Received: 30 March 2022 | Revised: 15 July 2022 | Accepted: 20 July 2022

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12131

ORIGINAL RESEARCH

Developing phoneme-based lip-reading sentences system for silent speech recognition

Randa El-Bialy[1,2] | Daqing Chen[1] | Souheil Fenghour[1] | Walid Hussein[2] | Perry Xiao[1] | Omar H. Karam[2] | Bo Li[3]

[1]School of Engineering, London South Bank University, London, UK

[2]Faculty of Informatics and Computer Science, British University in Egypt, Cairo, Egypt

[3]School of Electronics and Informatics, Northwestern Polytechnical University, Xi'an, China

Correspondence

Randa El-Bialy, School of Engineering, London South Bank University, London, UK.
Email: elbialyr@lsbu.ac.uk

Abstract

Lip-reading is a process of interpreting speech by visually analysing lip movements. Recent research in this area has shifted from simple word recognition to lip-reading sentences in the wild. This paper attempts to use phonemes as a classification schema for lip-reading sentences to explore an alternative schema and to enhance system performance. Different classification schemas have been investigated, including character-based and visemes-based schemas. The visual front-end model of the system consists of a Spatial-Temporal (3D) convolution followed by a 2D ResNet. Transformers utilise multi-headed attention for phoneme recognition models. For the language model, a Recurrent Neural Network is used. The performance of the proposed system has been testified with the BBC Lip Reading Sentences 2 (LRS2) benchmark dataset. Compared with the state-of-the-art approaches in lip-reading sentences, the proposed system has demonstrated an improved performance by a 10% lower word error rate on average under varying illumination ratios.

KEYWORDS

deep learning, deep neural networks, lip-reading, phoneme-based lip-reading, spatial-temporal convolution, transformers

1 | INTRODUCTION

In recent decades, decoding speech from visual cues imitating the human capability to perform lip-reading has drawn much research attention. Speech is an audio-visual signal consisting of audio vocalisation and its equivalent mouth movement. Visual Speech Recognition, also known as lip-reading or automatic lip-reading, is the process of understanding speech by analysing lip movements. Developing such systems depends on the information provided by the context of speech and the knowledge of the language being spoken. Accordingly, lip-reading is considered complementary information to compensate for the lack of audio information. Recently, the attention towards lip-reading has grown rapidly, given that visual information is not affected by acoustic noise and therefore is robust in noisy environments, and has led to significant improvements in performance [1, 2].

Compared to audio signals, decoding speech from visual signals faces significant challenges due to visual ambiguity and the absence of context. Visual ambiguity is considered one of the main problems lip-reading systems suffer from, which arises at the word level due to homophemes that produce the same or almost the same lip movement (e.g., [b], [*p*], and [m]).

Usually, phonemes, not characters, are considered the standard minimum units in speech processing and are defined as the minimum distinguishable sounds that can change the meaning of a word [3]. In the same way, a viseme is the minimum distinguishable speech used for analysing visual information [4].

Other challenges include poor temporal resolution, efficient encoding of spatial-temporal information, speaker dependency, head pose variation with different view angles, and illumination conditions, extracting lip's contours from different types of background, such as static and rotating backgrounds, and tackling the effect of different face structures, etc. [5]. Different pronunciations exist due to different dialects of people in different regions. Also, in reality, some people have short lip movements compared to others. Furthermore, people can talk from different angles towards a camera. Because of these issues, it is essential to create more robust models [6].

With the power of Deep Learning (DL) architectures and the availability of large-scale databases, it is possible to shift from the early lip-reading systems, which addressed simple word recognition tasks, to more realistic and complex tasks [7]. Accordingly, these DL architectures have led to current systems that target continuous lip-reading [8, 9] and to improve the performance of visual speech recognition in general. However, due to the complexity of image processing and the difficulty of training classifiers, it is difficult for traditional lip-reading systems to meet the requirements of real-time applications. As a result of the advancements in lip-reading systems, numerous applications are conceivable, for example, resolving multi-talker simultaneous speech [10], developing augmented lip views to assist people with hearing impairments [11], dictating messages to smartphones in noisy environments [12], transcribing and re-dubbing silent films [8], and discriminating between native and non-native speakers [13].

Currently, there are two leading approaches to solving the lip-reading problem. The first approach handles it as a word or phrase classification task. This approach uses video samples to predict a word or phrase label [14]. The second is a more recent one, which has gained its strength from the deep network's capability to perform text predictions, such as complete sentences. Accordingly, instead of predicting word labels for solving lip-reading problems, this approach predicts character sequences or a viseme sequence [8, 15, 16]. As for visual speech units, the two primary forms are phonemes and visemes. Phonemes are strongly linked to an acoustic speech signal [17]. A viseme, on the other hand, is the most basic visual unit of speech, reflecting a gesture of the mouth, face, and visible elements of the teeth and tongue, also known as visible articulators [4]. Even though phonemes represent distinct short sounds, some studies employ phoneme units to increase lip-reading accuracy [18], while others focus on visemes. The efficacy of phoneme or viseme units in lip-reading systems is a point of contention.

Using phonemes for lip-reading sentences has some advantages over other systems since it overcomes the cumulative loss of information caused by the mapping process from phoneme classes (the number of classes is between 45 and 53) to viseme classes (the number of classes is between 10 and 14) [19]. However, due to the reduction of a set of phonemes to a set of visemes, the complexity of the pronunciation dictionary increases due to the increasing volume of homophonic words, and the discriminative power of the classification model is reduced. Essentially, there is a trade-off between unit and model accuracy at the sentence level.

Another problem that a viseme-based system suffers is the large number of the proposed phoneme-to-visemes maps as a phoneme is related to one viseme class but a viseme may represent many phonemes. This cause ambiguity between phonemes when using viseme classifiers as an illustration the viseme class ‘FV’ can be mapped into ‘ae’ ‘eh’ ‘ay’ ‘ey’ ‘hh’ phoneme classes [20, 21]. In comparison, only two variations of phoneme dictionaries have been used. In addition, most English words have a one-to-one mapping to a word with only a few exceptions with a one-to-many relationship to a set of words. Therefore, converting recognized phonemes to words will have less complexity and require less computational effort.

