Image and Vision Computing 78 (2018) 53–72

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

Survey on automatic lip-reading in the era of deep learning☆

Adriana Fernandez-Lopez, Federico M. Sukno*

Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain

ARTICLE INFO

Article history:
Received 10 April 2018
Accepted 14 July 2018
Available online 30 July 2018

Keywords:
Automatic lip-reading
Audio-visual corpora
Visual speech decoding
Deep learning systems
Multi-view lip-reading

ABSTRACT

In the last few years, there has been an increasing interest in developing systems for Automatic Lip-Reading (ALR). Similarly to other computer vision applications, methods based on Deep Learning (DL) have become very popular and have permitted to substantially push forward the achievable performance. In this survey, we review ALR research during the last decade, highlighting the progression from approaches previous to DL (which we refer to as traditional) toward end-to-end DL architectures. We provide a comprehensive list of the audio-visual databases available for lip-reading, describing what tasks they can be used for, their popularity and their most important characteristics, such as the number of speakers, vocabulary size, recording settings and total duration. In correspondence with the shift toward DL, we show that there is a clear tendency toward large-scale datasets targeting realistic application settings and large numbers of samples per class. On the other hand, we summarize, discuss and compare the different ALR systems proposed in the last decade, separately considering traditional and DL approaches. We address a quantitative analysis of the different systems by organizing them in terms of the task that they target (e.g. recognition of letters or digits and words or sentences) and comparing their reported performance in the most commonly used datasets. As a result, we find that DL architectures perform similarly to traditional ones for simpler tasks but report significant improvements in more complex tasks, such as word or sentence recognition, with up to 40% improvement in word recognition rates. Hence, we provide a detailed description of the available ALR systems based on end-to-end DL architectures and identify a tendency to focus on the modeling of temporal context as the key to advance the field. Such modeling is dominated by recurrent neural networks due to their ability to retain context at multiple scales (e.g. short- and long-term information). In this sense, current efforts tend toward techniques that allow a more comprehensive modeling and interpretability of the retained context.

Contents

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail addresses: adriana.fernandez@upf.edu (A. Fernandez-Lopez), federico.sukno@upf.edu (F.M. Sukno).

https://doi.org/10.1016/j.imavis.2018.07.002
0262-8856/

1. Introduction

Speech is the most used communication method between humans, and it is considered a multi-sensory process that involves perception of both acoustic and visual cues. McGurk and McDonald demonstrated the influence of vision in speech perception in [1], where it was experimentally shown that when observers were presented with mismatched auditory and visual cues, they perceived a different sound from those presented in the stimulus, i.e. the syllable /ba/ was spoken over the lip movements of /ga/, and the perception was the intermediate syllable /da/. Since then, many authors have demonstrated that the use of visual information in speech recognition improves robustness [2,3].

Despite audio signals are in general much more informative than video signals, it has been demonstrated that most people use lip-reading cues to understand speech. However, these cues are often used unconsciously and to different degrees depending on aspects such as the hearing capability [4] or the acoustic conditions (e.g. the visual channel becomes more important in noisy environments) [5-8]. Furthermore, the visual channel is the only source of information for people with hearing disabilities to understand the oral language [2,9,10].

In the literature, much of the research has focused on Automatic Speech Recognition (ASR) systems, given that speech is primarily an acoustic form of communication. Nowadays, ASR systems are powerful systems able to understand the spoken language with very high recognition rates when the acoustic signal is not corrupted. However, when the acoustic signal is degraded, the performance of ASR drops and there is the need to rely also on the information provided by the visual channel. This has led to research in Audio-Visual Automatic Speech Recognition (AV-ASR) systems, which try to balance the contribution of the audio and the visual information channels to develop systems that are robust to audio artifacts and noise. AV-ASR systems have been shown to significantly improve the recognition performances of audio-based systems under adverse acoustic conditions [2,11].

