IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 15, 2021, accepted August 8, 2021, date of publication August 25, 2021, date of current version September 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107946

Deep Learning-Based Automated Lip-Reading: A Survey

SOUHEIL FENGHOUR[1], DAQING CHEN[1], KUN GUO[2], BO LI[3], AND PERRY XIAO[1]
[1]School of Engineering, London South Bank University, London SE1 0AA, U.K.
[2]Xi'an VANXUM Electronics Technology Company Ltd., Xi'an 710065, China
[3]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China

Corresponding author: Souheil Fenghour (fenghous@lsbu.ac.uk)

This study was supported under a joint scholarship by Chinasoft International Ltd., and London South Bank University.

ABSTRACT A survey on automated lip-reading approaches is presented in this paper with the main focus being on deep learning related methodologies which have proven to be more fruitful for both feature extraction and classification. This survey also provides comparisons of all the different components that make up automated lip-reading systems including the audio-visual databases, feature extraction, classification networks and classification schemas. The main contributions and unique insights of this survey are: 1) A comparison of Convolutional Neural Networks with other neural network architectures for feature extraction; 2) A critical review on the advantages of Attention-Transformers and Temporal Convolutional Networks to Recurrent Neural Networks for classification; 3) A comparison of different classification schemas used for lip-reading including ASCII characters, phonemes and visemes, and 4) A review of the most up-to-date lip-reading systems up until early 2021.

INDEX TERMS Visual speech recognition, lip-reading, deep learning, feature extraction, classification, computer vision, natural language processing.

I. INTRODUCTION

Research in automated lip-reading is a multifaceted discipline. Due to breakthroughs in deep neural networks and the emergence of large-scale databases covering vocabularies with thousands of different words, lip-reading systems have evolved from recognising isolated speech units in the form of digits and letters to decoding entire sentences.

Lip-reading systems typically follow a framework where there is a frontend for feature extraction, a backend for classification and some preprocessing at the start. Stages of automated lip-reading are outlined in Figure 1 and include the following steps:

- Visual Input - Videos of people speaking are sampled into image frames representing speech to be decoded.
- Pre-processing - This is where the region of interest (ROI), i.e., the lips are located and extracted from the raw image data. This involves detecting the face, locating the lips and extracting the lip region from the video image. Some basic transformations are applied to the ROI such as cropping to reduce the number of overall operations needed for training and validation.
- Feature Extraction (Frontend) - This involves extracting effective and relevant features from redundant features and the mapping of high dimensional image data into a lower dimensional representation.
- Classification (Backend) - This involves ascribing speech to facial movements that have been transformed into a lower dimensional feature vector.
- Decoded Speech - Speech is decoded in classes or units and eventually encoded as spoken words or sentences.

FIGURE 1. General framework for automated lip-reading.

Traditional non-deep learning methods with hand-crafted techniques were the first methods used for the automation of lip-reading and such methods include, for instance, Hidden Markov Models (HMMs) [1]–[5]. A variety of different feature extraction techniques have been used including Linear

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

Discriminant Analysis(LDA) [110], Principal Component Analysis(PCA) [6], Direct Cosine Transformations(DCTs) [107] and Active Appearance Models(AAMs) [109].

In recent years, more visual speech recognition systems have moved towards the use of deep learning networks for both feature extraction and classification and in 2011, Ngiam *et al.* [6] first proposed a deep audio-visual speech recognition system based on Restricted Boltzmann Machines(RBMs) [7]. This means that traditional feature extraction techniques like PCA have been superseded by the use of neural networks. Feed-forward networks, Autoencoders [76] and Convolutional Neural Networks(CNNs) are examples of networks that are used in lip-reading frontends. CNNs account for majority of neural network frontends as they are better at learning both spatial and temporal features, and more effective at extracting relevant features.

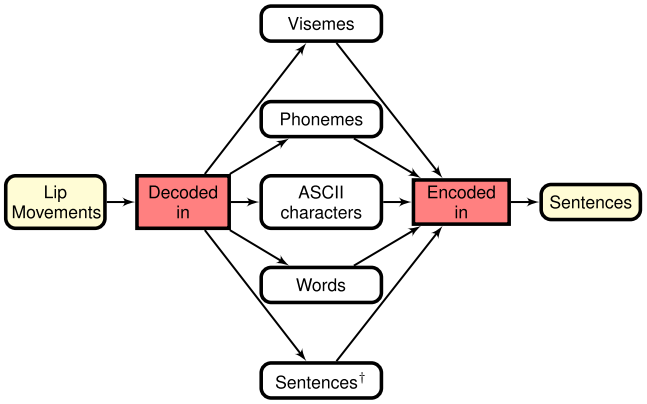

FIGURE 2. Different classification schema.

Lip movements can be interpreted in different ways and there are therefore different classification schemas that have been introduced into the domain such as phonemes [8] or visemes [9] classification. Figure 2 illustrates the various interpretations of lip movements and classification schemas used for lip-reading.

For classification, lip-reading backends predict speech sequential in nature like words or sentences and tend to use sequence processing networks like Recurrent Neural Networks(RNNs). RNNs take the form of either Long-Short Term Memory networks(LSTMs) [99] or Gated Recurrent Units(GRUs) [100]. Recently, alternative classification networks to RNNs such as Attention-based Transformers [103] and Temporal Convolutional Network(TCNs) [105] have been used in lip-reading backends.

There are other surveys on the topic of automated lip-reading with a particular focus on deep learning, for example, [10] and [11]. This survey has some unique insights in that there is a more in-depth comparison of some of the advantages of other alternative frontend networks to CNNs such as feedforward neural networks and autoencoders; and for classification, there is focus on lip-reading architectures with Attention-Transformers and TCNs which have advantages over RNNs; as well as there being a comparison of the different classification schema used in lip-reading. This paper also covers some of the most up-to-date approaches of late 2020 and early 2021.

The rest of the paper is organized as follows: First in Section II, the different audio-visual databases used to train and test lip-reading systems for decoding at the character, word and sentence levels are described; then in Section III, an overview of the different pre-processing aspects that make up lip-reading systems is given. This is followed by a comparison of the different frontend network architectures used for feature extraction in Section IV, a comparison of the different backend classification systems in Section V, and a comparison of the different classification schema in Section VI. In Section VII, a summary is given for performances of the best performing lip-reading systems on some of the most popular audio-visual datasets. Finally in Section VIII, concluding remarks are given along with suggestions for further research and a summary of current challenges faced in the domain of automated lip-reading.

II. DATASETS

As a data-driven process, the design and development of lip-reading systems has been inevitably affected by available data. Ideally, the data should be vocabulary rich, with variations in pose and illumination. Large data corpuses such as LRS2 [38], LRS3-TED [39], LSVSR [8] have been compiled from hours of programmes that have been streamed on the BBC, TED-X and YouTube. These corpuses consist of thousands of videos of people uttering sentences with thousands of different words. These datasets also consist of people speaking at different angles with varying levels of illumination.

Table 1 lists some of the main audio-visual datasets that have been utilized for lip-reading over the last thirty years. The first lip-reading datasets to be constructed were designed for classifying isolated speech segments in the form of digits and letters, with more recent datasets consisting of videos designed to classify longer segments in the form of words. Moreover, the most up-to-date lip-reading datasets consist not only of longer speech segments, but segments in continuous speech as opposed to isolated speech to better model visual speech in real time.

A further development of lip-reading data corpuses in addition to the nature of speech segments themselves is the ability to train lip-reading systems to classify speech from people speaking at various different angles(profile views), as opposed to frontally facing the cameras(frontal views). Additionally datasets such as LRW [40], LRS2 and LRS3 have moved on to gathering videos from multiple speakers as opposed to individual speakers, as one of the challenges facing the success of automated lip-reading systems is the inability to generalize to different people - especially unseen speakers who have not appeared in the training phase.

Other trends in the evolution of audio-visual corpuses include varying resolutions to accommodate for the fact that in real time, a person will often be speaking at varying

TABLE 1. Available audio-visual datasets. I stands for Isolated (one speech segment per recording) and C stands for Continuous recording.

Dataset	Language	Year	I/C	Segment	Speakers	Classes	Utterances	Resolution	Frame rate(fps)	Pose
TULIPS1 [54]	English	1995	Isolated	Digits	12	4	96	100×75	30	Frontal
DAVID [27]	English	1996	Isolated	Words	123			640×480	30	Frontal
M2VTS [43]	English	1997	Continuous	Digits	37	10	2920	286×350	25	Frontal
AVLetters [18]	English	1998	Isolated	Alphabet	10	26	780	376×288	25	Frontal
WAPUSK20 [61]	English	1999	Isolated	Sentences	20	52	2000	640×480	32	Frontal
XM2VTS [62]	English	1999	Continuous	Digits	295	10	1064	720×576	25	Frontal
CAVSR1.0 [23]	Chinese	2000	Isolated	Words	20	78	3120	352×228	25	Frontal
IBMViaVoice [3]	English	2000	Isolated	Sentences	290	10500	24325	704×480	30	Frontal
UNMC-VIER [55]	English	2002	Isolated	Sentences	123	12	2460	708×640	29	0, 90
BANCA [21]	Multiple	2003	Isolated	Digits	208	10	29952	720×576	25	Frontal
AV@CAR [15]	Spanish	2004	Isolated	Alphabet	20	26	800	768×576	25	Frontal
AV@CAR [15]	Spanish	2004	Isolated	Digits	20	10	600	768×576	25	Frontal
AV@CAR [15]	Spanish	2004	Isolated	Sentences	20	250	6000	768×576	25	Frontal
AVICAR [17]	English	2004	Continuous	Alphabet	86	26	59000	720×480	30	4 views
AVICAR [17]	English	2004	Continuous	Digits	86	10		720×480	30	4 views
AVICAR [17]	English	2004	Isolated	Sentences	86	20		720×480	30	4 views
AV-TIMIT [1]	English	2004	Isolated	Sentences	233	510	4660	720×480	30	Frontal
CUAVE [26]	English	2004	Isolated	Digits	36	10	7000	720×480	30	-90, 0, 90
IBMIH [34]	English	2004	Continuous	Digits	79	10	16197	720×480	30	Frontal
UWB-05-HSAVC [56]	Czech	2005	Isolated	Sentences	100	200	20000	720 × 576	25	Frontal
VALID [58]	English	2005	Isolated	Digits	106	10	590	576×720	25	Frontal
GRID [28]	English	2006	Isolated	Phrases	34	34000	34000	720×576	25	Frontal
CMU AVPFV [25]	English	2007	Isolated	Words	10	150	15000	640×480	30	0, 90
AVLetters2 [19]	English	2008	Isolated	Alphabet	5	29	910	1920×1080	50	Frontal
HIT-AVDB-II [32]	Multiple	2008	Isolated	Sentences	30	11	1980	720×576	25	0, 30, 60, 90
IBMSR [35]	English	2008	Continuous	Digits	38	10	1661	368×240	30	-90, 0, 90
IV2 [36]	French	2008	Isolated	Sentences	300	15	4500	780×576	25	0, 90
UWB-07-ICAV [57]	Czech	2008	Isolated	Sentences	50	7550	10000	720 × 576	<50	Frontal
MV-LRS [47]	English	2009	Isolated	Sentences	>1000	14960	74564	160×160	25	0 ~90
CENSREC-1-AV [24]	Japanese	2010	Continuous	Digits	42	10	3234	720×480	30	Frontal
LiLiR [37]	English	2010	Isolated	Sentences	12	200	2400	720×576	25	0, 30, 45, 60, 90
NDUTAVSC [48]	German	2010	Isolated	Digits	66	10	6907	640×480	100	Frontal
NDUTAVSC [48]	German	2010	Isolated	Words		6907	6907	640×480	100	Frontal
NDUTAVSC [48]	German	2010	Isolated	Sentences				640×480	100	Frontal
OuluVS2 [50]	English	2010	Continuous	Digits	53	10	159	1920×1080	30	0, 30, 45, 60, 90
OuluVS2 [50]	English	2010	Isolated	Phrases	53	10	1590	1920×1080	30	0, 30, 45, 60, 90
OuluVS2 [50]	English	2010	Isolated	Sentences	53	540	2120	1920×1080	30	0, 30, 45, 60, 90
VIDTIMIT [59]	English	2010	Isolated	Sentences	34	346	430	512×384	25	Frontal
BL [22]	French	2011	Isolated	Sentences	17	238	4046	640×480	30	0, 90
LTS5 [42]	French	2011	Isolated	Digits	20	10	180	1920x1080	25	0, 30, 60, 90
MIRACL-VC [44]	English	2011	Isolated	Words	15	10	1500	640×480	15	Frontal
TCD-TIMIT [53]	English	2011	Isolated	Sentences	20	62	5954	1920×1080	30	0, 30
AGH AV [12]	Polish	2012	Isolated	Digits	20	10	N/A	1920×1080	50	Frontal
MIRACL-VC [44]	English	2012	Isolated	Phrases	15	10	1500	640×480	15	Frontal
AVAS [16]	Arabic	2013	Isolated	Digits	50	10	13850	640×480	30	-90, -45, 0, 45, 90
AVAS [16]	Arabic	2013	Isolated	Words		24		640×480	30	-90, -45, 0, 45, 90
AVAS [16]	Arabic	2013	Isolated	Phrases		13		640×480	30	-90, -45, 0, 45, 90
AusTalk [13]	English	2014	Isolated	Digits	1000	10	24000	640×480	-	Frontal
AusTalk [13]	English	2014	Isolated	Words	1000	996	996000	640×480	-	Frontal
AusTalk [13]	English	2014	Isolated	Sentences	1000	59	59000	640×480	-	Frontal
LSVSR [8]	English	2014	Isolated	Sentences	>1000	127055	2934899	128×128	23-30	-30 ~30
IBM AV-ASR [33]	English	2015	Isolated	Sentences	262	10400	N/A	704×480	30	Frontal
MODALITY [46]	English	2015	Isolated	Words	35	182	231	1920×1080	100	Frontal
OuluVS [49]	English	2015	Isolated	Sentences	20	10	1000	720×576	25	Frontal
QuLips [51]	English	2015	Isolated	Digits	2	10	3600	720×576	25	-90 ~90
RM-3000 [52]	English	2015	Isolated	Sentences	1	1000	3000	360×640	60	Frontal
HAVRUS [31]	Russian	2016	Isolated	Sentences	20	1530	4000	640×460	200	Frontal
LRW [40]	English	2016	Continuous	Words	>1000	500	400000	256×256	25	-30 ~30
LRS2 [38]	English	2017	Isolated	Sentences	>1000	17428	118116	160×160	25	-30 ~30
MOBIO [45]	English	2017	Isolated	Sentences	150	N/A	N/A	640×480	16	Frontal
VLRF [60]	Spanish	2017	Isolated	Sentences	24	1374	10200	1280×720	50	Frontal
AV Digits [14]	English	2018	Isolated	Digits	53	10	795	1280×780	30	0, 45, 90
AV Digits [14]	English	2018	Isolated	Phrases	39	10	5850	1280×780	30	0, 45, 90
GRID-Lombard [29]	English	2018	Isolated	Phrases	54	5400	5400	720×480(face), 864×480(side)	24	0, 90
LRS3 [39]	English	2018	Isolated	Sentences	>1000	70000	165000	224×224	25	-90 ~90
LRW-1000 [41]	English	2018	Continuous	Words	>2000	1000	718018	Distributed	25	-90 ~90
AVSD [20]	Arabic	2019	Isolated	Phrases	22	10	1100	1920×1080	30	Frontal
VR Digits [63]	English	2020	Continuous	Digits	6	10	6000	1920×1080	25	Frontal
NSTDB [64]	Chinese	2020	Continuous	Words	N/A	349	N/A	64×64	25	-90 ~90

distances from a video camera. There have also been varying frame rates to accommodate for videos that are sampled at different frequencies as well having to contend with the possibility of there not being enough temporal information available due to the nature of videos having a low sampling frequency. The majority of corpuses uses the English language due to English being the World's lingua franca, though there are datasets that utilize other languages.