This paper uses phonemes as a classification schema for lip-reading sentences in the wild rather than character-based or visemes-based schemas. The main aim of this research is to explore an alternative schema and to enhance system's performance. The proposed system's performance has been validated using the BBC Lip Reading Sentences 2 (LRS2) benchmark dataset. The system displayed a 10% average reduction in a word error rate under varying illumination ratios compared to the state-of-the-art systems in lip-reading sentences.

The rest of this paper is organized as follows: Section 2 provides a literature review on phoneme-based lip-reading systems and discusses the different architectures used in feature extraction, phoneme labelling, and classifiers. In addition, the relevant works on Automatic Speech Recognition (ASR) are discussed. Section 3 discusses the methodology and the proposed system in detail, including its pre-processing steps, the structure of the visual front-end model, the phoneme recognition model, the pronunciation dictionary used, and the Recurrent Neural Network-based language model is explained. Section 4 briefly discusses the BBC Lip Reading Sentences 2 benchmark dataset. Section 5 presents Models comparison, which addresses the details of the state-of-the-art character-based lip-reading system and a viseme-based lip-reading system. Section 6 presents the experimental results and demonstrates the performance of the proposed model with evaluation. In addition, presenting how to add noise to testify the robustness of the proposed model, Gamma correction has also been considered in the experiments. Finally, concluding remarks and future work are given in Section 7.

2 | RELATED WORKS

Phonemes are mainly used with acoustic signals, considered the main building blocks of speech. However, in scenarios where audio signals are corrupted or unavailable, in noisy environments, or in the case of individuals with partial or total hearing loss, it will usually be challenging to detect audio signals. Accordingly, lip-reading is a complementary method to compensate for the lack of audio information. The literature can therefore be divided into two directions. The first one is where audio signals are absent, and only video signals are available. The second direction is where only audio signals are present. As phonemes have traditionally been associated with sound or audio, research studies are rare in relation to video-based lip-reading. This section reviews the literature on video-based phoneme recognition for lip-reading and some of the relevant works on Automatic Speech Recognition (ASR).

According to Howell et al. [22], an approach was proposed to treat visual speech as dysarthria to compensate for the gap of having a reduced phonemic repertoire. For visual feature extraction, the Active Appearance Model was used, the Hidden Markov Model was utilised for phoneme recognition, and Weighted Finite-State Transducers were employed for word recognition. The dataset was captured from a single female speaker who spoke six repetitions of a set of 112 isolated words. The word accuracy rate was 49.70%.

Noda et al. [23] were the first to apply Convolutional Neural Networks (CNN) for feature extraction in visual speech recognition systems and also for phoneme recognition as their purpose was to prove whether feature extraction mechanisms using CNN would outperform other models, which use the classic dimensionality reduction techniques. As for identifying isolated words, HMM was used. The dataset used contained six different speakers pronouncing 300 Japanese words. The average phoneme recognition rate was 58%, and the accuracy for word recognition was 37%.

Thangthai et al. [18] compared viseme-based and phoneme-based lip-reading systems to add more evidence to the argument that phonemes can surpass visemes lip-reading systems, and they suggested that phonemes are the current optimal class labels for lip-reading. Using the TCD-TIMIT corpus for sentences, Discrete Cosine Transformer and Eigenlip for feature extraction, and Weighted-Finite-state Transducer as word recogniser, the phoneme recognition accuracy acquired was 33.44%. Furthermore, the word accuracy rate was 48.74% in speaker-dependent tests using Eigenlip compared to 46.6% and 33.06% for viseme and word recognition, respectively.

To enhance the accuracy of phoneme-based lip-reading systems, Shillingford et al. [24] proposed a Deep Neural Network and a production-level speech decoder for both mapping videos into a sequence of phoneme distributions and generating the corresponding word sequences, respectively. A Large-Scale Visual Speech Recognition dataset was constructed and used in the study. Spatio-temporal Convolutions, Bi-directional Long Short-Term Memory, and Finite-State Transducers were utilised for feature extraction, phoneme recognition, and word recognition. The phoneme recognition accuracy rate was 66%, and the word accuracy rate was 60%.

The relevant works on Automatic Speech Recognition are discussed below.

Chiu et al. [25] presented a so-called Listen, Attend, and Spell (LAS) architecture, an attention-based encoder-decoder, in which traditional automatic speech recognition system components were included in a single neural network. They proved that graphemes could be substituted with a word piece model. The work has shown that the performance of ASR can be significantly improved by optimising the LAS model and introducing a multi-head attention architecture. Also, they improved the accuracy by exploring synchronous training, scheduled sampling, label smoothing, and minimum word error rate optimization. The experiments were conducted on a 12,500-h voice search task (Google Voice Search).

The work by Anjie Fang et al. [26] aimed to present a classification model that is robust to ASR errors and acquires pronunciation similarities ignored in word-level representations by creating an ASR transcription at the phoneme level. Four existing datasets were used in the research, including the Stanford Sentiment Treebank (SST), The TREC Question classification (TQ), SQuAD, and the subjectivity dataset (SUBJ) by generating noisy ASR transcriptions for them. The authors demonstrated the integration of phoneme embedding into existing neural network architectures and the improvement of classification models when handling data containing ASR errors. The accuracies for SST, TQ-50, and TQ6, were 41%, 65%, and 75%, respectively.

A comparison was conducted by Mohammad Zeineldeen et al. [27] between phoneme-based and grapheme-based output labels utilising the encoder-decoder-attention ASR model. Furthermore, the use of byte-pair-encoding (BPE)-based phonemes and single phonemes as output labels was investigated with a conclusion that both had a similar performance, and this has further proven that phoneme-based models are competitive to grapheme-based models. Switchboard 300 h and LibriSpeech 960 h benchmarks were used to conduct the experiments. As a result of these experiments, the accuracies obtained when using the switchboard 300 h dataset for BPE-based grapheme were 85%, and 86.2% was achieved for both single and BPE-based phonemes. As for the Librispeech 960 h dataset, the accuracies acquired were 89.44%, 86.2%, and 90.86 for BPE-based phoneme, single phoneme, and BPE-based grapheme, respectively. As such, it was observed that grapheme and phoneme-based BPE outperform single phonemes on Librispeech 960 h, which contradict the results of Switchboard 300 h.