On the other hand, in the last decades there has been an increased interest in decoding speech exclusively using visual cues, i.e. mimicking the human capability to perform lip-reading, leading to Automatic Lip-Reading (ALR) systems [11-20]. Nonetheless, ALR systems are still behind in performance compared to audio- or audio-visual systems. This can be partially explained by the greater challenges associated to decoding speech through the visual channel, when compared to the audio channel.

One of the main challenges in ALR systems resides on the visual ambiguities that arise at the word level due to homophemes, i.e characters that are easily confused because they produce the same or very similar lip movements (e.g. [p], [b] and [m]) [11,13,21]. Recall that the main objective of speech recognition systems is to understand verbal communication, which is structured in terms of sentences, words and characters, going from larger to smaller speech entities. More precisely, the standard minimum unit in speech processing is not the character, but the *phoneme*, defined as the minimum distinguishable sound that is able to change the meaning of a word [22]. Similarly, when analyzing visual information many researchers use the *viseme*, which is defined as the minimum distinguishable speech unit in the video domain [23], although there is no consensus on the precise definition of the different visemes nor their number, or even their actual usefulness and existence [23-26].

The fact that several phonemes produce lip movements that are visually indistinguishable implies that there is no direct or one-to-one correspondence between phonemes and visemes. For example, the phones /p/ and /b/ are visually indistinguishable because voicing occurs at the glottis, which is not visible. On the other hand, there are also phonemes whose visual appearance can change (or even disappear) depending on the context: this is the case of the velar consonants (e.g. /k/ or /g/)) which change the tongue's position in the palate depending on the previous or following phoneme [27]. For these reasons, many authors have proposed different phoneme-to-viseme mappings, with various definitions and numbers of visemes [28-33,18]. In contrast, other authors dispute the existence of visemes and defend that visual ambiguities can be completely resolved using context from neighboring characters, words or a language model [16,19,25,34]. They argue that working through visemes to understand speech is an irrecoverable loss of information. In any case, it is widely accepted that one of the most important challenges when designing ALR systems is how to make the system robust to visual ambiguities.

Other challenges associated to lip-reading include head pose variations, illumination conditions, poor temporal resolution (when compared to audio systems), efficient encoding of spatio-temporal information and speaker dependency [7,35,36]. Furthermore, human lip-readers argue that facial expressions help to decode the spoken message by adding context to the sentence. Thus, while most automatic systems focus only on the mouth region, it might be helpful to consider the whole face to decode visual speech [37].

Traditionally, ALR systems were based on the extraction of visual features and the classification and modeling of the spoken sequences. Thus, traditional ALR systems mainly consist of image transforms or appearance-based features combined with Hidden Markov Models (HMMs) that use short context information to model the temporal dynamics of the sequences. Early ALR systems addressed simple recognition tasks such as alphabet or digit recognition, but progressively shifted to more complex and realistic settings leading to several recent systems that target continuous lip-reading. To a large extent, these advances have been possible thanks to the construction of powerful systems based on Deep Learning (DL) architectures that have quickly started to replace traditional systems and to the availability of large-scale databases [16,19]. In this way, technological advances in ALR systems have made possible several novel applications such as dictating messages to smartphones in noisy environments [38,39], using visual silent passwords [40-42], discriminating between native and non-native speakers [43-45], transcribing and re-dubbing silent films [16,34], synthesizing voice for people with speech disabilities based on their lip movements [46-49], developing augmented lip views to assist people with hearing impairments [50] or resolving multi-talker simultaneous speech [51,52].

1.1. Contribution

In this survey, we review the research on ALR systems between 2007 and 2017[1], highlighting the progression from approaches previous to DL (which we refer to as traditional) toward end-to-end DL architectures. We provide a comprehensive list of the audio-visual

[1] We also include the works published so far during 2018.

databases available for lip-reading, describing what tasks they can be used for, their popularity and their most important characteristics, such as number of speakers, vocabulary size, recording settings and total duration. On the other hand, we summarize, discuss and compare the different ALR systems proposed in the last decade, separately considering traditional and DL approaches. We address a quantitative analysis of the different systems by organizing them in terms of the task that they target and comparing their reported performance in the most commonly used datasets.