A. LETTER AND DIGIT RECOGNITION

Because research in automated lip-reading started with simplest cases possible before gradually evolving to be suited

to lip-reading natural spoken language in real time, the first databases that were available for lip-reading were designed for the task of recognizing English letters and digits.

The AVLetters [18] dataset consists of 10 speakers (5 males and 5 females) uttering isolated letters from A to Z. Each letter was repeated three times by the speaker, and videos were recorded at a rate of 25 frames per second(fps) at an audio sampling rate of 22.5 kHz. A higher definition edition of the AVLetters database named AVLetters2 [19] was later compiled; and it includes 5 speakers uttering 26 isolated letters seven times with videos sampled at 50 fps, with an audio sampling rate of 48 kHz.

The AVICAR [17] dataset was recorded in a moving car with four cameras deployed on the dashboard for recording videos. The dataset consists of 100 speakers (50 males and 50 females) with 86 of them available for downloading. Each speaker was asked to first speak isolated digits and then letters twice, followed by 20 phone numbers with 10 digits each. Videos have a visual frame rate of 30 fps and an audio sampling rate of 16 kHz.

Tulips [54] which was released in 1995 is one of the oldest databases constructed for digit recognition. It consists of 96 grayscale image sequences pertaining to 12 speakers (9 males and 3 females) each uttering the first four English digits twice. Videos were sampled at 30 fps with resolution 100×75 pixels and the images contain only the mouth region of the speakers.

The M2VTS database [43] contains videos of 37 people (25 men and 12 women) uttering consecutive French numerals from 0-9, which were repeated five times by each person. The XM2VTSDB database [62] is an extension of the M2VTS database, and was constructed by getting 295 people to utter digits 0-9 in different orders. The VALID [58] database was designed to test a lip-reading system's robustness to light and noise conditions which is why the videos contain illumination, background and noise variations. Altogether, it contains 530 videos with 106 speakers speaking in five different environments.

AVDigits [14] is one of the largest datasets available for digit classification. It contains videos recorded with normal, whispered and silent speech and in it; participants read out 10 digits, from 0 to 9 in a random order five times in the three different modes of speech. They spoke at normal volume for the mode of normal speech, whispered for the whispering mode and remained silent in silent speech mode. 53 participants were recorded in total.

The CUAVE [26] (Clemson University Audio-Visual Experiments) database includes speaker movement and simultaneous speech from multiple speakers. It is split into two major sections: the first consists of individual speakers and the second consists of pairs of speakers. For the first section, 36 speakers (17 males and 19 females) were recorded with each speaker uttering 50 isolated digits while facing the front; another 30 isolated digits while moving the head and after that, the speaker was recorded from both profile views while speaking 20 isolated digits. Each individual then uttered 60 connected digits while facing the camera again. Videos were recorded at 30 fps with an audio sampling rate of 16 kHz.

Other corpuses constructed for digit recognition in speech recognition include AV@CAR [15] for Spanish digits, CENSREC-1-AV [24] for Japanese, NDUTAVSC [48] for German; LTS5 [42] databases for French, AGH AV [12] for Polish as well as other English datasets like IBMIH [34], IBMSR [35] and QuLips [51].

B. WORD AND SENTENCE RECOGNITION

The focus of compiling datasets for letter and digit recognition initially was not motivated solely by starting with simplest cases possible, but also due to the simplicity in the gathering of such data. Later, researchers focused more on the task of predicting words, phrases and sentences in continuous speech whereby they had to overcome the problem of trying to identify different words that look or sound identical when spoken.

The MIRACL-VC1 [44] database was released in 2014. It consists of videos from 15 participants who each uttered one of 10 possible words ten times, resulting in the availability of 1500 word videos. Videos were recorded using an RGBD camera with resolution 640×480 pixels and a frame rate of 15 fps. The videos were sampled into image frames with the images being divided into colour pictures and depth pictures - the latter of which contained more depth information.

Other isolated word datasets for the English language include MODALITY [46], AusTalk [13], CMU AVPFV [25] and DAVID [27]. Corpuses for other languages include AVAS [16] for Arabic, CAVSR1.0 [23] for Chinese and NDUTAVSC [48] for German.

Meanwhile, possibly the one of largest English word datasets we have available to us today, LRW [65] contains 1000 utterances of 500 different words, spoken by over 1000 different speakers. Videos were extracted from a number of BBC television programmes streamed between 2010 and 2016, and they are 1.16s long with a frame rate of 50 fps without any audio.

LRW-1000 [41] is possibly one of the largest continuous audio-visual datasets for words altogether consisting of over 700,000 samples of 1000 Chinese words spoken by over 2000 different speakers from Chinese CCTV programs. This dataset is unique in that it consists of videos with varying resolutions which makes it useful for the natural variability of people speaking in real-time where you will either have people speaking at varying distances from a video camera or videos that have been recorded with varying spatial dimensions.

The XM2VTSDB [62] corpus which consists of 295 speakers uttering digits, also consists of videos with the 295 speakers pronouncing the sentence "Joe too parents green shoe bench out". This makes it one of the oldest sentence-based corpuses. The MIRACL-VC1 [44] dataset in addition to having compiled word video data, also consists of sentence

videos whereby each of the 10 speakers uttered one of ten phrases ten times to generate 1500 phrase videos.

IBMViaVoice is one of the largest datasets available for lip-reading sentences and it contains videos with 290 speakers speaking a total of 24325 sentences with different 10500 words being spoken. It is however unavailable to the public.

The OuluVS1 [49] database consists of 10 phrases spoken by 20 speakers(17 males and 3 females), with each utterance repeated by the speaker up to nine times. Videos were recorded at 25 fps with an audio sampling rate of 48kHz. The OuluVS2 [50] database is an extension of OuluVS1 which also contains videos of these 10 phrases but spoken by with 52 different speakers.

The GRID [28] corpus consists of 34 speakers(18 males and 16 females) who each utter 1000 sentences [28] that follow a standard pattern of verbs, colours, prepositions, alphabet, digits, and adverbs [28]. ''Set white with p two soon'' is an example of one spoken sentence and each video has a duration of 3 seconds with a sampling rate of 25 fps and audio 25kHz.

The GRID-Lombard [29] database is an extension of the GRID corpus and consists of 54 speakers(30 females and 24 males) who altogether pronounce 5400 sentences that follow the GRID convention and take the form of ''<verb>, <colour>, <preposition>, <letter>, <number>, <adverb>'' using combinations that do not appear in the GRID corpus. It should be noted that the emphasis of this corpus is to not only include profile views of people speaking in addition to frontal views but to also provide videos of people speaking according to Lombard speech so that the Lombard effect can be modelled. The Lombard effect is the spontaneous habit of a speaker to increase their vocal effort when speaking in loud noise to enhance the audibility of their voice [30].

The TIMIT corpus is a dataset with audio recordings of 630 speakers each speaking 10 different sentences to give a total of 6300 sentences [66]. Several datasets with people uttering sentences following the TIMIT structure have been constructed.

The AV-TIMIT [1] database was constructed for performing speaker-independent audio-visual speech recognition and the corpus contains videos of 233 speakers (117 males and 106 females) uttering TIMIT sentences [66]. Each speaker was asked to utter 20 sentences, and each sentence was spoken by at least 9 different speakers with one sentence that was uttered by all the speakers. Videos were recorded at 30 fps with a resolution of 720 × 480 pixels and an audio sampling rate of 16 kHz.

Similarly, the Vid-TIMIT [59] database is comprised of videos of 43 speakers (19 females and 24 males), each pronouncing 10 different TIMIT sentences. The videos were recorded at 25 fps with resolution 512 × 384 pixels and an audio-sampling rate of 32kHz. Meanwhile, the TCD-TIMIT [53] database consists of videos of resolution 1920 × 1080 pixels from 62 female speakers of whom 3 are professional lip readers and the other 59 are volunteers. The three professionals say 377 sentences each while the remaining speakers speak 98 sentences each.

In recent years, more challenging datasets consisting of spoken sentences that are more random and less structured have been constructed which consist of thousands of sentences spoken by limitless people, with extensive vocabularies covering thousands of different possible words so that lip-reading systems can be generalised to natural spoken language. The LRS2 [65] dataset is a sentence-based dataset of videos without audio which was compiled by extracting videos from BBC television programmes much like the LRW corpus. Altogether the corpus covers 17,428 different words with a total of 118,116 samples.

MV-LRS [47] is also a sentence-based dataset constructed from videos from BBC programs with a total of 74,564 samples covering 14,960 words. However, unlike the LRS2 corpus which only includes frontal shots, MV-LRS includes both profile and frontal shots. The LRS3-TED [67] dataset is another sentence-based dataset compiled in a similar fashion by extracting videos from Ted-X videos where 150,000 sentences were extracted from TED programs. LSVSR [68] was built using YouTube videos with 140,000 hours of audio, approximately 3,000,000 speech utterances and over 127,000 words making it the largest database to date.

Lip-reading datasets with people pronouncing sentences in other languages have also been created too. Examples include AV@CAR [15] and VLRF for Spanish, AVAS [16] and AVSD [20] for Arabic, BL [22] and IV2 [36] for French, UWB-05-HSAVC [56], and UWB-07-ICAV [57] for Czech, the German NDUTAVSC [48] dataset, the Russian HAVRUS [31] corpus and the HIT-AVDB-II [32] database that covers Chinese and English.

C. MULTIVIEW DATABASES

In an ideal situation, an automated lip-reading would only need videos of people speaking from frontal poses. However, in practice it is impossible to always guarantee that the input images will be exclusively from frontal shots.

Another challenge with pose is when a video with a talking person consists of that very person speaking at different angles. When there is a static camera, a speaker may rotate their face while speaking which results in the data that is present consisting of a person speaking at multiple angles in the very same video. Some datasets provide image data recorded at various angles whilst a speaker is speaking, though this is not always the case.

Many researchers argue that the frontal shots are not necessarily the best angles to use for lip-reading. One reason for this is that a slight angle deviation can be beneficial because lip-protrusion and the rounding of the lip can be better observed [11], [69].

III. PREPROCESSING

One of the stages of automated lip-reading is to extract the region of interest and in the case of automated lip-reading, the ROI that needs to be extracted is the person's lips. The lip

movements will be given a speech class label according to the hierarchy of speech data explained in Section VI.

There are different feature representation methods that can be used to represent lip movements and they can typically be divided into four categories as summarized by Dupont and Luettin [70]: geometric-based, image-based, model-based and motion-based. A more detailed comparison of feature representation can be found in the following works [70], [71].

The overwhelming majority of deep learning classification methods use image-based feature representation and the input will either be an image with channels of red, green and blue pixel intensities or an input with grayscale images. A general advantage of being able to use raw pixel data as a neural network input is that there is less pre-processing involved as there is no need to device hand-crafted models for extracting facial contours or the representations of lip motion.

For a recorded video of a person speaking, an automated lip-reading system will first need to sample the video into image frames. Once the video has been sampled, the face must be detected as part of a face localization step which involves facial landmarks needing to be located in order to extract just the speaking person's lips as the ROI and feature input to the visual frontend. Figure 3 outlines the process of extracting the ROI of an individual speaking in a video, while Figure 4 shows an example of an image frame and its corresponding ROI.

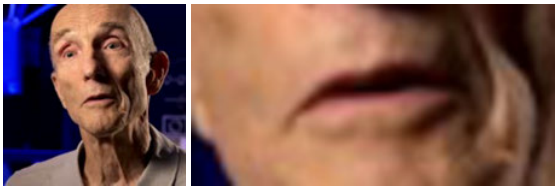

FIGURE 3. A person's face on the left with the extracted ROI shown on the right.

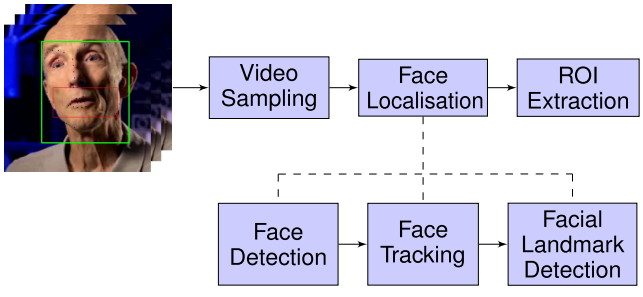

FIGURE 4. Procedure for video processing.

