysis of the viola-jones face detection algorithm. *Image Processing on Line*, *4*, 128–148.

Wang, H., Pu, G., & Chen, T. (2022). A lip reading method based on 3D convolutional vision transformer. *IEEE Access*, *10*, 77205–77212.

Zacharaki, K., & Sebastian-Galles, N. (2021). The ontogeny of early language discrimination: Beyond rhythm. *Cognition*, *213*, Article 104628.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, *4*(1), 31.