While there exists another literature review on ALR in [13], it only covers papers up to 2013. The big growth of research in visual speech architectures during the last few years (see Fig. 1) has considerably expanded the literature of the field, producing a shift of the state-of-the-art toward systems based on DL architectures and justifying the need for an up-to-date review as the one presented here.

The reminder of this survey is organized as follows: in Section 2 we summarize the available corpora for lip-reading and their main characteristics, grouped by recognition task and viewing angle. In Section 3) we review the progression of ALR systems in the last decade in terms of system's architecture and performance, including i) a review of traditional architectures grouped by task and dataset, and ii) a review of recent ALR systems based on DL architectures. Conclusions are provided in Section 4.

2. Audio-visual databases

Reviewing the literature, the early databases designed to develop ALR systems, starting from the nineties, focused on specific and simple recognition tasks with restricted vocabularies, such as alphabet or digit recognition. These datasets have been widely analyzed because they allow to quickly train prototype systems given that they tackle lip-reading from well controlled settings with a pre-defined vocabulary and multiple repetitions. However, the typically low numbers of subjects and limited amount of recorded data make it difficult to construct robust ALR models that generalize well to more realistic application settings. Thus, subsequent databases focused on increasing the amount of captured data and addressing more complex tasks, going toward ALR systems targeting continuous speech.

Acquisition of large audio-visual databases is challenging due to the several factors that could be addressed (subjects, repetitions, illumination, head pose, vocabulary, resolution, etc.). Thus, some efforts were made to create datasets providing moderately large amounts of data focusing just on a few factors, while giving up other aspects. For example, the GRID corpus [53] contains a big number of utterances but very similar and constrained sentences and the RM-3000 database [54] contains only one speaker but it has a huge vocabulary. More recent efforts have led to large-scale databases collected from TV broadcasts with the objective to provide a wide vocabulary under increasingly realistic settings (LRW [19], LRS [16], MV-LRS [25]). The biggest dataset for continuous speech recognition, named LRS, consists of more than 100,000 utterances spoken by over a thousand different people. Thus, the field is growing toward large databases with a lot of variability to train robust ALR systems.

In the following subsections we compare the available databases for training ALR systems, classifying them by task (e.g. letters, digits, words and sentences) and by viewing angle. Despite audio-visual datasets have been dominated by frontal-view recordings, ALR systems should deal with multi-view lip-reading to decode speech in realistic scenarios. Table 1 provides a list of audio-visual databases for ALR with frontal-view data, while Table 3 provides a similar list for datasets captured under multiple viewpoints. For each database we summarize its key features, including year of creation; Google scholar citations; language; number of speakers; recognition task being considered; number of classes; number of utterances; resolution and total duration. In addition, representative snapshots from some of these databases are shown in Fig. 2.

2.1. Alphabet and digit recognition

Early works in ALR focused on simple recognition tasks such as alphabet or digit recognition. The available databases differ in several aspects, such as number of speakers, language, number of utterances and spatial and temporal resolutions.

For alphabet recognition, AVLetters (1998) [55] is one of the most used databases. In contains recordings from 10 speakers repeating each letter 3 times, at a resolution of 376 × 288 pixels and 25 fps. Later on, AVLetters2 [56] and AVICAR [57] solved some weaknesses of AVLetters, such as the low resolution or the limited number of speakers. Specifically, AVLetters2 increased the number of utterances (from 3 to 7 repetitions per speaker) and the resolution (1920 × 1080 pixels and 50 fps). Nonetheless, the number of speakers was reduced to just 5. On the other hand, AVICAR is a large multi-speaker database with high resolution. It contains 100 speakers, although only 86 are available.