A variety of face localization methods can be used for extracting facial landmarks from people's faces and such approaches include Naive Bayes classifiers [72], neural networks [73], HMMs [74] and Principal Component Analysis [75] to name a few. A more detailed review of face localization procedures can be found in [74], though they all typically use the standard iBug landmark convention where 68 landmarks are detected for the face. The procedure for locating facial landmarks and to extract the ROI is shown in Figure 5.

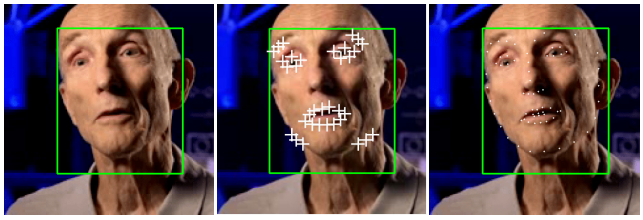

FIGURE 5. Facial landmark extraction stages which include face detection(left), face tracking(middle) and facial landmark detection(right).

For the first deep learning-based lip reading systems, the ROI extraction was often performed as part of preprocessing, but modern end-to-end lip reading systems now perform ROI extraction during the feature extraction stages whereby a frontend will have been trained to locate the ROI and this means that video frames do not need cropping beforehand [102], [104].

After locating and extracting the ROI, a series of pre-processing steps will typically be applied to the image and this is done to not only improve the efficiency of training and validation by reducing the number of overall operations but also to limit variation as much as possible. Pre-processing will often consists of processes such as grayscale conversion, z-score normalization and some augmentation techniques; though augmentation is implemented during the training phase.

Images naturally consist of three pixel channels in the red-green-blue(RGB) format with red, green and blue pixel components. The challenge with images having multiple colour channels is that there will be huge volumes of data to work with, making the process computationally intensive. So as a result, lip-reading systems will often consist of a grayscale conversion stage where RGB pixels are converted to a grayscale format beforehand.

Another pre-processing step is the Normalization process. Normalizing helps to ensure consistency of scale when processing images, which can improve a model's ability to learn if the scales for different features are very different. Z-score normalization is the simplest of such techniques where a correction is applied to all of the pixels by subtracting from every pixel x the mean pixel value $\bar{x}$ and then dividing by the standard deviation σ to give a corrected pixel value x' with zero-mean and unit-variance according to Eq. 1.

$$x' = \frac{x - \bar{x}}{\sigma} \tag{1}$$

In summary, the training of a good classification model for speech recognition requires a lot of data and the lack of the labelled training data leads to poor generalization. A greater availability of training data will invariably lead to a better classification model. However, when there is an

insufficient supply of data available to begin with, augmentation can be a useful strategy which is where existing training data is extended by adding modified or augmented samples. New training samples can be created by applying various transformations to existing labelled samples. Examples of image-based augmentation techniques include rotation, scaling, flipping, cropping, spatial or temporal pixel translation and even the addition of Gaussian noise.

IV. FEATURE EXTRACTION

Feature extraction for visual speech recognition has two main purposes. The first is to separate redundant features in the images from relevant features and the second is to convert images from high-dimensional space into low-dimensional space. A variety of techniques such as Active Appearance Models, Active Shape Models, Discrete Cosine Transformation, Linear Discriminant Analysis, Principal Component Analysis and Locality Discriminant Graphs have been deployed for feature extraction in lip-reading and more detailed information about such approaches can be found in Zhou's work [71]. Non-deep learning methods of feature extraction will not be discussed in this Section. For most of the up-to-date state-of-the-art lip-reading systems, deep learning methods are preferred to traditional methods because feature extraction can be automated.

Convolutional Neural Networks are one family of neural networks that have been deployed for feature extraction in neural network architectures for automated lip-reading. They are a supervised learning method and they account for majority of networks used for feature extraction. The other family of architectures used for feature extraction include Autoencoders, Restricted Boltzmann Machines and Deep Belief Networks which are all unsupervised methods mainly used in dimensionality reduction tasks.

A. FEED-FORWARD NEURAL NETWORKS

A feed-forward neural network is the most basic neural network that can be used for feature extraction. Wand *et al.* used a feed-forward network as part of a frontend for three of their approaches where 51 different possible variants of words from the GRID corpus were decoded with an LSTM configuration used in the backend. A 40×40 pixel window containing the lips was extracted from each video frame before being converted to grayscale and flattened into a 1D vector. This was performed for every frame that made up the video and so videos were inputted into the frontend in the form of 2D matrices.

Feed-forward neural networks are limited in comparison to other architectures that can be used for feature extraction including Autoencoders and CNNs because image frame pixels from videos have to be stacked together. This means that feed-forward neural networks simply compress image data without being able to learn the spatial and temporal features needed for processing sequential inputs.

B. AUTOENCODERS AND RBMs

An Autoencoder is a network used for learning compressed distributions of data. Autoencoders consist of an encoder and decoder. The encoder converts data in higher-dimensional space to lower-dimensional space, while the decoder transforms the lower-dimensional data into higher-dimensional data. For input data x, the autoencoder tries learning identity relationship $x_{out} = x$ by tuning the network weights and biases when the network is being trained. The loss function is simply the difference between x_{out} and x which the network tries to minimise. The operations performed by the encoder and a decoder are given in Eqs. 2 to 5 respectively. W is the encoder weight matrix, b is the encoder bias matrix, W^T is the decoder weight matrix, and b' is encoder bias matrix [76].

$$Encoder(x) = Wx + b \tag{2}$$
$$Decoder(x) = W^T x + b' \tag{3}$$
$$min(f_{loss} : W^T(Wx + b) + b', x) \tag{4}$$
$$C_{AE} = W_{AE} I + b \tag{5}$$

The Decoder section of the Autoencoder is only used for training and discarded for validation as it the compressed represented learned by the Encoder that is used for feature extraction in lip-reading [76].

Real Boltzmann Machines have an identical structure to Autoencoders, but they differ in that they use stochastic units with a particular distribution(usually Binary of Gaussian) instead of deterministic distribution. The learning procedure consists of several steps of Gibbs sampling where the weights are adjusted to minimize the loss function [76].

Petridis *et al.* proposed lip-reading systems in a number of works that use bottleneck RBMs to do the feature extraction for lip-reading sentences. Their work in [77] decoded phrases from the OuluVS2 using an LSTM backend with two visual input streams. The first input stream uses inputs of 2D image frames converted into grayscale, while the second stream uses the difference between two consecutive frames as the input where feature extraction is performed for that input. For the outputs of both bottlenecks, the first and second derivatives are processed and added to the bottleneck outputs. Each overall output stream is then is fed into an LSTM layer with both LSTM outputs then concatenated and passed into a Bidirectional LSTM with their information combined. The output layer is a softmax layer that performs the classification.

Petridis *et al.*'s architecture in [78] is similar to that of [77] except the second input stream takes audio as an input as opposed to taking in the differential of two consecutive images frames, as well using bidirectional LSTMs at the end of each input instead of unidirectional LSTMs. Petridis *et al.* [79] presented a third system for tackling multi-view lip-reading for sentence prediction. There are three architecturally identical streams to extract features from three images captured from different angles. The outputs are concatenated and passed into a Bidirectional LSTM and a softmax layer that perform classification in an identical manner to [77] and [78]. Meanwhile, Petridis *et al.*'s fourth proposed

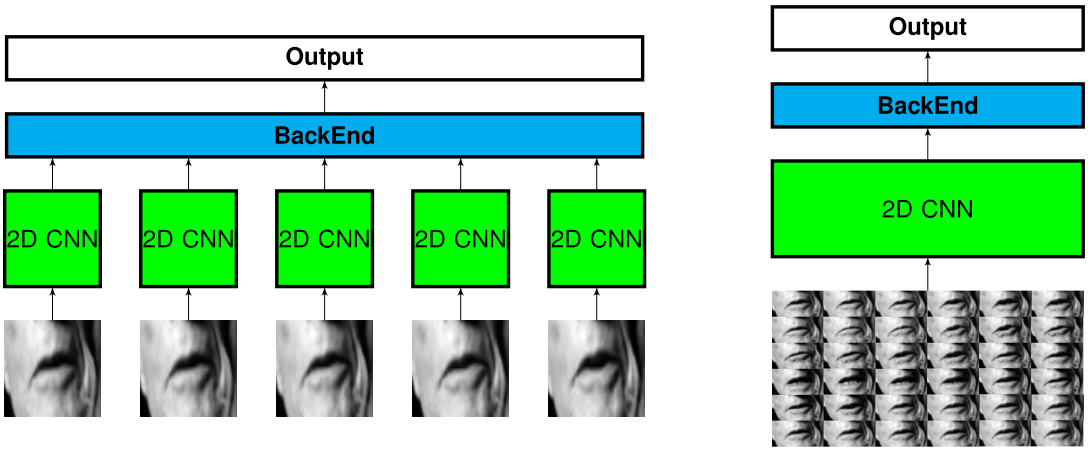

FIGURE 6. CNN diagrams with 2D kernel CNN shown on the left and Concatenated Image Frame CNN on the right.

architecture [14] is similar to [78] except that the system uses only visual inputs with no audio for assistance.

Autoencoders and RBMs do have advantages over CNNs; one is that they are unsupervised learning techniques and can map data from higher dimensions to lower dimensions in isolation without the need for any labelled classification. They also have simpler topologies to tune and are quicker and more compact for backpropagation [80].

Autoencoders and RBMs do have limitations in their feature extraction capabilities. Whilst Autoencoder or RBMs try to capture as much information as possible, they can be inefficient if information that is most relevant the classifier makes up only a small part of the input and so an autoencoder or RBM may lose a lot of it. CNNs are better at separating relevant information from redundant information [80].

C. 2D CNNs

It is common to have a series of 2D CNN kernels whereby feature extraction is performed for each individual image frame(Figure 6). A CNN will extract features using architectural layers for convolution, pooling and normalization; and for a 2D CNN, the convolution stage involves convolving an input y with a weight ω of pixel width w and pixel width h over the different channels. For the expression shown in Eq. 6, C represents the different channels for the image. There will be three channels for RGB pixels and 1 channel for grayscale pixels and the convolution may consist of an arbitrary bias b.

$$(y \otimes \omega)_{wh} = \sum_{c=1}^{C} \sum_{w'=1}^{k_w} \sum_{h'=1}^{k_h} y_{cw'h'} \omega_{c,w'+w,h'+h} \tag{6}$$

Noda *et al.* [81] were among the first group to use CNNs for lip-reading in a task of extracting visual feature sequences for 6 people speaking 300 Japanese words whereby the output formed the input of a Gaussian Mixture Model-Hidden Markov Model(GMM-HMM) used for classification. Their results demonstrated that the visual features acquired by CNNs were significantly better than those acquired using traditional methods like PCAs. They later proposed a lip-reading system that incorporated audio as an input for assistance to create an audio-visual speech recognition system.

Garg *et al.* [82] were the first to use Concatenated Frame Images(CFIs) as shown in Figure 6 where a 2D CNN with the VGG topology was used as their frontend(the structure of a VGG CNN is shown in Figure 7). Image frames were intertwined within one giant image frame to form the input of the LSTM that was utilised for classification where they effectively transformed the temporal information per data-point into spatial information. Their model was trained and tested on videos from MIRACL-VC1 dataset and their best performance was achieved when freezing the VGG parameters and then training the LSTM, rather than training both the backend and frontend simultaneously.

Li *et al.* [83] acknowledged that dynamic features are a better representation of moving lips than static features, so they represented lip movements not in the form of static images, but in the form of dynamic images. Dynamic images are obtained by calculating the first-order regression coefficients of every three consecutive image frames. The extracted features formed the input of an HMM which classified words from the Japanese word-based ATR dataset that consisted of 2620 words for training and 216 words for testing.

Chung and Zisserman proposed SyncNet [84], a CNN consisting of 5 convolution layers and 5 fully-connected layers. Grayscale images are the input, with a feature vector as the frontend output. The output of each CNN kernel is then concatenated and inputted into a single LSTM and their overall model performs the classification of phrases from the OuluVS dataset. The LSTM processes the feature vector as a temporal sequence and with a Softmax layer, a class is predicted. They repeat the same task using almost the same architecture except with a VGG-M topology for the CNN kernels that was already pre-trained in ImageNet with its weights being frozen for training as opposed to the SyncNet. An accuracy rate for validation of the initial SyncNet model of 92.8% was recorded in comparison to a validation accuracy rate of just 25.4% and the main reason for the former model

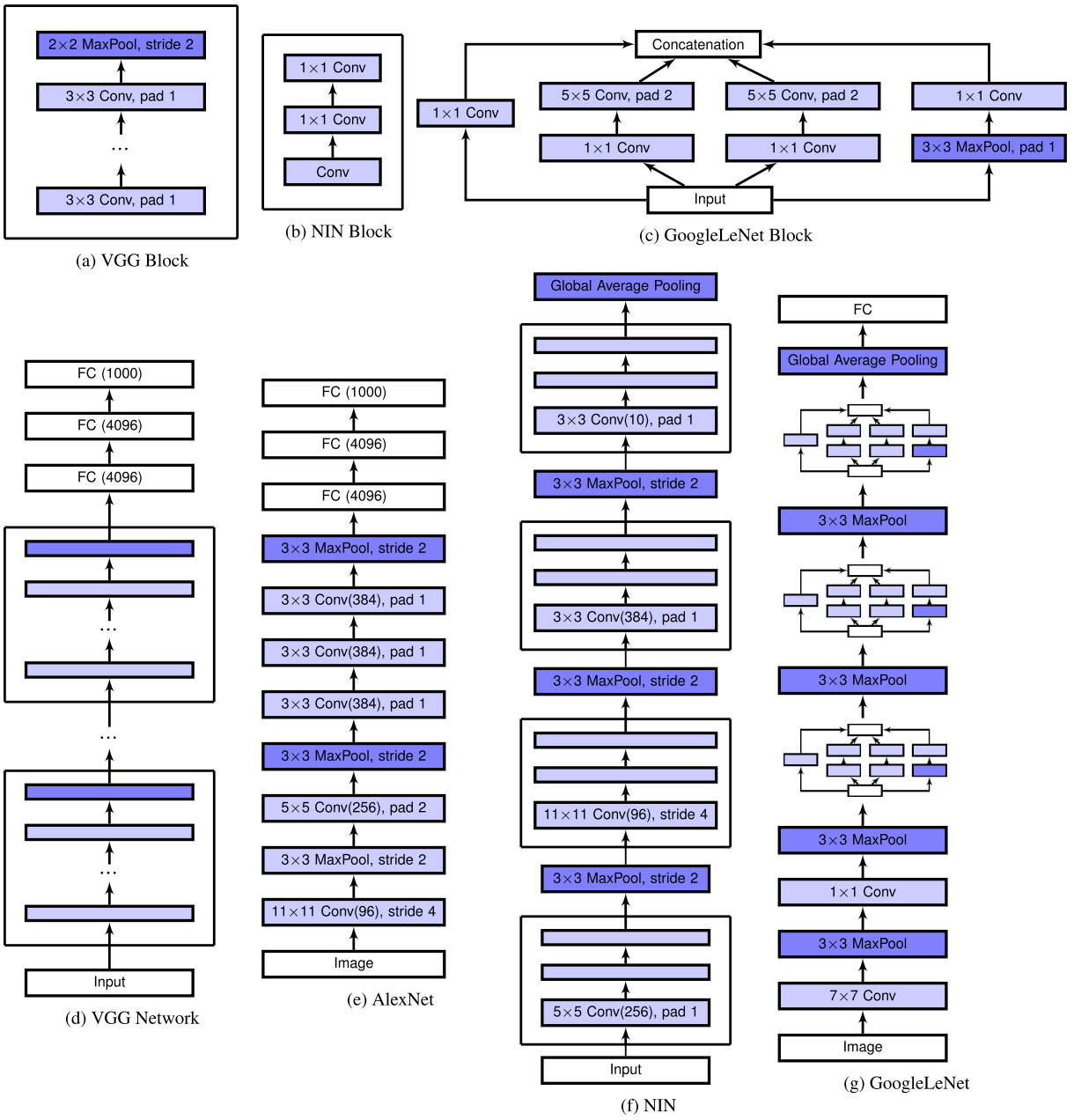

FIGURE 7. CNN architectures corresponding to VGG Network, AlexNet, NIN and GoogleLeNet.

performing significantly better was that the SyncNet kernels were trained directly on the lip-reading data as opposed to the VGG-M kernels which were not.