Wei Zhou et al. [28] adopted a simple competitive approach for phoneme-based neural transducer modelling, sustaining the advantages of both classical and end-to-end approaches. In order to maintain the sequence-to-sequence modelling consistency, a simplified neural network structure along with direct integration with an external word-level language model was presented by utilising the local dependencies of phonemes. Furthermore, augmentation for word-end-based phoneme labels was proposed to improve the system performance. Furthermore, frame-wise cross-entropy loss was used for an efficient training procedure. The proposed model was evaluated on both TED-LIUM release 2 (TLv2) and Switchboard (SWBD) corpora, and the word error rate obtained was 6.3% and 11.5, respectively (Tables 1 and 2,).

As shown in the literature, research on phoneme-based lip-reading systems is very limited, and to the best of the authors' knowledge, this study is the first work that purely uses phonemes from videos for lip-reading sentences in the wild. Most of the time, using phonemes is associated with audio signals; however, in this research, the audio is not presented/provided as in some scenarios and potential applications, such as CCTV footage analysis, forensic investigations, silent dictation in public places, wearable optical technologies to aid hearing, animation and digital avatars, silent movies restoration, and last but not least, humanoid robotics.

TABLE 1 Phoneme-based lip-reading systems

References	Dataset	Year	Signal	Feature extraction	Phoneme recognition	PAR	Classification	Classifier	WAR (%)
Howell et al. [22]	---------	2013	Video	AAM	HMM	------	Isolated words	WFTS	49.7
Noda et al. [23]	-------	2014	Video	CNN	CNN	58%	Isolated words	HMM	37
Thangthai et al. [18]	TCD-TIMIT	2017	Video	Eigenlip	Hybrid DNN-HMM	33.4%	Sentences	WFTS	48.7
Shillingford et al. [24]	LSVSR	2018	Video	Spatio-temporal convolutions	Bi-LSTM	66%	Sentences	FST	60

TABLE 2 Works on ASR in the literature.

Reference	Dataset	Year	Signal	Model	WAR
Chiu et al. [25]	a 12, 500-h voice search task	2018	Audio	Attention-based encoder-decoder	94%
Fang et al. [26]	Amazon ALEXA DATA	2020	Audio	CNN	76%
Mohammad Zeineldeen et al. [27]	Switchboard 300h	2020	Audio	Attention-based encoder-decoder	BPE-grapheme 85%
					Single-phoneme 86.2%
					BPE-phoneme 86.2%
	LibriSpeech				BPE-grapheme 90.8%
					Single-phoneme 86.3%
					BPE-phoneme 90.7%
Wei Zhou et al. [28]	TLv2	2021	Audio	LSTM-T	93%
	SWBD				88.5

3 | METHODOLOGY

The objective of the proposed system is to predict sentences being spoken from silent videos by extracting lip movements and decoding the movements into phonemes. This section discusses the different processing steps of the lip-reading system under consideration.

The first step is pre-processing, where facial landmark detection is utilised to extract the lip contour as the region of interest from videos. In the second step, a spatio-temporal visual front-end uses a sequence of images of loosely cropped lip regions as input to generate and outputs a feature vector per frame as outputs. Finally, in the third step, a sequence processing module inputs the sequence of per-frame feature vectors to the phoneme classifier to recognise the phonemes; then, a language model is used to convert the phonemes into words and outputs a sentence. The overall system diagram is shown in Figure 1.

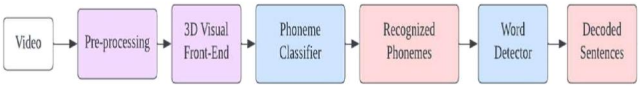

FIGURE 1 Overall lip-reading system components

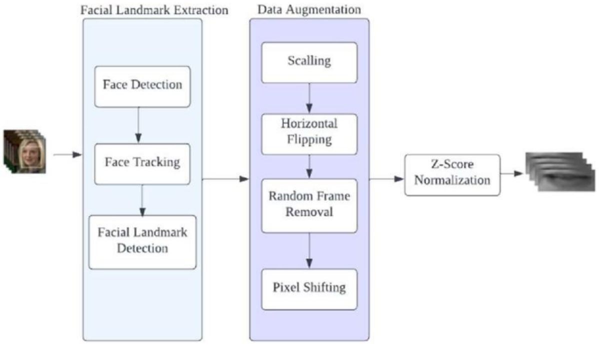

FIGURE 2 Pre-processing steps for video

3.1 | Pre-processing

All of the videos are pre-processed as shown in Figure 2. With 25 frames per second framing rate, images with red, green, and blue pixel values with a resolution of 160 pixels by 160 pixels are utilised. Because the region of interest (ROI) and feature input to the visual front end are speaker's lips, the following steps for video pre-processing are:

- Step 1: Sample videos are into image frames.
- Step 2: Identify face landmarks with the videos sampled. Based on iBug [29], a Convolutional Neural Network detector known as The Single Shot Multi-Box Detector (SSD) [34], facial landmarks are extracted by detecting face presence in every single frame.

- Step 3: Generate image dimensions of 112 × 112 × T dimensions (where T is the number of image frames) by converting each video frame to a greyscale, followed by scaling and cropping in the centre of the facial landmark boundary.
- Step 4: Conduct data augmentation with horizontal flipping, random frame removal [30, 31], and random shifts in the temporal and spatial dimensions of ±2 frames and ±5 pixels, respectively.
- Step 5: Normalise each pixel in a frame to its overall mean and variance.