For digit recognition, XM2VTS [58] is one of the biggest multi-speaker databases with 295 participants. It was especially designed for personal identification. Each subject was asked to pronounce two continuous digit strings and one phonetically balanced sentence. Other databases such as VALID [65] or BANCA [60] followed a similar structure to the XM2VTS database. In particular, VALID was designed for comparing speaker identification experiments under controlled and uncontrolled illumination and acoustic noise. This database includes recordings from 106 speakers in five scenarios. Similarly, the BANCA database was especially designed for identity verification under 3 different scenarios (controlled, degraded and adverse). It consists of 208 subjects covering 4 different languages (English, French, Italian

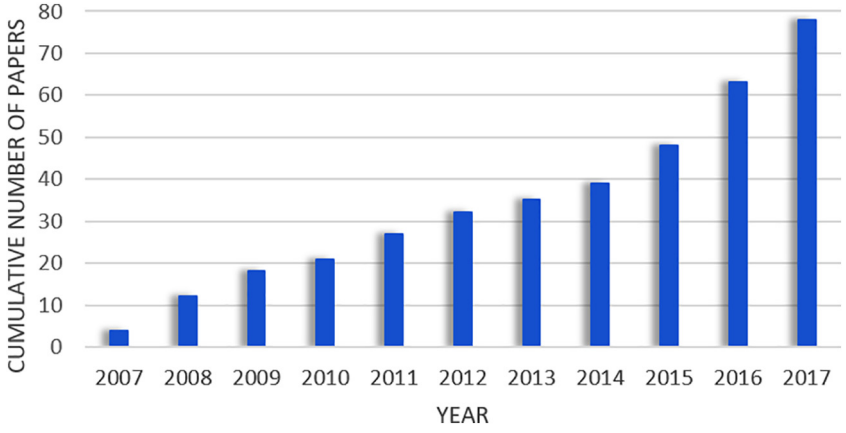

Fig. 1. Cumulative number of papers on ALR systems published between 2007 and 2017.

Table 1
Audio-visual corpora, in chronological order.