Lee *et al.* [85] devised a multi-view lip-reading system and experimented with three scenarios: single-view, cross-view, and multiple-view. Their system consisted of a frontend with two layers of CNN kernels and a backend with a two-layer LSTM, that was trained and validated on the OuluVS2 corpus. The corpus contains images of lip movements that were divided into 5 groups including frontal, 30°, 45°, 60° and profile. For the single-view scenario, training and testing was performed for each group separately. For the cross-view scenario, data from all angles were trained together and validated on each one of the 5 groups separately. For the multiple-view scenario, images of each of the poses where merged into one frame and the network was trained and tested on the merged data.

Lu and Li [63] introduced a hybrid neural network architecture composed of a CNN and an attention-based LSTM for lip-reading. Firstly, they extracted key frames (numbers zero to nine in English for three males and three females) from an independent database they created. They then implemented the VGG network to extract lip image features and found that the image feature extraction results were fault-tolerant and effective. Lastly, they used two fully connected layers and a SoftMax layer to classify the test samples. The approach they

proposed was superior to traditional lip-reading recognition methods. Specifically, in the test dataset, the accuracy of the proposed model was 88.2%, which was 3.3% higher than the general CNN-RNN.

Saitoh *et al.* [86] devised a system that takes CFIs as an input, where lip images are merged into one single frame like the approach of Garg *et al.* [82]. They used three different CNN models with three different topologies to extract features from CFIs that include the Network in Network(NIN) [87], AlexNet, and GoogLeNet. The NIN uses 4 four Mlpconv blocks with a max pooling layer between each block, and a softmax layer at the end of the network; AlexNet uses five convolution layers and three fully connected layers; while GoogLeNet is a 22-layer neural network that uses a sparse connection architecture to avoid computational bottlenecks(diagrams of the overall networks are shown in Figure 7). On a classification task of decoding digits and phrases from the OuluVS2 corpus, the system that used the GoogLeNet CNN attained the best performance result.

Chung and Zisserman [40] used VGG-based CNNs for feature extraction when lip-reading words in continuous speech from the LRW dataset. They proposed two different structures including Early Fusion(EF) and Multiple Tower (MT), which both concatenate the outputs of the different CNN kernel streams at different stages. The EF model involves applying 2D CNN kernels to every grayscale ROI and concatenating the outputs before applying convolution layers and pooling layers. Whereas the MT model uses extracted ROIs with RGB pixels and applies one stage of convolution and pooling to the outputs of every stream individually before concatenating the streams. Performance results indicated that the MT model performed the best.

Mesbah *et al.* [88] proposed a CNN structure (HCNN) based on Hahn moments and Hahn moments are effective in the sense that they can be used to extract the most useful information in image frames to reduce redundancy. Hahn moments are applied to the frames at the input to extract moments and input them to the CNN-based frontend and this helps to reduce the dimensionality of video images so that images can be represented with fewer dimensions. The frontend which takes moment matrices as the input, consists of three convolutional layers and two fully connected layers. A softmax layer was used for backend and the lip-reading system performed word classification on the LRW dataset whereby each word was encoded as an individual class.

Zhang *et al.* [89] proposed a visual speech recognition system called LipCH-Net, for recognizing Chinese sentences from the challenging CCTV dataset in two stages. The first stage involved the conversion of image sequences as an input, to Pinyin as an output; while for the second stage, the decoded Pinyin was converted to Hanzi. The inputs take the form of fixed-size grayscale images where CNN kernels following the VGG-M topology extract the features which are then followed by a 14-layer ResNet (each block consists of two convolutional layers, plus batch normalization and rectified linear units). The backend consists of two LSTMs with a CTC. The architecture for performing the second stage of Pinyin-to-Hanzi conversion uses an attention-based GRU.

Lu *et al.* [90] used a CNN and RNN to construct a speech training system for hearing impaired individuals and dysphonic people. First and foremost, a speech training database was built which stored mouth shapes of normal people and the corresponding gesture language vocabulary. The overall system combines the MobileNet and the LSTM networks to performs lip-reading and then, the system finds the correct lip shape matching the sign language vocabulary from the speech training database and compares the result with the lip shape of the hearing impaired. Finally, the system will compare and analyze the lips size, opening angle, lip shape and other information of the hearing impaired, and provide a standard lip-reading sequence for the learning and training of the hearing impaired.

It should be noted that the use of 2D CNNs for feature extraction in lip-reading when dealing with sequential inputs is limited because such an architecture would only learn spatial features without learning temporal features. Even if dynamic frames were to be used as opposed to static frames, the architecture would still be compromising on the loss of spatial features, so it is necessary to learn both spatial and temporal information. It is for this reason that 3D or spatiotemporal CNNs were introduced into lip-reading.

D. 3D CNNs

The obvious difference between 2D and 3D networks is the extra dimension involved in the convolution process with the time dimension so the expression for convolution in Eq. 7 for a 3D CNN will be similar to that of Eq. 6 but with a time parameter t. Figure 8 shows an outline a lip-reading system with a 3D CNN frontend.

$$(y \otimes \omega)_{wht} = \sum_{c=1}^{C} \sum_{t'=1}^{k_t} \sum_{w'=1}^{k_w} \sum_{h'=1}^{k_h} y_{ct'w'h'} \omega_{c,t'+t,w'+w,h'+h} \quad (7)$$

Assael *et al.* [91] proposed an architecture with a frontend consisting of a spatiotemporal CNN, which extracts features from lip images with RGB pixels once pre-processing had been applied to videos from the GRID dataset which the architecture was trained and tested on. The backend consisted of 2 bidirectional GRUs, a softmax layer using ASCII characters as classes and a CTC for temporal alignments. Fung and Mak [92] proposed an architecture for decoding 10 sentences from the OuluVS2 corpus and they used a similar network for their backend, though their frontend used

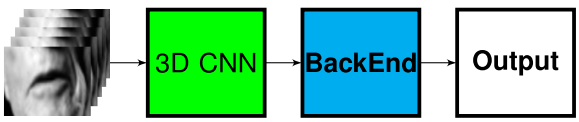

FIGURE 8. Diagram with 3D CNN frontend.

more 3D convolution layers and used max-out activation function instead of pooling. Their backend consisted of two bidirectional LSTMs with a softmax layer for classification whereby sentences were treated as individual classes, unlike Assael *et al.*'s [91] system which predicted sentences as sequences of ASCII characters.

Torfi *et al.* [92] proposed an audio-visual speech system that uses a coupled 3D CNN for the visual stream with grayscale images as the input and four layers of 3D convolution in total. For the audio stream, the first layer uses a 3D convolutional layer to extract spatiotemporal features after extracting MFCC features from speech signals; whereas the second layer uses 2D convolution to extract spatiotemporal features. The outputs of both streams are then combined into a representation space, so that the correspondence between the audio and visual streams can be evaluated.

Chung *et al.* [38] constructed an audio-visual speech recognition system called Watch, Listen, Attend, and Spell (WLAS) which consists of four components: Watch, Listen, Attend, and Spell. The front-end consists of a ''Watch'' component for the visual stream and a ''Listen'' for the audio component, with ''Attend'' and ''Spell'' components making up the back-end. The Watch component processes 5 consecutive grayscale images at a time with five 3D convolution layers, one fully connected layer, and three LSTM layers. Each LSTM at every timestep is part of an overall encoder LSTM configuration. The Listen component for the audio stream follows a similar structure except that MFCCs are used to extract features from the audio inputs as opposed to CNNs. The Spell component of the back-end network consists of three LSTMs, two attention mechanisms [93], and a Multi-layer Perceptron(MLP). The attention mechanisms process the context information of Watch and Listen to generate the context vectors for the Watch and Listen components. The decoder LSTM network in Spell uses the previous step output, the previous decoder LSTM state and the previous context vectors of Watch and Listen to generate the decoder state and output vectors. Finally, a MLP and softmax layer predict the outputted sentence by generating probability distribution of possible output ASCII characters.

Xu *et al.* [94] proposed a network called LCANet specifically designed to encode rich semantic features, that was trained on the GRID corpus and decodes sentences on an ASCII character-level. The frontend of the LCANet entails 3D convolutional layers and a highway network, while the backend uses Bidirectional GRU networks with a Cascaded Attention-CTC. The LCANet takes in images frames and uses the 3D-CNN to encode both spatial and temporal information with two layers of highway networks [95] on top of the 3D-CNN. The highway network module has two gates that allows the neural network to transfer some input information directly to the output.

Yang *et al.* [41] proposed an architecture called the D3D model for lip-reading Chinese words from the LRW-1000 dataset. It consists of a front-end with a spatiotemporal CNN following a similar topology to that of DenseNet that has stages of Convolution, Batch Normalization and pooling at the beginning; followed by three combinations of a DenseBlock and Trans-Block, plus a final Dense-Block at the end. Each Dense-Block contains two successive layers of convolution and batch normalisation while the Trans-Block contains three layers that include Batch Normalization, Convolution and Average Pooling. The backend consists of two Bidirectional GRUs with a softmax layer of 100 classes for each of the 100 words in the LRW-1000 dataset.

Chen *et al.* [64] constructed a neural network for Mandarin sentence-level lipreading consisting of two sub-networks. To predict the Hanyu Pinyin sequence for the input lip sequence, they combined a 3D CNN and a DenseNet with a two layer resBi-LSTM for the first part of the network, which was trained by a CTC loss function. The second part of the network converted Hanyu Pinyin into Chinese characters, and it consisted of a set of multi-headed attention that was trained using the cross-entropy loss function. The procedure in converting Hanyu Pinyin to Chinese characters does result in an 8% drop in accuracy rate. In consideration of the result, Chinese characters would be diverse on account of the different contexts whether Hanyu Pinyin is same or not.

3D CNNs can extract both spatial and temporal features more effectively than 2D CNNs. However, one drawback of 3D CNNs is that they require more powerful hardware and thus require high computation and storage costs. A compromise that is often made is to alleviate the limitations of both scenarios by using a 3D + 2D convolution neural network which consists of a mixture of 2D and 3D convolution layers. This helps to extract the necessary temporal features of lip movements and to limit the hardware capabilities required in performing feature extraction for lip-reading.

E. 2D + 3D CNNs

Frontends with a mixture of 2D and 3D CNNs will perform a combination of operations given in Eqs. 6 and 7. Figure 9 shows an outline a lip-reading system with a frontend containing 2D and 3D CNNs.

Stafylakis and Tzmiropoulos [96] proposed a visual speech recognition system for decoding words from the LRW corpus using grayscale images as an input. The front-end network consists of a 3D CNN and 2D ResNet, in which the 3D CNN has just one layer with which to extract short-term features of lip movements. The 2D ResNet has 34 layers which includes a max-pooling layer for reducing the feature vector's spatial dimensionality until the output is a one-dimensional feature vector. The backend is a two-layer Bidirectional LSTM with a softmax layer to classify one of 500 word classes.

Stafylakis and Tzmiropoulos proposed a visual speech system in [97] similar to that of [96] but with modifications to the architecture which included the use of word embeddings, to summarize the information of the mouth region that is relevant to the problem of word recognition, while suppressing other varying attributes such as speaker, pose and illumination. Other modifications from their architecture

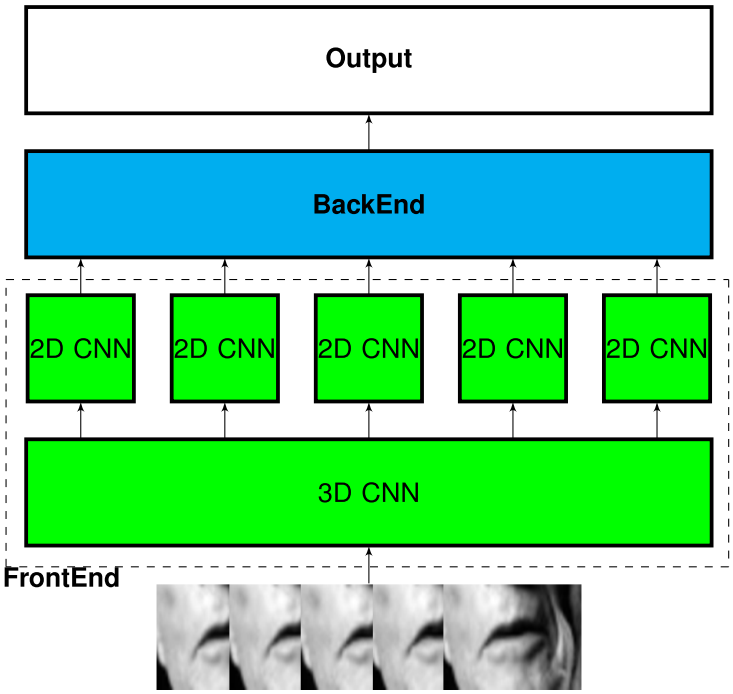

FIGURE 9. Diagram with frontend composed of 2D and 3D CNN kernels.

of [96] include the use of a smaller ResNet to reduce the total number of parameters from ∼ 24 million to ∼ 17 million, and of word boundaries passed to the backend as an additional feature.