3.2 | Visual front-end model

The model presented in [32] has served as the foundation for the spatial-temporal visual front-end model. As the outputs of the pre-processing step, the frame sequence that contains the cropped lips is then inputted to a spatial-temporal (3D) convolution with a filter width of 5 frames to best capture the short-term dynamics of the mouth region and outputs a 3D feature map. A 2D ResNet is then employed to use these feature maps to reduce the spatial dimensions. The output is a $T \times \frac{W}{32} \times \frac{H}{32} \times 512$ tensor for an input sequence of $T \times W \times H$ (*Time/width/Height*) frames, and it is then average-pooled over the spatial dimensions, producing a 512-dimensional feature vector for every input video frame.

Illustrated in Figure 3, the input image frames are fed into a 3D CNN, and the network captures the short-term dynamics of a mouth area. Further, the 3D feature maps are fed into a 2D ResNet, which reduces the spatial dimensions to a single-dimensional tensor at every time step.

3.3 | Phoneme recognition model

In this study, the Carnegie Mellon Pronunciation Dictionary [33] is utilised to map (convert) a sequence of words into a sequence of phonemes to produce labels for phoneme classification. A Transformer uses 44 classes in total to predict phonemes in silent videos. Also, a padding character where each video is added to 180 characters is to ensure that all videos are of the same length, a space character, Start of Sentence <SoS>, and End of Sentence <EoS>. Table 3 illustrates the classes used by the phoneme classifier.

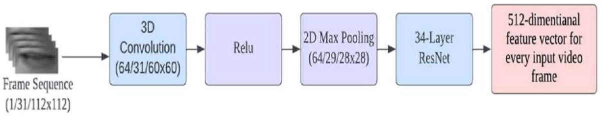

FIGURE 3 Visual Front-end Model. The input image frames are input to a 3D CNN to capture the short-term dynamics of a mouth region, and then the outputted 3D feature maps are used as input to a 2DResNet to reduce the spatial dimensions to a single-dimensional tensor per time step

TABLE 3 Phonemes as classes

{[pad], <sos> ‘AA’ , ‘AE’, ‘AH’, ‘AH’, ‘AO’, ‘AW’ , ‘AY’ , ‘B’ , ‘CH’ , ‘D’ , ‘DH’ , ‘EH’ , ‘EH’ , ‘ER’, ‘EY’, ‘F’, ‘G’ , ‘HH’, ‘IH’, ‘IY’, ‘JH’,’K’ , ‘L’ , ‘M’ , ‘N’ , ‘NG’, ‘OW’ , ‘OY’ , ‘P’, ‘R’ , ‘S’ , ‘SH’, ‘T’, ‘TH’,’UH’, ‘UW’, ‘V’, ‘W’, ‘Y’, ‘Z’, ‘ZH’, <eos>, [space]}.

As the primary building block in an encoder-decoder architecture, as illustrated in Figure 4, multi-headed attention is implemented by Transformers [34]. A stacked self-attention layer with the input tensor as attention queries and keys and values constitutes the encoder. As for the decoder, it follows the model presented in [15] and consists of a dense layer, batch normalisation, RELU, and a dropout layer probability of 0.1 for each of the three fully connected layers; 1024 nodes in the

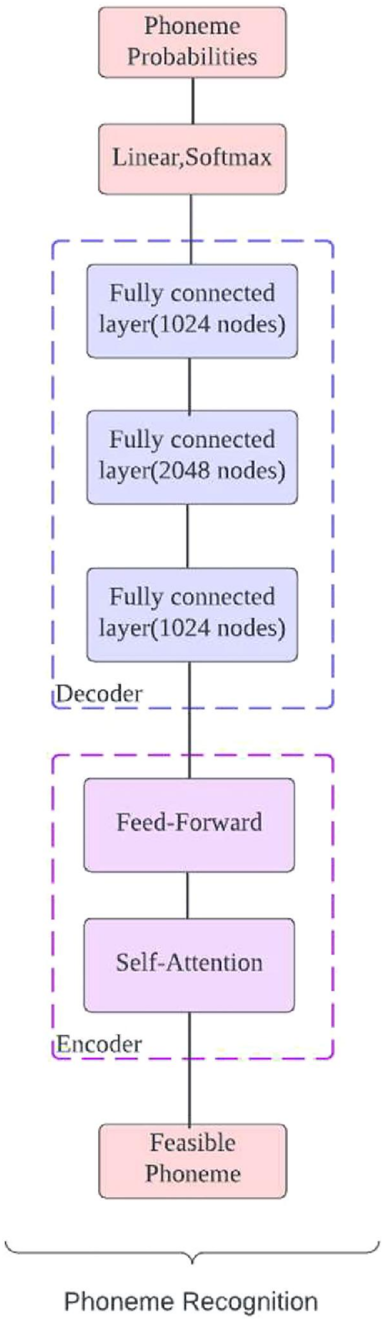

FIGURE 4 Phoneme transformer architecture

first and the last fully connected dense layers, and as for the dense middle layer, it consists of 2048 nodes. Furthermore, the decoder generates phoneme probabilities with a cross-entropy loss function corresponding to the ground truth table. The encoder utilises the [34] base model, which has six layers, a model size of 512, eight attention heads, and a dropout probability of 0.1.

3.4 | Language model

An attention-based RNN is utilised to convert the recognized phonemes into meaningful sentences [35]. As shown in Figure 5, the network consists of two multilayer RNNs: an encoder for the source phoneme sequences and a decoder for the target word sequences. By initialising the decoder with the last hidden state of the encoder, the decoder will gain access to the source information. The main goal of the attention mechanism is to create direct connections between the target and the source. At every decoder time step, the following process of attention computations takes place. First, all source states and current target hidden states are compared to produce/drive attention weights as in Equation (1). Second, a context vector is computed based on the attention weights as shown in Equation (2). Third, as shown in Eqution (3), in order to produce the attention vector, the current target hidden state is combined with the context vector and then fed as input to the following time step, where α_{ts} represents attention weights, h_t is the target hidden state, h_s is the source hidden state, C_t is the context vector, and a_t is the attention vector.

$$\alpha_{ts} = \frac{\exp(\text{score}\ (h_t, h_s))}{\sum_{s'}^{S} \exp(\text{score}\ (h_t,\ h_{s'}))} \tag{1}$$

$$C_t \ = \ \sum_{s} \alpha_{ts} h_s \tag{2}$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \tag{3}$$