Name	Year	Cites	Language	Speakers	Task	Classes	Utterances	Resolution	Duration
AVLetters [55]	1998	455	English	10	Alphabet	26	780	376 × 288, 25 fps	13 min
XM2VTS [58]	1999	1466	English	295	Digits	10	885	720 × 576, 25 fps	59 min
IBMViaVoice [30]	2000	295	English	290	Sentences	10,500	24,325	704 × 480, 30 fps	50 h
VIDTIMIT [59]	2002	45	English	43	Sentences	346	430	512 × 384, 25 fps	30 min
BANCA [60]	2003	507	Multiple	208	Digits	10	29,952	720 × 576, 25 fps	∼14 h
IBMIH [61]	2004	37	English	79	Digits	10	16,197	720 × 480, 30 fps	N/A
AVOZES [62]	2004	55	English	20	Digits	10	200	720 × 480, 30 fps	∼2 h
					Sentences	3	60		
AV@CAR [63]	2004	26	Spanish	20	Alphabet	26	800	768 × 576, 25 fps	∼1 h
					Digits	10	600		50 min
					Sentences	250	6000		∼8 h
AVICAR [57]	2004	150	English	86	Alphabet	26	59,000	720 × 480, 30 fps	∼33 h
					Digits	13			
					Sentences	1317[a]			
CUAVE [64]	2004	248	English	36	Digits	10	7000	720 × 480, 30 fps	14 min
AV-TIMIT [29]	2004	112	English	233	Sentences	510	4660	720 × 480, 30 fps	4 h
VALID [65]	2005	33	English	106	Digits	10	1590	576 × 720, 25 fps	N/A
GRID [53]	2006	520	English	34	Phrases	51	34,000	720 × 576, 25 fps	∼28 h
IBMSR [66]	2008	15	English	38	Digits	10	1661	368 × 240, 30 fps	N/A
AVLetters2 [56]	2008	44	English	5	Alphabet	26	910	1920 × 1080, 50 fps	15 min
IV^2 [67]	2008	13	French	300	Sentences	15	4500	780 × 576, 25 fps	∼8 h
UWB-07-ICAV [68]	2008	9	Czech	50	Sentences	7550	10,000	720 × 576, 50 fps	25 h
OuluVS [69]	2009	164	English	20	Phrases	10	1000	720 × 576, 25 fps	16 min
CENSREC-1-AV [70]	2010	20	Japanese	42	Digits	10	3234	720 × 480, 30 fps	N/A
QuLips [71]	2010	11	English	2	Digits	10	3600	720 × 576, 25 fps	N/A
NDUTAVSC [72]	2010	11	German	66	Digits	6907	6907	640 × 480, 100 fps	∼11 h
					Words				
					Sentences				
WAPUSK20 [73]	2010	12	English	20	Phrases	52	2000	640 × 480, 32 fps	20 h
LILiR [74]	2010	49	English	12	Sentences	200	2400	720 × 576, 25 fps	N/A
BL [75]	2011	7	French	17	Sentences	238	4046	640 × 480, 30 fps	∼6 h
UNMC-VIER [76]	2011	5	English	123	Sentences	12	2460	708 × 640, 29 fps	N/A
MOBIO [77]	2012	128	English	150	Sentences	N/A	N/A	640 × 480, 16 fps	61 h
AGH AV [78]	2012	5	Polish	20	Digits	N/A	N/A	1920 × 1080, 50 fps	∼3 h
MIRACL-VC [79]	2014	10	English	15	Words	10	1500	640 × 480, 15 fps	N/A
					Phrases	10	1500		
AusTalk [80]	2014	6	English	1000	Digits	10	24,000	640 × 480	∼3000 h
					Words	966	966,000		
					Sentences	59	59,000		
MODALITY [81]	2015	2	English	35	Words	182	231	1920 × 1080, 100 fps	N/A
OuluVS2 [82]	2015	17	English	53	Digits	10	1590	1920 × 1080, 30 fps	∼1 h
					Phrases				∼1 h
					Sentences	530	530		13 min
RM-3000 [54]	2015	4	English	1	Sentences	1000[a]	3000	360 × 640, 60 fps	∼4 h
IBM AV-ASR [83]	2015	47	English	262	Sentences	10,400[a]	N/A	704 × 480, 30 fps	∼40 h
TCD-TIMIT [84]	2015	20	English	62	Sentences	5954	6913	1920 × 1080, 30 fps	∼6 h
HAVRUS [85]	2016	3	Russian	20	Sentences	1530	4000	640 × 460, 200 fps	N/A
LRW [19]	2016	30	English	1000+	Words	500	400,000	256 × 256, 25 fps	∼111 h
LRS [16]	2017	29	English	1000+	Sentences	17,428[a]	118,116	160 × 160, 25 fps	∼33 h
VLRF [37]	2017	1	Spanish	24	Sentences	1374[a]	10,200[a]	1280 × 720, 50 fps	∼3 h
MV-LRS [25]	2017	1	English	1000+	Sentences	14,960	74,564	160 × 160, 25 fps	∼20 h
AV Digits [86]	2018	0	English	53	Digits	10	795	1280 × 780, 30 fps	N/A
				39	Phrases		5850		

h: hours, min: minutes.
[a] Number of words.

and Spanish). There are 12 sessions per subject in which they were instructed to say a random 12 digit number, his/her name, their address and birth date (∼30,000 utterances).