Margam *et al.* [98] devised a 3D+2D CNN architecture configuration for decoding ASCII characters to predict spoken sentences from the GRID corpus, taking in RGB-pixelated images frames as an input. Their frontend consisted of two blocks of 3D CNNs followed by two blocks of 2D CNNs; where each 3D CNN block consists of a layer for convolution, pooling and batch normalisation, and each 2D CNN block will consist of layers for convolution and batch normalisation. Their backend consists of two Bidirectional LSTMs with a CTC for temporal alignment.

In summary, CNNs are the most widely used network for feature extraction techniques in deep learning-based automated lip-reading. They have advantages over Autoencoders, RBMs and Feed-forward networks in that they are more effective at learning both spatial and temporal features as well as being the most effective in extracting relevant features from any redundant features. For spatio-temporal data, frontends will either deploy 2D CNNs, 3D CNNs or 2D+3D CNNs; but the use of 2D+3D CNNs appears to be the most preferred as they are a compromise between being able to extract the necessary temporal features of lip movements in the most effective way, and to limit the hardware capabilities required in performing feature extraction.

V. CLASSIFICATION

The first neural network-based lip-reading systems were designed to classify isolated speech units such as individual letters, digits and words; where each speech segment or word was codified a class. This approach was sufficient for classifying visual speech that was limited to a limited number of discrete classes. For many systems that classified individual words such as Saitoh *et al.* [86] or Ngiam *et al.* [6], it was sufficient to use a backend that was composed of only a softmax layer for classification. Both of their architectures consisted of a frontend with a CNN for feature extraction a softmax layer backend to classify one of the possible words that had been uttered from the list of possible words contained within either of the OuluVS2 and LRW corpuses respectively.

A backend with solely a softmax layer would be sufficient for classifying speech in the form of a limited number of phrases where each phrase is treated as a class like Saitoh *et al.* [86] did with their approach. However when people utter phrases or even longer words, there is temporal information that can be exploited by neural networks to decipher between phrases and long words, which is why many visual speech recognitions systems use backends with networks for processing temporal sequences such as Recurrent Neural Networks(RNNs). They give a neural network architecture greater discriminative power when distinguishing between classes by learning conditional dependencies. Table 2 lists many of the automated lip-reading approaches which use deep-learning classification networks respectively and many of them are listed in the works of [10] and [11].

A. RECURRENT NEURAL NETWORKS

RNNs are a sequence-based neural network used in many tasks including language modelling, machine translation and speech recognition. Recurrent Neural Networks(RNNs) can be used to predict sequences based on the output of particular timesteps which is what makes them useful for natural

TABLE 2. Performance of lip-reading systems with deep learning-based classification algorithms.

Year	Reference	Feature Extractor	Classifier	Dataset	Class	Segment	Accuracy(%)
2011	Ngiam et al. [6]	Sparse Tensor PCA	Autoencoder	AVLetters	Alphabet	Alphabet	64.40
2013	Huang and Kingsbury [107]	DCT *plus* LDA	Deep Belief Network	Own data	Digits	Digits	35.70
2015	Moon et al. [108]	Deep Belief Network		AVLetters	Alphabet	Alphabet	55.30
2015	Mroueh et al. [33]	Scattering coefficients *plus* LDA	Feed-forward	IBM AV-ASR	Phonemes	Sentences	30.64[P]
2015	Thangthai et al. [109]	AAM	Feed-forward	RM-3000	Phonemes	Sentences	77.49
2015	Thangthai et al. [109]	HiLDA	Feed-forward	RM-3000	Phonemes	Sentences	84.67
2016	Almajai et al. [110]	LDA *plus* MLLT *plus* SAT	Feed-forward	LILiR	Phonemes	Phrases	53.00
2016	Assael et al. [91]	3D-CNN	Bidirectional GRU *plus* CTC	GRID	ASCII	Phrases	93.40
2016	Chung and Zisserman [84]	VGG-M	LSTM	OuluVS2	Phrases	Phrases	31.90
2016	Chung and Zisserman [84]	SyncNet	LSTM	OuluVS2	Phrases	Phrases	94.10
2016	Chung and Zisserman [40]	CNN		LRW	Words	Words	61.10
2016	Chung and Zisserman [40]	CNN		OuluVS	Phrases	Phrases	91.40
2016	Chung and Zisserman [40]	CNN		OuluVS2	Phrases	Phrases	93.20
2016	Lee et al. [85]	CNN	LSTM	OuluVS2	Phrases	Phrases	81.10
2016	Petridis and Pantic [111]	DBNF *plus* DCT	LSTM	AVLetters	Visemes	Alphabet	58.10
2016	Petridis and Pantic [111]	DBNF *plus* DCT	LSTM	OuluVS	Visemes	Phrases	81.80
2016	Saitoh et al. [86]	CFI *plus* NIN		OuluVS2	Phrases	Phrases	81.10
2016	Saitoh et al. [86]	CFI *plus* AlexNet		OuluVS2	Phrases	Phrases	82.80
2016	Saitoh et al. [86]	CFI *plus* GoogLeNet		OuluVS2	Phrases	Phrases	85.60
2016	Garg et al. [82]	CFI *plus* VGG	LSTM	MIRACL-VC	Words+Phrases	Words+Phrases	76.00
2016	Wand et al. [112]	Feed-forward	LSTM	GRID	Words	Phrases	79.50[b]
2017	Chung and Zisserman [38]	CNN	LSTM *plus* attention	OuluVS2	ASCII	Phrases	91.10
2017	Chung and Zisserman [38]	CNN	LSTM *plus* attention	MV-LRS	ASCII	Sentences	43.60
2017	Chung et al. [47]	CNN	LSTM *plus* attention	LRW	ASCII	Words	76.20
2017	Chung et al. [47]	CNN	LSTM *plus* attention	GRID	ASCII	Phrases	97.00
2017	Chung et al. [47]	CNN	LSTM *plus* attention	LRS	ASCII	Sentences	49.80
2017	Petridis et al. [77]	Autoencoder	LSTM	OuluVS2	Phrases	Phrases	84.50
2017	Petridis et al. [78]	Autoencoder	Bidirectional LSTM	OuluVS2	Phrases	Phrases	91.80
2017	Petridis et al. [79]	Autoencoder	Bidirectional LSTM	OuluVS2	Phrases	Phrases	94.70
2017	Torfi et al. [92]	3D CNN	Contrastive Loss	LRW	Words	Words	98.50
2017	Stafylakis and Tzimiropoulos [96]	3D-CNN *plus* ResNet	Bidirectional LSTM	LRW	Words	Words	83.00
2017	Stafylakis and Tzimiropoulos [97]	3D-CNN *plus* ResNet *plus* word boundaries	Bidirectional LSTM	LRW	Words	Words	88.08
2017	Wand and Schmidhuber [113]	Feed-forward	LSTM	GRID	Words	Phrases	42.40
2018	Afouras et al. [102]	3D-CNN *plus* ResNet	Bi-LSTM *plus* Language Model	LRS2	ASCII	Sentences	37.80
2018	Afouras et al. [102]	3D-CNN *plus* ResNet	Depthwise CNN	LRS2	ASCII	Sentences	45.00
2018	Afouras et al. [102]	3D-CNN *plus* ResNet	Attention-Transformer	LRS2	ASCII	Sentences	50.00
2018	Fung and Mak [78]	3D-CNN	Bidirectional LSTM	OuluVS2	Phrases	Phrases	87.60
2018	Hashmi et al. [114]	CFI *plus* CNN		MIRACL-VC	Words+Phrases	Words+Phrases	52.90
2018	Petridis et al. [106]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW	Words	Words	82.00
2018	Petridis et al. [13]	Autoencoder	Bidirectional LSTM	AV Digits	Phrases	Phrases	69.70
2018	Petridis et al. [13]	Autoencoder	Bidirectional LSTM	AV Digits	Digits	Digits	68.00
2018	Wand et al. [115]	Feed-forward	LSTM	GRID	Words	Phrases	84.70
2018	Xu et al. [94]	3D-CNN *plus* Highway	Bidirectional GRU *plus* Attention	GRID	ASCII	Phrases	97.10
2018	Afouras et al. [116]	3D-CNN *plus* ResNet	Transformer-CTC	LRS2	ASCII	Sentences	45.30
2018	Afouras et al. [116]	3D-CNN *plus* ResNet	Transformer-Seq2seq	LRS2	ASCII	Sentences	51.70
2018	Yang et al. [41]	2D CNN	Bidirectional GRU	LRW-1000	Words	Words	25.76
2018	Yang et al. [41]	DenseNet3D	Bidirectional GRU	LRW-1000	Words	Words	34.76
2018	Yang et al. [41]	2D+3D CNN	Bidirectional GRU	LRW-1000	Words	Words	38.19
2018	Mattos et al. [117]	CNN		GRID	Visemes	Visemes	64.80
2018	Oliveira et al. [118]	CNN		GRID	Visemes	Visemes	67.30
2019	Lu et al. [63]	CNN	LSTM *plus* Attention	Own data	Digits	Digits	88.20
2019	Shillingford et al. [8]	3D-CNN	Bidirectional LSTM *plus* Finite-state transducer	LSVSR	Phonemes	Sentences	59.10
2019	Shillingford et al. [8]	3D-CNN	Bidirectional LSTM *plus* Finite-state transducer	LRS3-TED	Phonemes	Sentences	44.90
2019	Courtney and Sreenivas [119]	Res-Bi-Conv-LSTM		LRW	Words	Words	85.20
2019	Jang et al. [120]	CFI *plus* QVGG *plus* Committee		OuluVS2	Phrases	Phrases	90.90
2019	Zhou et al. [121]	CNN	Bidirectional LSTM *plus* Modality Attention Mechanism	Chinese TV	Chinese + ASCII	Sentences	93.15
2019	Mesbah et al. [38]	CFI *plus* Hahn CNN		OuluVS2	Phrases	Phrases	93.72
2019	Mesbah et al. [88]	CFI *plus* Hahn CNN		LRW	Words	Words	58.20
2019	Margam et al. [98]	2D+3D CNN	Bidirectional LSTM *plus* CTC	GRID	Words	Sentences	98.70
2019	Margam et al. [98]	2D+3D CNN	Bi-LSTM *plus* CTC	Indian English	Words	Sentences	87.70
2019	Weng and Kitani [122]	3D-CNN	Bi-LSTM	LRW	Words	Words	84.11
2019	Zhang et al. [89]	VGG-M *plus* ResNet *plus* Bi-LSTM *plus* CTC	GRU *plus* Attention	CCTC	Pinyin-to-Hanzi	Sentences	50.20
2019	Wang et al. [123]	3D-CNN	Bi-Conv-LSTM	LRW	Words	Words	83.34
2019	Wang et al. [123]	3D-CNN	Bi-Conv-LSTM	LRW-1000	Words	Words	36.91
2020	Lu et al. [90]	CNN *plus* ResNet	LSTM	Own data	Digits	Digits	87.00
2020	Chen et al. [64]	3D-CNN	resBi-LSTM	NSTDB	Pinyin-to-Hanzi	Words	49.56
2020	Zhang et al. [124]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW	Words	Words	85.20
2020	Zhang et al. [124]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW-1000	Words	Words	45.24
2020	Xiao et al. [125]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW	Words	Words	84.13
2020	Xiao et al. [125]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW-1000	Words	Words	41.93
2020	Luo et al. [126]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW	Words	Words	83.50
2020	Luo et al. [126]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW-1000	Words	Words	38.70
2020	Zhao et al. [127]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW	Words	Words	84.41

TABLE 2. ***(Continued.)*** **Performance of lip-reading systems with deep learning-based classification algorithms.**

2020	Zhao et al. [127]	3D-CNN *plus* ResNet	Bidirectional GRU	LRW-1000	Words	Words	38.79
2020	Fenghour et al. [9]	3D-CNN *plus* ResNet	Linear Decoder Transformer *plus* GPT Transformer	LRS2	Visemes	Sentences	64.60
2020	Martinez et al. [105]	3D-CNN *plus* ResNet	Temporal CNN	LRW	Words	Words	85.30
2020	Martinez et al. [105]	3D-CNN *plus* ResNet	Temporal CNN	LRW-1000	Words	Words	41.40
2020	Ma et al. [128]	3D-CNN *plus* ResNet	Temporal CNN	LRW	Words	Words	88.36
2020	Ma et al. [128]	3D-CNN *plus* ResNet	Temporal CNN	LRW-1000	Words	Words	43.65
2021	Ma et al. [104]	3D-CNN *plus* ResNet *plus* Conformer Encoder	Decoder Transformer	LRS2	Pinyin-to-Hanzi	Sentences	62.10
2021	Ma et al. [129]	3D-CNN *plus* ResNet	Temporal CNN	LRW	Words	Words	88.50
2021	Ma et al. [129]	3D-CNN *plus* ResNet	Temporal CNN	LRW-1000	Words	Words	46.60

b - Speaker Dependent V - Viseme accuracy P - Phoneme accuracy C - Correctness

language processing tasks where in language models for instance, they can predict the next character in a word or the next word in a sequence of words [5]. A vanilla RNN is the simplest form of RNN, but Vanilla RNNs do suffer from the problem of vanishing gradients when trying to learn long-term dependencies. This is why RNNs used for lip-reading generally take the form of LSTMs or GRUs which consist of gates to control information that is transmitted through the network cells to control the gradient's value.