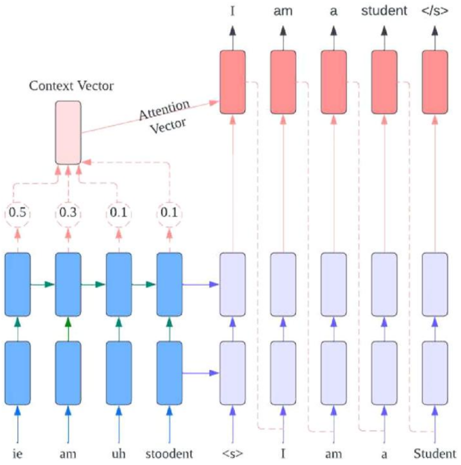

FIGURE 5 Encoder-decoder-attention architecture

4 | DATASET USED

The BBC Lip-reading Sentences 2 (LRS2) [36] dataset has been used for this work. There are over 46,000 videos in total in the dataset with over two million-word occurrences and a vocabulary of approximately 40,000 words. The video with the most extended duration is 180 frames long with each video having a frame rate of 25 frames per second. The dataset consists of spoken sentences, where each sentence is up to 100 ASCII characters extracted from BBC videos, with a range of facial positions from frontal to profile. The dataset is quite challenging because of the variety of perspectives, lighting settings, genres, and speakers.

5 | MODELS COMPARISON

Recently, two prominent techniques are used to tackle the lip-reading challenge. The first technique treats it as a word or phrase recognition problem, where video samples are analysed to predict a word or phrase label. The second technique, the latest solution, addresses the lip-reading problem by predicting a viseme sequence or a character sequence rather than a word label. In this section, a comparison is provided regarding the work of both [15, 37].

Lip-reading systems consist of 3 stages; the first stage is the pre-processing stage, where videos are input into the system and apply facial landmark extraction, which consists of face detection, face tracking, and facial landmark detection, grey-scale conversion, scaling, central cropping horizontal flipping, random frame removal, pixel shifting, and Z-score normalisation to extract the felicitous region of interest (ROI). Next, the extracted ROI as a sequence of frames is input into the visual front-end model, where a spatial-temporal (3D) convolution is applied. Subsequently, a 2D ResNet is utilised to decrease the spatial dimensions with depth. Accordingly, the output would be a 512-dimensional feature vector for each input video frame.

The second stage depends on the classification scheme [37] using characters for labelling the videos with 26 classes; the authors in [15] use visemes with 13 classes; the authors in [37] discussed three models for this task: the first is a recurrent model consisting of stacked Bidirectional LSTM layers, the second model is fully convolutional, and the third is a Transformer model that follows an encoder-decoder structure with multi-head attention layers as a building block. The authors in [15] presented a Transformer model with a different decoder and a dense layer structure than the work presented in [37] due to the difference in nature between visemes and characters.

The third stage is the language model, where the input is the labels for each sequence of frames and outputs of the uttered sentence. For example, the author [37] uses a character-level external language model consisting of four unidirectional layers of a Recurrent Neural Network with 1024 LSTM cells each that outputs a sentence character by character. As for [15], a word detector consists of two steps: the first step is a word lookup step and the second is Perplexity Calculations.

The datasets used by [37] are Lip Reading in the Wild (LRW) and the Lip Reading Sentences 2 (LRS2); the authors in [15] used the Lip Reading Sentences 2 (LRS2). Both [15, 37] trained their models on a single GeForce GTX 1080 Ti GPU with 11 GB memory and implemented all operations in TensorFlow.

According to [37], Transformer is the best performing network among the three presented networks with a 50% word error rate. While in [15], the model presented achieved better performance with a 35.4% word error rate.

6 | EXPERIMENTS AND RESULTS

In this section, we provide some experiments to compare the performance of our proposed model using phonemes as classifiers with the work presented in [37] using characters as classifiers and the authors in [15] use visemes as classifiers. The matrix used for model evaluation includes Word Error Rate (WER) and Character Error Rate (CER) as discussed below.

All the simulations have been implemented with TensorFlow and on a GeForce GTX 1080 Ti GPU with 11 GB memory for the first set of simulations with 90% training to 10% validation and the second set of simulations using GeForce RTX-3070 GPU with 16 GB memory for 70% training to 30% testing.

6.1 | Evaluation criteria

In order to evaluate lip-reading systems, many matrices have been used, such as Word Accuracy Rate (WAR) and Sentence Accuracy Rate (SAR). In addition, word, character, viseme, or phoneme is another category that can be assisted using the Error Rate (ER) matrices, which is shown in Equation 4. The overall distance is calculated by comparing the decoded text to the original text as follows:

$$ER = \frac{S + D + I}{N} \tag{4}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of characters in the ground truth.

As for our model, the matrix used includes phoneme Error Rate (PER), Character Error Rate (CER), and Word Error Rate (WER) as expressed as follows:

$$\mathrm{PER} = \frac{P_{S+}P_D \ + P_I}{V_N} \tag{5}$$

$$\mathrm{CER} = \frac{C_{S+}C_D \ + C_I}{C_N} \tag{6}$$

$$\mathrm{WER} = \frac{W_{S+}W_D \ + W_I}{W_N} \tag{7}$$

6.2 | Experimental results

In this section, the proposed model is evaluated and compared to different state-of-the-art classification schemas. Different ratios of training and testing are used to verify the robustness of the model. The model/phoneme classifier was trained for 2000 epochs. The results for the first simulations are shown in Table 4, and the plots for the loss and the PER for both the training and validation for 2000 epochs are given in Figures 6 and 7. The confusion matrix for classification of ASCII characters is provided in Figure 8.

The phoneme-based lip-reading system achieved an overall WER of 40%, a reduction of 10% compared to the highest result of 50% of the previous state-of-the-art model [37] as presented in Table 5.