An LSTM is one variant of RNN which uses three gates to regulate the state and output at different timesteps [99]. An LSTM uses its gate structure to combine long and short-term memory to alleviate the problem of vanishing gradients. GRUs [100] are a more simplified form of RNN in comparison to LSTMs as they use just two gates instead of three. A diagram of an LSTM cell is shown in Figure 10 while a diagram of a GRU cell is shown in Figure 11.

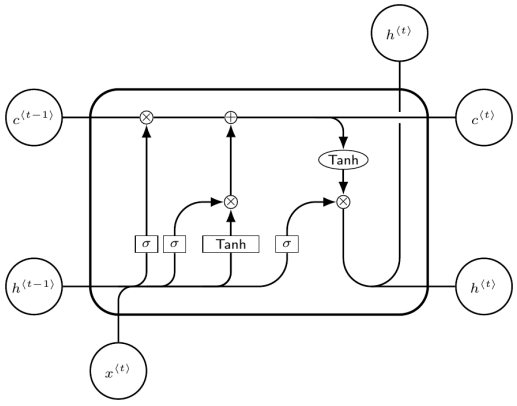

FIGURE 10. **Long-Short Term Memory Cell.**

Unidirectional RNNs rely on just forward transmission, whereby the output depends on the input at that particular timestep and the output of the previous timestep. Bidirectional RNNs however rely on both forwards and backwards transmission where the output of a particular timestep relies not just on the current input and previous timestep output, but also on the successive timestep output too. A speech segment can be dependent on the successive segment as well as the previous one however. Bidirectional RNNs do use roughly double the number of parameters and so take longer to train.

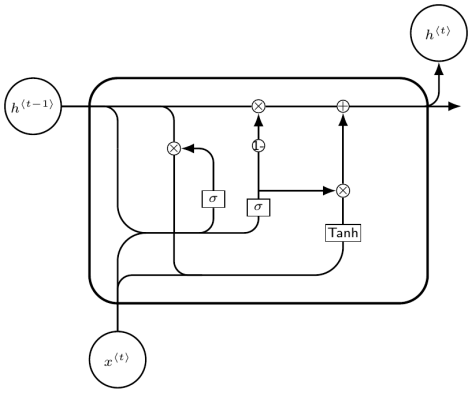

FIGURE 11. **Gated Recurrent Unit Cell.**

For lip-reading sentences that are more random and not repetitive such as those in the TIMIT and LRS2 corpuses, it is not possible to encode each sentence as a class and even to encode each word as a class is not feasible because of there are thousands of different possible words to account for. Visual speech recognition systems that decode sentences will often use ASCII characters to decode sentences by learning conditional dependence relationships of how they appear in words.

When automating speech recognition in real time, information about where a particular character starts and ends in the image frame sequence will generally be unavailable and the use of RNNs to learn sequences of characters will not be sufficient without being able to learn the temporal alignment of the sequence.

B. ATTENTION MECHANISMS + CTCs

An Attention mechanism is one way of learning to temporally align predictions of an input sequence. An attention-based RNN will predict a decoder state s and for every timestep, a context vector c_i will be generated which is an indicator of how dependant the output at a timestep is to the output of another particular timestep.

The context vector of a timestep is generated by calculating an alignment model e_{ij} which scores how well the input around position j and the output at position i match. This alignment model is then exponentiated and normalised by dividing by the sum of exponentiated alignment models to give a weight α_{ij}. Finally, the context vector for the timestep is calculated by summing over the all weights and annotations

for that timestep. Using the decoder state and context vectors, the RNN can construct an output probability distribution to predict an output sequence. Relationships between all the variables are shown in Eqs. 8 to 12.

$$e_{ij} = a\left(s_{i-1}, h_j\right) \tag{8}$$

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T} \exp\left(e_{ik}\right)} \tag{9}$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{10}$$

$$p(y_i|y_1, \ldots, y_{i-1}, x) = g\left(y_{t-1}, s_t, c\right) \tag{11}$$

$$s_i = f\left(s_i, y_{i-1}, c_i\right) \tag{12}$$

There are two main problems posed by using attention mechanisms for temporal alignment in automated lip-reading. The first is the length variation between the input and output sequences in speech recognition that makes it more difficult to track the alignment and secondly, the basic temporal attention mechanism is too flexible and allows for extremely non-sequential alignments.

A Connectionist Temporal Classification (CTC) [101] model predicts frame labels and then looks for the optimal alignment between the frame predictions and the output sequence. A CTC can resolve the problem of input sequences and output sequences not being equivalent in length because of people speaking at different speeds.

If T is taken to the number of time steps in the sequence model, for example $T = 3$, a CTC defines the probability of the string "me" as $p(mme) + p(m\epsilon e) + \ldots + p(mee)$ and there exists a ϵ symbol in the case of repeated characters to make sure that the CTC does not group symbols when there are supposed to be repetitions.

For an input sequence $X = [x_1, x_2 \ldots, x_T]$ to a backend, an output sequence $Y = [y_1, y_2, \ldots, y_U]$ is predicted and the aim is to find the most likely sequence Y^*. A label l will have a set of possible paths with each path π corresponding to a possible frame prediction sequence. Eqs. 13 to 15 indicate how the CTC loss L_{CTC} is calculated.

$$p(\pi|x) = \prod_{t=1}^{T} p(\pi_t|x) \tag{13}$$

$$p(l|x) = \sum_{i} p(\pi_i|x) \tag{14}$$

$$L_{CTC} = -\ln p(l|x) \tag{15}$$

Assael *et al.* [91] were the first to introduce CTCs into lipreading when ASCII characters were used as units of classification. Bidirectional GRUs were used in the backend along with a CTC for temporal alignment and a CTC loss function to train the system.

The use of CTCs do have constraints, one being that input sequences must be longer than output sequences. CTCs also assume that character labels are conditionally independent and that each output is the probability of observing one particular label at a particular timestep. CTCs therefore focus more on local information from nearby frames than global information from all frames. It for this reason that lip-reading systems that use attention mechanisms perform better than those with CTCs for visual only speech recognition; whereas those that use CTCs are the better option for audio-visual speech recognition when there is available audio.

Xu *et al.* [94] tackle the problem of the conditional independence limitation in CTCs by using a Cascaded Attention-CTC which tries to capture information from a longer context. Their frontend follows an Encoder-Decoder structure with two bidirectional GRUs in the Encoder and an Attention-CTC configuration with a hidden layer in between the Encoder and Decoder. The Decoder alleviates the conditional independence limitation by cascading the CTC with attention. This not only serves to address limitations of the CTC but also the limitations of using an Attention mechanism by itself because a Cascaded Attention-CTC can reduce uneven alignments during training in order to eliminate unnecessary non-sequential predictions between the decoded result and ground truth.

C. TRANSFORMERS

RNNs account for the majority of frontend networks in neural network based lip-reading systems. However, a new trend in the use of Transformers has emerged in some of the most recent approaches to classification in lip-reading and they are appear to be replacing RNNs in many lip-reading systems.

Transformers are designed to allow parallel computation by processing entire inputs as at once rather than processing them sequentially like RNNs. Transformers require less time to train than RNNs because they avoid recursion, and they are better at capturing long term dependencies.

Afouras *et al.* [102] proposed three architectures that perform ASCII character-level classification for lip-reading sentences from the BBC LRS2 dataset. All three systems consist of an identical frontend with a 3D-CNN followed by a ResNet. The first architecture consisted of a backend with three stacked Bidirectional LSTMs trained with a CTC loss, and where decoding was implemented using a beam search that utilised information from an external language model. The second system used an attention-based transformer with an encoder-decoder structure that follows the baseline model of [103]. The Transformer model was the best performing model and it attained better word accuracies than the Bidirectional LSTM for every evaluation scenario and the author observed for instance that the Transformer model was far better at generating to longer sequences than the Bidirectional LSTM model - particularly for sequences longer than 80 frames. Moreover, the Bidirectional LSTM model had a limited capacity for learning long-term, non-linear dependences and modelling complex grammar rules because of the CTC's assumption of timestep outputs being conditionally independent.

Ma *et al.* [104] proposed an audio-visual lip-reading system with a frontend composed of a spatiotemporal CNN and a ResNet-18 network. The visual backend uses the

"Conformer" variant of the Transformer which follows a similar structure to that of Vaswani *et al.* [103]. It is convolution-augmented in that it uses convolutional layers in the Encoder because whilst Transformers are good at modelling long-range global context, they are less capable of extracting fine-grained local feature patterns - whereas CNNs can exploit local information.

A MLP is used to concatenate the outputs of the audio and visual streams whereby the output of the MLP forms the input of the Transformer Decoder which uses a hybrid CTC/Attention model that is specifically designed to address the individual limitations to the use of either a CTC or Attention model individually. This is done by generating a loss for the CTC and for the Conformer Encoder individually and adding them together using aggregated loss function [104] (Eq. 16).

$$Loss = \alpha \log p_{CTC}(\boldsymbol{y}|\boldsymbol{x}) + (1-\alpha) \log p_{CE}(\boldsymbol{y}|\boldsymbol{x}) \quad (16)$$

D. TEMPORAL CONVOLUTIONAL NETWORKS

Temporal Convolutional Networks(TCNs) are another form of neural network that have emerged as an alternative to RNNs for sequence classification. Recently in many NLP tasks there has been a move towards the use of purely convolutional models for sequence modelling.

Like Transformers, TCNs have an advantage over RNNs in that they can process inputs in parallel as opposed to processing the input at every timestep sequentially. They are also advantageous because they are flexible in changing receptive field size; which can be done by stacking more convolutional layers, using larger dilation factors, or increasing filter size which allows for better control of the model's memory size. Furthermore, TCNs do not suffer from the problem of exploding or vanishing gradients because they have a backpropagation path different from the temporal direction of the sequence, as well as lower memory requirement for training - particularly for long input sequences.

The third backend system used by Afouras *et al.* [102] for lip-reading sentences from the BBC LRS2 corpus was a Fully Convolutional(FC) model containing depth-wise separable convolution layers, which consists of layers for performing convolution along the spatial and temporal channel dimensions. The network contains 15 convolutional layers that were trained with a CTC loss where the decoding was performed in the same way as the Bidirectional LSTM system [102]. The FC model has advantages over the other two systems namely the transformer-based and Bi-LSTM-based systems, in that it uses fewer parameters and is quicker to train. Afouras *et al.* also noted that the FC model gave them greater control over the amount of future and past context by adjusting the receptive field. The FC model performed better than the Bidirectional LSTM model, though it did deliver diminishing returns on performance for sequences longer than 80 frames.

Martinez *et al.* [105] constructed a word-based lip-reading system similar to that of Petridis *et al.* [106] with a similar frontend that entails a spatiotemporal CNN followed by a ResNet-18 CNN. For the backend, the Bidirectional GRU has been substituted with a network in its place that they proposed called a Multi-Scale Temporal Convolutional Network (MS-TCN); devised to tailor the receptive field of a TCN so that long and short term information can be mixed up. A MS-TCN block consists of a series of TCNs, each with a different kernel size whereby the outputs are concatenated. Their system was trained and evaluated on the English datasets LRW and Mandarin dataset LRW-1000 achieving word accuracies of 85.3% and 41.4% respectively. In addition to improving on the accuracy of the system for Petridis *et al.* [106], they also noted a reduction in the overall GPU training time which was reduced by two thirds.

Ma *et al.* propose modifications to the system of Martinez *et al.* by using a Densely Connected Temporal Convolutional Network (DC-TCN) instead of the MS-TCN contained within the frontend for the aim of providing denser and more robust temporal features. Two variants are used including Fully-Dense(FD) and Partially-Dense(PD) architectures, as well as an additional "Squeeze and Excitation" block within the network which is a lightweight attention mechanism to further enhance the model's classification power. They improve on the word accuracies of Martinez *et al.* to record word accuracies on the LRW and LRW-1000 datasets of 88.4% and 43.7%.

In summary of classification techniques, RNNs are the most frequently used backend network for predicting spoken sentences and are often used in conjunction with mechanisms for learning temporal alignment such as CTCs or Attention mechanisms. CTCs align sequences based on the conditional independence assumption, whereas attention mechanisms are better at modelling conditional dependence and this is why CTCs are the better option for audio-assisted speech recognition and why attention mechanisms are more effective for visual only speech recognition. RNNs however have started to be superseded by the use of Attention-Transformers and TCNs which both have advantages over RNNs in that they can perform parallel computation and are better at learning long-term dependencies. Out of all three networks, Attention-Transformers appear to have attained the best classification performance results when predicting sentences. However, TCNs do have advantages over both RNNs and transformers in that they take less time to train and are more flexible in changing receptive field size.

VI. CLASSIFICATION SCHEMA

The first automated approaches to lip-reading started off with recognising a limited number of speech units in the form of digits, letters and words; especially as the first audio-visual datasets that were available for training lip-reading systems were limited and only focused on the classification of small isolated speech segments. For this reason it was sufficient to encode each speech segment as a class.

Eventually, the emergence of more audio-visual training data covering a wider range of vocabulary saw the development of lip-reading systems with entire words a classes. Some

approaches encoded entire phrases when performing the task of speech recognition in videos of people uttering a limited number of structured and repetitive phrases.

Some of the largest and most recent of lip-reading corpuses consist of people speaking in a continuous manner with vocabularies coverings thousands of different words, and so many lip-reading systems that have been trained to predict entire sentences have opted for the use ASCII characters as a classification schema as opposed to encoding every word as a single class. This allows for fewer classes to be used and for a reduction in the creation of computational bottleneck [130]. The use of ASCII characters also allows for natural language to be modelled due to the conditional dependence relationships that exist between ASCII characters. This makes it easier to predict characters and words [38], [91], [96].

However, even the use of ASCII characters for automated lip-reading of speech covering an extensive range of vocabulary has its limitations. Neural networks for speech recognition systems that use either words or ASCII characters as classes are only able to predict words that the system has been trained to predict, because in the case of using words as a class, the word needs to be encoded as a class and have been present in the training phase. While for the case of ASCII characters, the prediction of words is based on combinations of characters having been observed in training as patterns.

Furthermore, the models must be trained to cover a wide range of vocabulary which would require a significant number of parameters, lots of hyperparameters to be optimised and a significant volume of training data to be used. This is in addition to the requirement of curriculum learning-based strategies [131], [132] which involve further pre-processing, such as the clipping of training videos with individuals speaking so that the models can be trained on single word examples to begin with, before gradually incrementing the length of the sentences being spoken.

Other less frequently used classification schema include visemes and phonemes. The usage of visemes for decoding speech when trying to predict sentences has some unique advantages. Firstly, the prediction of speech as sequences of visemes as classes as opposed to sequences of either words or ASCII characters would require a smaller overall number of classes which alleviates computational bottleneck. In addition, the use of visemes does not require pre-trained lexicons, which means that a lip-reading system which classifies visemes can in theory be used to classify words that have not been seen during training. A lip-reading system that predicts speech using visemes as classes can be generalised to decoding speech from people speaking in other languages because many different languages often share identical visemes.

The general classification performance for recognising individual segmented visemes has been less satisfactory compared with the classification of words. This is due to the natures of visemes tending to have a shorter duration than words which results in there being less temporal information available to distinguish between different classes, as well as there being more visual ambiguity when it comes to class recognition [118].

Moreover, the eventual prediction of words and sentences based on decoding visemes requires a two-stage procedure where visemes will be decoded as the first stage and with a viseme-to-word conversion process being performed as the second stage. One set of visemes can correspond to multiple different sets of phonemes or sounds; unlike the use of ASCII characters where there is one-to-one mapping relationship when mapping characters to possible words or sentences.

The viseme-to-word conversion is a challenge because once visemes have been classified, there is a need to disambiguate between homopheme words(words that look identical when spoken but sound different [133]). This bottleneck exists because of the one-to-many mapping correspondence between visemes and phonemes. The conversion process requires a language model to determine the most likely words that have been uttered.

Phonemes have been more frequently used than visemes as an intermediate classification schema in lip-reading where speech is decoded in the form of phonemes, which are then converted to words [8], [109], [134]–[136]. The classification of phonemes as individual units using only visual speech can never be done with as much precision as classifying individual visemes due to the fact that many phonemes share identical visemes and therefore look the same so context is needed to resolve that problem.

Phonemes are more preferred to visemes though because the conversion of phonemes to words will always comprise of less ambiguity than the conversion of visemes to words. This is because there are significantly fewer homophone words, or words that sound the same in the English language than homopheme words. Some of the language models used to perform the phoneme-to-word conversion such as WFSTs and HMMs use Markov chains and are limited in performing viseme-to-word conversion with good precision due to their inability to detect semantic and syntactic information needed to discriminate between words with identical visemes.

It is still remains to be seen which is the most accurate classification scheme to utilise out of visemes, phonemes and ASCII characters. The performance of a lip-reading system that uses ASCII characters can itself be enhanced by the inclusion of a language model which means the decoding of ASCII characters in predicting sentences can be performed as a two-stage procedure. Afouras *et al.* [102] do include a character-based language model to increase the likelihood of a word being correctly predicted however, some of the sentences that the model does not predict correctly are not as grammatically sound as the ground-truth sentences. So the model's performance itself could be enhanced by including a word-based language model to ensure that sentences being predicted are the most likely given the combination of words using a word-based language model to calculate sentence perplexity.

VII. PERFORMANCES IN LIP-READING

The AVLetters database is the most widely used corpus for alphabet recognition. Zhao *et al.* [49] used LBP-TOP for feature extraction and a Support Vector Machine(SVM) for classification and they attained a 62.80% word accuracy rate(WAR). Pei *et al.* [137] recorded the highest WAR of 69.60% with a RFMA based lip-reading system. Petridis and Pantic [111] used a frontend that combined Deep Belief Network features and DCT features, with an LSTM for the backend achieving a 58.10% classification accuracy. Hu *et al.* [138] proposed a system based on multimodal RBMs called Recurrent Temporal Multimodal Restricted Boltzmann Machines and achieved a WAR of 64.63%.

CUAVE is the most frequently used database for digit recognition. Papandreou *et al.* [139] used an AAM for feature extraction with a HMM for classification for performing digit recognition and they recorded a 83.00% word recognition rate. Ngiam *et al.* [6] achieved a 68.70% word recognition rate using an RBM-Autoencoder. Rahmani and Almasganj [140] extracted deep bottleneck features, and then used a GMM-HMM for the language model to achieve a WAR of 63.40%. Petridis *et al.* [77] achieved a WAR of 78.60% using the dual flow method.

GRID is one of the oldest and most frequently used databases for predicting phrases. Wand *et al.* [112] experimented with three different feature extraction techniques for their backend that included Eigenlips, HOG, and feedforward neural networks. The lip-reading systems that used Eigenlips and HOG for the respective frontends utilised an SVM for the backend, while the lip-reading system with the feedforward network in the frontend used an LSTM for the backend. Performance results indicate that the combination of the feedforward network with an LSTM was the best model. Assael *et al.* [91], Xu *et al.* [94] and Margam *et al.* [98] obtained word accuracies of 95.20%, 97.10%, and 98.70% respectively through the use of spatiotemporal convolutional networks and Bidirectional RNNs.

OuluVS2 is the most widely used multi-view database. Lee *et al.* [85] used a frontend that combined DCT and PCA features, and an HMM to attain a 63.00% word accuracy rate for phrase prediction. They also constructed a lip-reading system that utilised a CNN for feature extraction and an LSTM for classification achieving a 83.80% word accuracy rate. Wu *et al.* [141] combined SDF features with STLP features while using an SVM for classification, to achieve a 87.55% classification accuracy. Petridis *et al.* [65] obtained a 96.90% word recognition rate based on the three-stream method.

LRW is one of the most challenging datasets there is for word classification which Chung and Zisserman [40] used for training and validation. They obtained a word accuracy rate(WAR) of 61.10% with a spatiotemporal CNN, while Torfi *et al.* [92] used a coupled 3D CNN for their lip-reading system achieving a WAR of 98.50%. Stafylakis and Tzimiropoulos [96] used a 3D CNN and ResNet for their frontend with a Bidirectional LSTM backend obtaining a WAR of 83.00%. In recent years; Zhang *et al.* [124], Xiao *et al.* [125], Luo *et al.* [126] and Zhao *et al.* [127] have all used a frontend with a 3D CNN and ResNet along with a Bidirectional GRU for the backend and they all recorded state-of-the-art performance results on the LRW corpus with WARs of 85.20%, 84.13%, 83.50% and 84.41% respectively. The best results that were recorded for the validation on the LRW set were for the systems proposed by Martinez *et al.* [105] and Ma et el. [128], [129] who all used a 3D CNN and ResNet for the frontend with a TCN for the backend and they correspondingly achieved WARs of 85.30%, 88.36% and 88.50%. As discussed in Section V, TCNs have advantages over RNNs and they are set to replace RNNs for many sequence processing tasks.

For the BBC-LRS2 database, Chung *et al.* [38] proposed a Watch-Attend-and-Spell system that achieved a WAR of 49.80%. Afouras *et al.* [116] proposed two approaches which both used a 3D CNN plus ResNet for the front-end. One of their approaches used an attention-transformer for the backend that trained with a CTC loss achieving a WAR of 45.30%. Their other approach also used a backend with a Transformer, but that was trained with a seq2seq loss and achieved a WAR of 51.70%. Ma et al [104] proposed a frontend with a 3D-CNN, ResNet plus Conformer Encoder in tandem with a backend that used Decoder Transformer and accomplished a word accuracy rate of 62.1%. Finally, Fenghour *et al.* [9] devised a system that decoded videos in two stages where visemes were predicted for the first stage using a 3D-CNN plus ResNet with a Linear Decoder Transformer, and then words where predicted using a converter that calculated perplexity scores using the pre-trained GPT transformer. Fenghour *et al.* [9] achieved a WAR of 64.0%.

For the task of recognising shorter speech segments, traditional methods have outperformed deep learning-based methods in terms of performance. This is because deep learning requires large numbers of training samples and because the focus of automated lip-reading research has moved towards classifying larger speech units in the form of words and entire sentences in continuous speech, plus there is very little demand and effort to attempt to increase the volume of training samples for people uttering isolated digits and letters. For sentences prediction, deep learning methods significantly outperform traditional methods. For word and sentence prediction, Transformers and TCNs are starting to replace RNNs due to their ability to better perform parallel computation and learn long-term dependencies.

VIII. CONCLUSION

This survey reviews automated lip-reading systems running from 2007 to 2021. One can see a progressions of visual speech recognition systems moving from the use of traditional algorithms for letter and digit classification to the use of deep neural networks for predicting words and sentences thanks to the development of more advanced corpuses such as BBC-LRS2, LRS3-TED, LSVSR and LRW-1000. New datasets not only cover larger vocabularies covering

thousands of words and uttered by thousands of people, they also feature people speaking in varying poses, lighting conditions and resolutions.

Lip-reading systems consist of components for feature extraction and classification. 2D+3D CNNs are the most preferred network for frontends because of their ability to learn spatial and temporal features though Autoencoders do have the advantage of being able to map visual feature data from higher dimensional space into lower dimensional space without the need for any labelled classification.

RNNs in the form of LSTMs and GRUs form the majority of classification networks. In recent years though, Transformers and TCNs have started to replace RNNs due to their ability to better perform parallel computation, learn long-term dependencies and be trained in a shorter period of time.

A variety of different classification schema have been deployed where earlier classification networks encoded single words as a class and later networks have used ASCII characters to predict sentences covering huge lexicons. In theory, the use of phonemes and visemes could mean that lip-reading systems could be lexicon-free whereby a lip-reading system could predict a word spoken by an individual that did not appear in the training phase.

Other challenges inhibiting the progress of automated lip-reading still remain. These include the need to predict unseen words, i.e. predict spoken words that did not appear in training phase and are not covered by the lexicon as well as visual ambiguities where the semantic and syntactic features of words can be learned for words that look the same when spoken. From a visual perspective, there remains challenges such as speaker dependency, especially when attempting to generalise to speakers who have not appeared in the training data; the need to generalise to videos of varying spatial resolution and the need to generalise to videos of different frame rates while consisting of varying quantities of temporal data.

REFERENCES

[1] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. 6th Int. Conf. Multimodal Interface (ICMI)*, 2004, pp. 235–242.

[2] J. Jeffers and M. Barley, *Speechreading (Lipreading)*. Springfield, IL, USA: Charles C Thomas Publisher Limited, 1971.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Martigny, Switzerland, Tech. Rep., 2000.

[4] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, and M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in *Proc. 3DTV Conf.*, May 2007, pp. 1–4.

[5] S. Lee and D. Yook, "Audio-to-visual conversion using hidden Markov models," in *Proc. 7th Pacific Rim Int. Conf. Artif. Intell., Trends Artif. Intell.*, 2002, pp. 563–570.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn., (ICML)*, 2011, pp. 1–8.

[7] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 873–880.

[8] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. de Freitas, "Large-scale visual speech recognition," 2018, *arXiv:1807.05162*. [Online]. Available: https://arxiv.org/abs/1807.05162

[9] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020.

[10] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image Vis. Comput.*, vol. 78, pp. 53–72, Oct. 2018.

[11] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A survey of research on lipreading technology," *IEEE Access*, vol. 8, pp. 204518–204544, 2020.

[12] M. Igras, B. Ziolko, and T. Jadczyk, "Audiovisual database of Polish speech recordings," *Studia Inform.*, vol. 33, no. 2B, pp. 163–172, 2012.

[13] D. Estival, S. Cassidy, F. Cox, and D. Burnham, "AusTalk: An audio-visual corpus of Australian English," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 1–13.

[14] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6219–6223.

[15] A. Ortega, F. Sukno, E. Lleida, A. F. Frangi, A. Miguel, L. Buera, and E. Zacur, "AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2004, pp. 1–4.

[16] S. Antar and A. Sagheer, "Audio visual Arabic speech (AVAS) database for human-computer interaction applications," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 9, p. 7, 2013.

[17] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Interspeech*, Oct. 2004, pp. 2489–2492.

[18] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[19] S. J. Cox, R. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2008, pp. 179–184.

[20] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Comput. Sci.*, vol. 163, pp. 400–409, Jan. 2019.

[21] E. Bailly-Bailliere, "The BANCA database and evaluation protocol," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, 2003, pp. 625–638.

[22] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraa-Labastie, and F. Bimbot, "BL-database: A French audiovisual database for speech driven lip animation systems INRIA," Ph.D. dissertation, Nat. Inst. Res. Digit. Sci. Technol., Le Chesnay-Rocquencourt, France, 2011.

[23] X. Yanjun, D. Limin, L. Guoqiang, Z. Xin, and Z. Zhi, "Chinese audiovisual bimodal speeeh database CAVSR1.0," *Acta Acustica-Peking*, vol. 25, no. 1, pp. 42–44, 2000.

[24] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, and M. Fujimoto, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2010, pp. 85–88.

[25] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. IV-429–IV-432.

[26] Patterson, Gurbuz, Tufekci, and Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2002, pp. II-2017–II-2020.

[27] C. C. Chibelushi, F. Deravi, and J. S. Mason, "BT David database-internal rep," Dept. Elect. Electron. Eng., Speech Image Process. Res. Group, Univ. Swansea, Swansea, Wales, Tech. Rep., 1996. [Online]. Available: http://www-ee.swan.ac.uk/SIPL/david/survey.html

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[29] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL523–EL529, Jun. 2018.