Comparing the results of the phoneme-based lip-reading system with those of a viseme-based system that uses the LRS2 dataset with the same ratio of the number of training

TABLE 4 Results of phoneme-based lip-reading system

Epochs	Validation samples	PER (%)	CER (%)	WER (%)
2000	1500	30	32	40

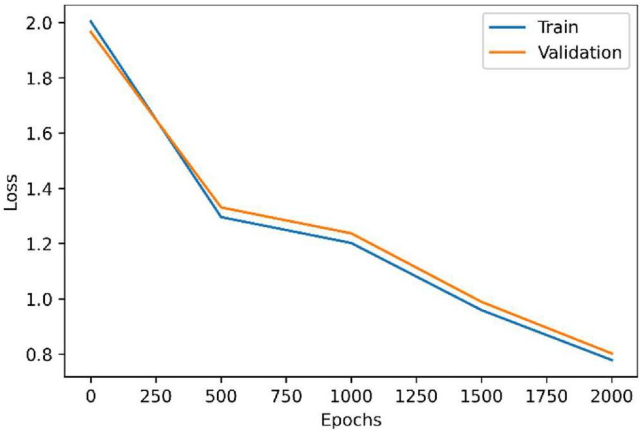

FIGURE 6 Loss curve for training and validation

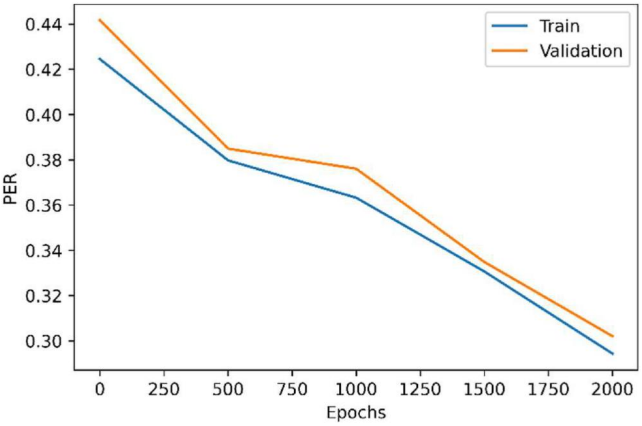

FIGURE 7 PER curve for training and validation

samples to test samples, the observed accuracy of the phoneme-based model was lower than that of the viseme-based. Table 6 shows the PER, VER (Viseme Error Rate,) and WER results. VER is calculated as

$$\mathrm{VER} = \frac{V_{S+}V_D \ + V_I}{V_N} \tag{8}$$

After running more simulations with a different ratio of training to testing till no further convergence was recorded, the achieved results are reported below in Table 7, and the plots for the loss and the PER for both the training and the validation are shown in Figures 9 and 10.

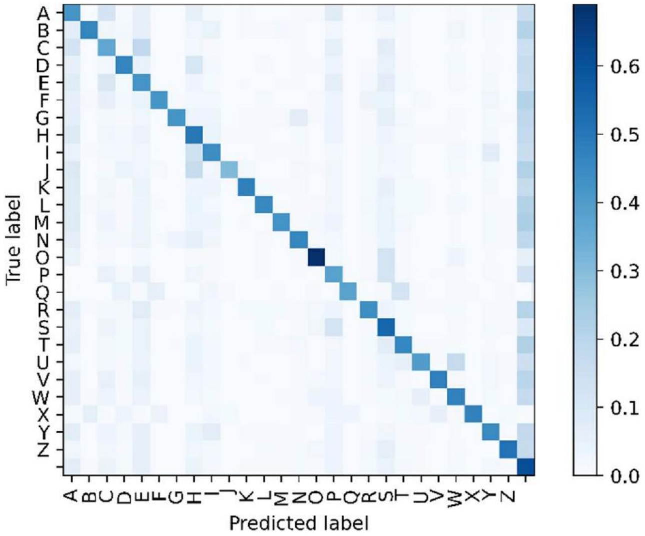

FIGURE 8 Confusion matrices of ASCII characters

TABLE 5 Performance comparison of phoneme and best results of character-based Lip-reading systems

Phoneme-based lip reading		Character-based lip reading [37]	
PER (%)	**WER (%)**	**CER (%)**	**WER (%)**
30	40	34	50

TABLE 6 Performance comparison of phoneme and viseme-based lip-reading systems

Phoneme-based lip-reading		Viseme-based lip-reading [15]	
PER (%)	**WER (%)**	**VER (%)**	**WER (%)**
30	40	5	35

TABLE 7 Results of the phoneme-based lip-reading system

Epochs	Validation samples	PER (%)	CER (%)	WER (%)
4500	4500	16	18	26

6.3 | Gamma correction

The pixel brightness has been altered to provide illumination to image frames in order to test the robustness of the proposed model. Videos consist of images with red, green, and blue pixels, and the intensity of the numerical values ranges from 0 as a minimum to 255 as a maximum. Normalisation is the first step used to map the pixel values ranging from a minimum of 0 to a maximum of 1. The next step is to apply gamma correction according to Equation (9):

$$V_o = A V_I^{\gamma} \tag{9}$$

where A represents a constant equals to 1, V_I represents the matrix of pixels, γ, when given a value of less than <1 makes dark parts lighter, and when given values larger than >1 makes shadowed parts darker. The last step is re-normalisation, where all the pixels are re-normalised to values from 0 to 255.

Tables 8 and 9 show the performance of the phoneme-based lip-reading system under varying illumination ratios compared to the state-of-the-art model [37]. The proposed system has an improved performance with a 10% lower word error rate. It can be seen that the lip-reading system is generally