[30] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *J. Speech Hearing Res.*, vol. 14, no. 4, pp. 677–709, Dec. 1971.
[31] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Zelezny, "HAVRUS corpus: High-speed recordings of audio-visual Russian speech," in *Proc. Int. Conf. Speech Comput.*, Aug. 2016, pp. 338–345.
[32] X. Lin, H. Yao, X. Hong, and Q. Wang, "HIT-AVDB-II: A new multi-view and extreme feature cases contained audio-visual database for biometrics," in *Proc. 11th Joint Conf. Inf. Sci. (JCIS)*, 2008, pp. 1–8, doi: 10.2991/jcis.2008.61.
[33] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2130–2134.
[34] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-visual speech recognition using an infrared headset," *Speech Commun.*, vol. 44, nos. 1–4, pp. 83–96, Oct. 2004.
[35] P. J. Lucey, G. Potamianos, and S. Sridharan, "Patch-based analysis of visual speech from multiple views," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2008, pp. 69–74.
[36] D. Petrovska-Delacretaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, M. Ardabilian, E. Krichen, M.-A. Mellakh, A. Chaari, S. Guerfi, J. D'Hose, and B. B. Amor, "The IV^2 multimodal biometric database (including iris, 2D, 3D, stereoscopic, and talking face data), and the IV^2-2007 evaluation campaign," in *Proc. IEEE 2nd Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep. 2008, pp. 1–7.
[37] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lip-reading," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2010, pp. 1–6.
[38] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.
[39] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," 2018, *arXiv:1809.00496*. [Online]. Available: https://arxiv.org/abs/1809.00496
[40] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 87–103.
[41] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.
[42] V. Estellers and J. P. Thiran, "Multipose audio-visual speech recognition," in *Proc. 19th Eur. Signal Process. Conf.*, Sep. 2011, pp. 1065–1069.
[43] O. Vanegas, K. Tokuda, and T. Kitamura, "Location normalization of HMM-based lip-reading: Experiments for the M2 VTS database," in *Proc. Int. Conf. Image Process.*, Oct. 1999, pp. 343–347.
[44] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," in *Proc. Int. Conf. Image Anal. Recognit.*, 2014, pp. 21–28.
[45] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernock, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2012, pp. 635–640.
[46] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, "An audio-visual corpus for multimodal automatic speech recognition," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 167–192, Oct. 2017.
[47] J. S. Son and A. Zisserman, "Lip reading in profile," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.
[48] A. G. Chitu, K. Driel, and L. J. Rothkrantz, "Automatic lip reading in the Dutch language using active appearance models on high speed recordings," in *Proc. Int. Conf. Text, Speech Dialogue*, 2010, pp. 259–266.
[49] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.
[50] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–5.
[51] A. Pass, J. Zhang, and D. Stewart, "AN investigation into features for multi-view lipreading," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2417–2420.
[52] D. L. Howell, "Confusion modelling for lip-reading, University of East Anglia," Ph.D. dissertation, School Comput. Sci., Dept. Sci., Univ. East Anglia, Norwich, U.K., 2015.
[53] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.
[54] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 851–858.
[55] Y. W. Wong, S. I. Ch'ng, K. P. Seng, L.-M. Ang, S. W. Chin, W. J. Chew, and K. H. Lim, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1503–1510, Oct. 2011.
[56] P. Cisar, M. Zelezny, Z. Krnoul, J. Kanis, J. Zelinka, and L. Müller, "Design and recording of Czech speech corpus for audio-visual continuous speech recognition," in *Proc. Auditory-Visual Speech Process Int. Conf.*, Jul. 2005, pp. 1–4.
[57] J. Trojanova, M. Hruz, P. Campr, and M. Zelezny, "Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2008, pp. 1239–1243.
[58] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, 2005, pp. 777–786.
[59] C. Sanderson, "The VidTIMIT database," IDIAP, Martigny, Switzerland, Tech. Rep., 2002.
[60] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 208–215.
[61] A. X. V. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20—A database for robust audiovisual speech recognition," in *Proc. Int. Conf. Lang. Resour. Eval.*, May 2010, pp. 1–4.
[62] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, 1999, pp. 965–966.
[63] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Appl. Sci.*, vol. 9, no. 8, p. 1599, Apr. 2019.
[64] X. Chen, J. Du, and H. Zhang, "Lipreading with DenseNet and resBi-LSTM," *Signal, Image Video Process.*, vol. 14, no. 5, pp. 981–989, Jul. 2020.
[65] T. Mohammed, R. Campbell, M. Macsweeney, F. Barry, and M. Coleman, "Speechreading and its association with reading among deaf, hearing and dyslexic individuals," *Clin. Linguistics Phonetics*, vol. 20, nos. 7–8, pp. 621–630, Jan. 2006.
[66] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, Aug. 1990.
[67] J. B. Millar and R. Goecke, "The audio-video Australian English speech data corpus AVOZES," in *Proc. Interspeech*, Oct. 2004, pp. 2525–2528.
[68] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition," in *Proc. Int. Conf. Multimodal Process. Interact.*, 2008, pp. 423–435.
[69] Y. Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 432–437.
[70] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.
[71] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, Sep. 2014.
[72] S. L. Phung, A. Bouzerdoum, D. Chai, and A. Watson, "Naive Bayes face/nonface classifier: A study of preprocessing and feature extraction techniques," in *Proc. Int. Conf. Image Process. (ICIP)*, Oct. 2004, pp. 1385–1388.
[73] M. Saaidia, A. Chaari, S. Lelandais, V. Vigneron, and M. Bedda, "Face localization by neural networks trained with Zernike moments and Eigenfaces feature vectors. A comparison," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 377–382.
[74] Q. M. Rizvi, "A review on face detection methods," *J. Manage. Develop. Inf. Technol.*, vol. 11, no. 2, pp. 1–11, 2011.
[75] S.-J. Lee, S.-B. Jung, J.-W. Kwon, and S.-H. Hong, "Face detection and recognition using PCA," in *Proc. IEEE Region 10 Conf. TENCON Multimedia Technol. Asia–Pacific Inf. Infrastruct.*, vol. 1, Sep. 1999, pp. 84–87.

[76] X. Li, T. Zhang, X. Zhao, and Z. Yi, "Guided autoencoder for dimensionality reduction of pedestrian features," *Int. J. Speech Technol.*, vol. 50, no. 12, pp. 4557–4567, Dec. 2020.
[77] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2592–2596.
[78] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end audiovisual fusion with LSTMs," in *Proc. 14th Int. Conf. Auditory-Visual Speech Process.*, Aug. 2017, pp. 36–40.
[79] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–14.
[80] J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2011, pp. 52–59.
[81] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Proc. Conf. Int. speech Commun. Assoc.*, 2014, pp. 1149–1153.
[82] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM," Stanford Univ., Stanford, CA, USA, Tech. Rep. CS231n Project, 2016.
[83] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–6.
[84] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 251–263.
[85] D. Lee, J. Lee, and K.-E. Kim, "Multi-view automatic lip-reading using neural network," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 290–302.
[86] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen, "Concatenated frame image based CNN for visual speech recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 277–289.
[87] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400
[88] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip reading with Hahn convolutional neural networks," *Image Vis. Comput.*, vol. 88, pp. 76–83, Aug. 2019.
[89] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu, and M. Liu, "Understanding pictograph with facial features: End-to-end sentence-level lip reading of Chinese," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9211–9218.
[90] Y. Lu, S. Yang, Z. Xu, and J. Wang, "Speech training system for hearing impaired individuals based on automatic lip-reading recognition," in *Proc. AHFE Virtual Conf. Hum. Factors Syst. Interact.*, 2020, pp. 250–258.
[91] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence level lipreading," in *Proc. ICLR Conf.*, 2016, pp. 1–13.
[92] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.
[93] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[94] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 548–555.
[95] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
[96] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Proc. Interspeech*, Aug. 2017, pp. 3652–3656.
[97] T. Stafylakis and G. Tzimiropoulos, "Deep word embeddings for visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4974–4978.
[98] D. Kumar Margam, R. Aralikatti, T. Sharma, A. Thanda, P. A K, S. Roy, and S. M Venkatesan, "LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models," 2019, *arXiv:1906.12170*. [Online]. Available: http://arxiv.org/abs/1906.12170
[99] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[100] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1–14.
[101] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data With Recurrent Neural Networks*. New York, NY, USA: ACM, City, 2006.
[102] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: A comparison of models and an online application," in *Proc. Interspeech*, Sep. 2018, pp. 1–8.
[103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
[104] P. Ma, S. Petridis, and M. Pantic, "End-To-end audio-visual speech recognition with conformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7613–7617.
[105] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6319–6323.
[106] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6548–6552.
[107] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7596–7599.
[108] S. Moon, S. Kim, and H. Wang, "Multimodal transfer deep learning with applications in audio-visual recognition," in *Proc. MMML Workshop Neural Inf. Process. Syst.*, 2015.
[109] K. Thangthai, R. Harvey, S. Cox, and B.-J. Theobald, "Improving lipreading performance for robust audiovisual speech recognition using DNNs," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, Sep. 2015, pp. 127–131.
[110] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2722–2726.
[111] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2304–2308.
[112] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6115–6119.
[113] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Proc. Interspeech*, Aug. 2017, pp. 3662–3666.
[114] S. NadeemHashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda, and S. Gupta, "A lip reading model using CNN with batch normalization," in *Proc. 11th Int. Conf. Contemp. Comput. (IC3)*, Aug. 2018, pp. 1–6.
[115] M. Wand, J. Schmidhuber, and N. T. Vu, "Investigations on end-to-end audiovisual fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3041–3045.
[116] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2018, doi: 10.1109/TPAMI.2018.2889052.
[117] A. Britto Mattos, D. A. B. Oliveira, and E. da silva Morais, "Improving CNN-based viseme recognition using synthetic data," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6, doi: 10.1109/ICME.2018.8486470.
[118] D. A. B. Oliveira, A. B. Mattos, and E. da Silva Morais, "Improving viseme recognition using GAN-based frontal view mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2148–2155.
[119] L. Courtney and R. Sreenivas, "Learning from videos with deep convolutional LSTM networks," 2019, *arXiv:1904.04817*. [Online]. Available: http://arxiv.org/abs/1904.04817
[120] D.-W. Jang, H.-I. Kim, C. Je, R.-H. Park, and H.-M. Park, "Lip reading using committee networks with two different types of concatenated frame images," *IEEE Access*, vol. 7, pp. 90125–90131, 2019.
[121] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, "Modality attention for end-to-end audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6565–6569.
[122] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–13.
[123] C. Wang, "Multi-grained spatio-temporal modelling for lip-reading," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–11.
[124] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 356–363.

[125] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 364–370.
[126] M. Luo, S. Yang, S. Shan, and X. Chen, "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 273–280.
[127] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual information maximization for effective lip reading," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 420–427.
[128] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2857–2866.
[129] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7608–7612.
[130] A. Botev, B. Zheng, and D. Barber, "Complementary sum sampling for likelihood approximation in large scale classification," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54. PMLR, 2017, pp. 1030–1038.
[131] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[132] L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, pp. 71–99, Jul. 1993.
[133] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous automatic speech recognition by lipreading," in *Motion-Based Recognition*. Springer, 1997, pp. 321–343.
[134] D. Howell, S. Cox, and B. Theobald, "Visual units and confusion modelling for automatic lip-reading," *Image Vis. Comput.*, vol. 51, pp. 1–12, Jul. 2016.
[135] K. Thangthai and R. Harvey, "Improving computer lipreading via DNN sequence discriminative training techniques," in *Proc. Interspeech*, Aug. 2017, pp. 1–5.
[136] K. Thangthai, H. L. Bear, and R. Harvey, "Comparing phonemes and visemes with DNN-based lipreading," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.
[137] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 129–136.
[138] D. Hu, X. Li, and X. Lu, "Temporal multimodal learning in audiovisual speech recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3574–3582.
[139] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 423–435, Mar. 2009.
[140] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 195–199.
[141] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 326–338, Mar. 2016.

SOUHEIL FENGHOUR received the M.Sci. degree in physics from Imperial College, London, U.K., in 2012. He is currently pursuing the Ph.D. degree in computer science with London South Bank University. From 2012 to 2016, he worked in various internet companies as a Data Analyst doing data mining and analytics. His research interests include lip reading, deep learning, natural language processing, computer vision, and heuristic search optimization.

DAQING CHEN received the bachelor's degree in systems engineering from Northwestern Polytechnical University, Xi'an, China, in 1982, the M.Phil. degree in automatic control engineering from the National University of Defense Technology, Changsha, China, in 1990, and the Ph.D. degree in automatic control engineering from Northwestern Polytechnical University, in 1993. From 1994 to 1997, he worked as a Postdoctoral Researcher and then an Associate Professor with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an. From 1997 to 1998, he was a Research Associate with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. From 1998 to 1999, he worked as a Research Fellow with the System, Electronics and Information Laboratory, IRESTE, University of Nantes, Nantes, France. Since 1999, he has been working with London South Bank University, where he is currently a Senior Lecturer in informatics with the School of Engineering. His research interests include deep learning algorithms with applications in lip reading, medical image diagnosis, high-dimensional data embedding and visualization, high-volume data labeling, and business intelligence.

KUN GUO received the bachelor's degree in detection, guidance and control technology and the master's degree in systems engineering from Northwestern Polytechnical University, Xi'an, China, in 2007 and 2010, respectively, and the Ph.D. degree in engineering systems and design from London South Bank University, U.K., in 2013. From 2014 to 2016, he worked as a Senior Data Analyst with ZTE, Xi'an. Since 2016, he has been working as a Senior Algorithm Engineer with Xi'an VANXUM Electronics Technology Company Ltd., China. His research interests include applications of deep learning in computer vision and natural language processing, data mining, and computer graphics compression.

BO LI received the B.E. degree in electronic information technology and the M.E. and Ph.D. degrees in systems engineering from Northwestern Polytechnical University, Xi'an, China, in 2000, 2002, and 2008, respectively. From 2014 to 2015, he was a Visitor Scholar with London South Bank University, London, U.K. He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. His current research interests include intelligent decision and control, deep reinforcement learning, and uncertain information processing.

PERRY XIAO received the bachelor's degree in opto-electronics and the master's degree in solid state physics from Jilin University of Technology, China, in 1990 and 1993, respectively, and the Ph.D. degree in photophysics from the University of Strathclyde and London South Bank University, in 1998. From 1998 to 2000, he worked as a Research Fellow with the School of Engineering, London South Bank University, where he held various posts, since 2000. He is currently the Co-Founder and the Director of Biox Systems Ltd., a successful university spin-out company that designed and manufactured-AquaFlux and Epsilon, novel instruments for water vapour flux density and permittivity imaging measurements, which have been sold to more than 200 organizations worldwide, including leading cosmetic companies, such as Unilever, L'Oreal, Philips, GSK, Johnson and Johnson, and Pfizer. His research interests include development of novel infra-red and electronic measurement technologies for biomedical applications, including skin characterization, trans-dermal drug diffusion, and medical diagnosis.