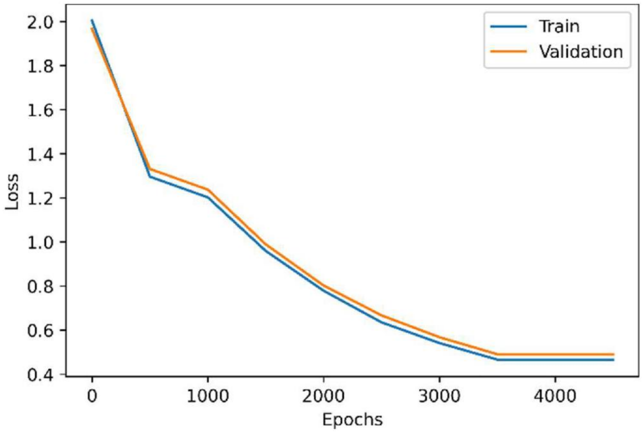

FIGURE 9 Loss curve for training and validation for 4000 epochs

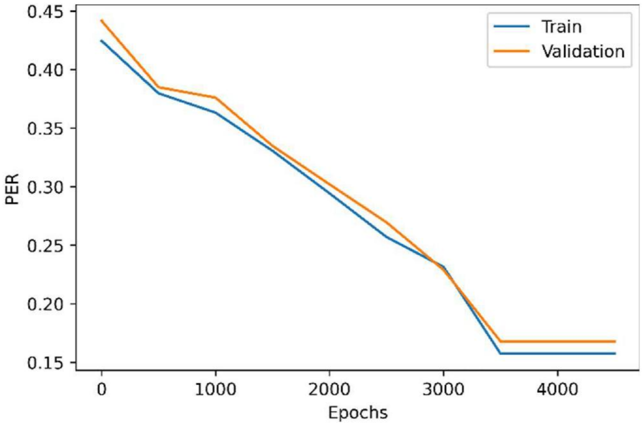

FIGURE 10 PER curve for both training and validation for 4000 epochs

TABLE 8 Performance of the phoneme-based lip-reading system under different gamma ratios

Gamma ratio	PER (%)	CER (%)	WER (%)	WAR (%)	SAR (%)
0.5	34.1	37.2	45.9	54.1	12.8
0.8	31.0	33.2	42.7	57.3	14.6
1	30.0	32.1	40.3	59.8	15.3
1.1	30.5	32.5	40.9	59.1	14.6
1.2	30.9	33.1	41.5	58.5	14.8
1.5	33.3	36.2	45.0	55.0	14.1

TABLE 9 Performance of the phoneme-based lip-reading system compared to the state-of-the-art model

	Phoneme-based lip-reading system			Afouras et al.		
Gamma ratio	PER (%)	WER (%)	SAR (%)	CER (%)	WER (%)	SAR (%)
0.5	34.1	45.9	12.8	35.8	53.9	18.4
0.8	31.0	42.7	14.6	33.9	51.0	20.3
0.9	30.5	41.2	14.9	33.7	50.9	20.6
1	30.0	40.3	15.3	33.7	50.8	20.8
1.1	30.5	40.9	14.6	33.7	50.8	20.2
1.2	30.9	41.5	14.8	34.1	51.4	20.6
1.5	33.3	45.0	14.1	36.2	51.4	20.2

robust to varying levels of illumination, like that reported in [37], and this has been expected given that videos in the BBC LRS2 corpus were recorded in varying lighting conditions.

7 | CONCLUSION

Using phonemes is usually associated with audio signals; however, in this research, audio signals are not presented/provided. A purely phoneme-based lip-reading system from videos using spatial-temporal Convolution Neural Network as the front end and Recurrent Neural Network as the back end has been proposed in this study. The advantage of using phonemes for lip-reading sentences is to overcome the cumulative loss of information caused by the mapping process from phoneme to viseme. Another advantage is having only two variations of dictionaries used in phoneme recognition compared to the large number of phoneme-to-viseme maps; that means that the conversion part in the phoneme system has less complexity, that is, the required computational effort is lower. With the BBC LRS2 benchmark dataset, the proposed model has demonstrated an improved performance by an 18% lower word error rate on average compared with the state-of-the-art lip-reading sentences. The results prove that using phonemes as a classification schema is a promising alternative to other classification schemas.

Future research includes an in-depth investigation on how to improve the performance of the phoneme recognition model and to investigate further how to enhance the language model, particularly in speech impediments, such as in the case of dysarthria speech.

As observed in the results presented in Tables 4 and 7, the phoneme recognition accuracy is 70%, and the word accuracy after the language model decreases to 60% due to the RNN architecture and the fact that in some cases, the current output depends on subsequent sequence factors, not only the previous factors. Furthermore, because the difference in speech speed varies from person to person (the input sequence and the output sequence do not correspond to a one-to-one relationship), the duration of the image frame sequences also differs. Information regarding the beginning and finishing of a word in an image sequence is not obtainable and the fact that RNNs process the input in a sequential behaviour. Because of these reasons, we intend to enhance the language model by substituting the RNN with Attention-Transformer to elevate the accuracy further, primarily when the input consists of distorted phoneme representation.

In our future work, we will further develop our system in the same direction and comprehensively verify the results using more than one dataset. Specifically, we will use the BBC LRS3 and apply a systematic ablation study to investigate the network behaviour further.

DATA AVAILABILITY STATEMENT

BBC Lip Reading Sentences 2 Dataset LRS2 (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html).

ORCID

Randa El-Bialy https://orcid.org/0000-0002-4693-9246
Daqing Chen https://orcid.org/0000-0003-0030-1199
Bo Li https://orcid.org/0000-0002-1415-4444

REFERENCES

1. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimed. 2(3), 141–151 (2000). https://doi.org/10.1109/6046.865479
2. Zhou, Z., et al.: A review of recent advances in visual speech decoding. Image Vis Comput. 32(9), 590–605 (2014). https://doi.org/10.1016/j.imavis.2014.06.004
3. Twaddell, W.F.: On defining the phoneme. Linguistic Society of America 11(1), 5–62 (1935). https://doi.org/10.2307/522070
4. Fisher, C.G.: Confusions among visually perceived consonants. J. Speech Hear. Res. 11(4), 796–804 (1968). https://doi.org/10.1044/jshr.1104.796
5. Fernandez-lopez, A., Sukno, F.M.: Survey on automatic lip-reading in the era of deep learning. Image Vis Comput. 78, 53–72 (2018). https://doi.org/10.1016/j.imavis.2018.07.002
6. Mestri, R., et al.: Analysis of feature extraction and classification models for lip-reading. In: Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI, pp. 911–915 (2019). https://doi.org/10.1109/icoei.2019.8862649
7. Hao, M., et al.: A survey of research on lipreading technology. IEEE Access 8, 204518–204544 (2020). https://doi.org/10.1109/ACCESS.2020.3036865
8. Chung, J.S., et al.: Lip reading sentences in the wild. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3444–3450 (2017). https://doi.org/10.1109/CVPR.2017.367

9. Fenghour, S., et al.: Deep learning-based automated lip-reading: a survey. IEEE Access, 9–121205 (2021). https://doi.org/10.1109/ACCESS.2021.3107946
10. Noda, K., et al.: Audio-visual speech recognition using deep learning. Appl. Intell. 42(4), 722–737 (2015). https://doi.org/10.1007/s10489-014-0629-7
11. Mattos, A.B., Oliveira, D.A.B.: Multi-view mouth renderization for assisting lip-reading. In: Proceedings of the 15th Web for All Conference: Internet of Accessible Things (2018). W4A 2018, (August). https://doi.org/10.1145/3192714.3192824
12. Gabbay, A., et al.: Seeing through noise: visually driven speaker separation and enhancement. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2018-April(December), 3051–3055 (2018). https://doi.org/10.1109/ICASSP.2018.8462527
13. Georgakis, C., Petridis, S., Pantic, M.: Visual-only discrimination between native and non-native speech. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4828–4832 (2014). https://doi.org/10.1109/ICASSP.2014.6854519
14. Chung, J.S., Zisserman, A.: Lip Reading in the Wild (2018). https://doi.org/10.1007/978-3-319-54184-6
15. Fenghour, S., Chen, D., Xiao, P., et al.: Disentangling Homophemes in Lip Reading Using Perplexity Analysis, pp. 1–17 (2020)
16. Fenghour, S., Chen, D., Xiao, P.: Decoder-encoder LSTM for lip reading. In: Proceedings of the Conference: 8th International Conference on Software and Information Engineering. ICSIE (2019)
17. Henning, P., Finn, U., Tønnessen, E.: The status of the concept of phoneme in psycholinguistics the status of the concept of phoneme in psycholinguistics. Speech.Lang.Hear.Res. (2014). https://doi.org/10.1007/s10936-010-9149-8
18. Thangthai, K., Bear, H.L., Harvey, R.: Comparing phonemes and visemes with DNN-based lip-reading', (September) (2018). http://arxiv.org/abs/1805.02924
19. Bear, H.L., et al.: Some observations on computer lip-reading: moving from the dream to the reality. In: Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology, vol. XI, p. 92530G (2014). https://doi.org/10.1117/12.2067464.9253
20. Bear, H.L., Harvey, R.: Alternative visual units for an optimized phoneme-based lip-reading system. Appl. Sci. 9(18), 3870 (2019). https://doi.org/10.3390/app9183870
21. Montgomery, A.A., Jackson, P.L.: Physical characteristics of the lips underlying vowel lip-reading performance. J. Acoust. Soc. Am. 73(6), 2134–2144 (1983). https://doi.org/10.1121/1.389537
22. Howell, D., Theobald, B.-J., Cox, S.J.: Confusion modelling for automated lip-reading using weighted finite-state transducers. In: Proceedings of the International Conference on Auditory-Visual Speech Processing, 197–203 (2013)
23. Noda, K., et al.: Lipreading using convolutional neural network. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 1149–1153. INTERSPEECH (2014). (September). https://doi.org/10.21437/interspeech.2014-293
24. Shillingford, B., et al.: Large-scale visual speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 4135–4139. INTERSPEECH (2019). https://doi.org/10.21437/Interspeech.2019-1669
25. Chiu, C.C., et al.: State-of-the-Art speech recognition with sequence-to-sequence models. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4774–4778 (2018). https://doi.org/10.1109/ICASSP.2018.8462105
26. Fang, A., et al.: Using phoneme representations to build predictive models robust to ASR errors. In: SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 699–708 (2020). https://doi.org/10.1145/3397271.3401050
27. Zeineldeen, M., et al.: A systematic comparison of grapheme-based vs. phoneme-based label units for encoder-decoder-attention models. (2020). http://arxiv.org/abs/2005.09336
28. Zhou, W., et al.: Phoneme Based Neural Transducer for Large Vocabulary Speech Recognition Human Language Technology and Pattern Recognition, pp. 5644–5648. Computer Science Department, AppTek GmbH, 52062 Aachen, Germany, Proc. ICASSP (2021)
29. Fenghour, S., et al.: An effective conversion of visemes to words for high-performance. Sensors 21(November), 7890 (2021). https://doi.org/10.3390/s21237890
30. Fenghour, S., et al.: Lip Reading Sentences Using Deep Learning with Only Visual Cues. IEEE Access (2020). https://doi.org/10.1109/ACCESS.2020.3040906
31. Afouras, T., et al.: Deep Audio-Visual Speech Recognition, pp. 1–13. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018). https://doi.org/10.1109/TPAMI.2018.2889052
32. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lip-reading. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 3652–3656. INTERSPEECH (2017). https://doi.org/10.21437/Interspeech.2017-85
33. Treiman, R., Kessler, B., Bick, S.: Memory and Language Context sensitivity in the spelling of English vowels. Memory and Language 47(3), 448–468 (2002). https://doi.org/10.1016/s0749-596x(02)00010-4
34. Vaswani, A.: Attention is all you need. In: Neural Information Processing Systems (NIPS), pp. 5998–6008 (2017)
35. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Conference Proceedings—EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015). https://doi.org/10.18653/v1/d15-1166
36. Chung, J.S., et al.: Lipreading sentences in the wild. In: Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3444–3450 (2017). https://doi.org/10.1109/CVPR.2017.367
37. Afouras, T., Chung, J.S., Zisserman, A.: Deep Lip Reading: A Comparison of Models and an Online Application Deep Lip Reading: A Comparison of Models and an Online Application, pp. 3514–3518. Interspeech (2018). https://doi.org/10.21437/Interspeech.2018-1943

How to cite this article: El-Bialy, R., et al.: Developing phoneme-based lip-reading sentences system for silent speech recognition. CAAI Trans. Intell. Technol. 1–10 (2022). https://doi.org/10.1049/cit2.12131