However, the most popular database for training ALR systems in digit recognition is CUAVE [64] despite it contains considerably less speakers than XM2VTS and VALID. CUAVE contains 36 speakers but it provides a large number of utterances, organized in sessions of single and dual speakers. In single-speaker sessions, the speaker pronounced 50 isolated digits while standing naturally in front of the camera. After that, the speaker was captured from both profile views while uttering 20 isolated digits, and then 60 connected digits facing the camera again. For dual-speaker sessions, two speakers were recorded at the same time; while one speaker was talking the other one would remain silent, but both were captured by the camera. Speakers were asked to utter two repetitions of connected-digit sequences, alternating their turns. Subsequent datasets were presented dealing with digit recognition such as AV@CAR [63] for Spanish, AVOZES [62], AVICAR [57] and AusTalk [80] for English, the AGH AV Corpus [78] for Polish and the CENSREC-1-AV [70] for Japanese. They were recorded with moderate spatial and temporal resolutions and at least 20 speakers. Other datasets such as IBMIH [61] and IBMSR [66] were designed for digit recognition with huge numbers of speakers and utterances, but unfortunately they are not publicly available. In 2015, the multi-view OuluVS2 database [82] was presented with high resolution, 52 subjects and near 1600 utterances. More recently, in 2018 the multi-view AV Digit database [86] was presented also with high resolution, 53 subjects and close to 800 utterances of digit sequences.

Table 2
Sentence examples of audio-visual databases.

Name	Year	Language	Sentences or phrases
AVICAR	2004	English	This was easy for us. First add milk to the shredded cheese. Tofu is made from processed soybeans.
GRID	2006	English	Bin blue at A 1 again. Lay green by B 2 now. Place red in C 3 please.
OuluVS	2009	English	Excuse me. Nice to meet you. How are you.
VIDTIMIT	2009	English	She had your dark suit in greasy wash water all year. Don't ask me to carry an oily rag like that. The clumsy customer spilled some expensive perfume.
UNMC-VIER	2011	English	Joe took father's green shoe bench out. She had your dark suit in greasy wash water all year. Mum strongly dislikes appetizers.
OuluVS2	2015	English	Military personnel are expected to obey government orders. Agricultural products are unevenly distributed. Chocolate and roses never fail as a romantic gift.
TCD-TIMIT	2015	English	She had your dark suit in greasy wash water all year. The prospect of cutting back spending is an unpleasant one for any governor. Don't ask me to carry an oily rag like that.
VLRF	2017	Spanish	Eligieron una casa allí con las mismas condiciones. Los gusanos son animales invertebrados sin extremidades. A las ocho de la mañana ya estaba haciendo pasteles.
LRS	2017	English	When you're cooking chips at home. The traditional chip pan often stays on the shelf. Through what they call a knife block.

2.2. Word and sentence recognition

Datasets for digit and alphabet recognition have been very popular because they allow dealing with ALR under controlled settings with a constrained vocabulary and large numbers of instances per class. While this is useful to analyze the effectiveness of algorithms at early design stages, the resulting models tend to be of limited scope and difficult to extrapolate to more complex tasks such as word or sentence recognition. However, the aim of ASR systems is to understand natural speech, which is mainly structured is terms of sentences, which has made it necessary the acquisition of databases containing words, phrases and phonetically balanced sentences.

One of the earliest audio-visual databases containing sentences is IBMViaVoice™ [30], which consists of 290 subjects uttering continuous speech read from a script with a vocabulary size of approximately 10,500 words and 24,325 sentence utterances. Unfortunately, this corpus is not publicly available. Among the available corpora we find VIDTIMIT (2002) [59], designed to target person verification. It consists of 43 subjects reciting 10 sentences each, selected from a pool of 346 different sentences. Similarly, AV-TIMIT [29] was published in 2004 for audio-visual speech recognition. It contains 233 speakers and 510 different sentences. Other datasets already described in Section 2.1 for digit recognition also contain specific sessions with sentences: AV@CAR provides 250 phonetically balanced sentences, AVICAR sentences with more than 1300 different words, and AVOZES three different sentences designed to contain almost all phonemes and visemes of Australian English.

Several other databases were published between 2008 and 2014. Most of them were recorded in English [69,76,77,79,80,87] but we can also find two databases recorded in French [67] and one recorded in Czech [68]. Among the English-based corpora, the OuluVS database [69] is one of the most used databases for evaluating ALR systems. It contains 20 speakers uttering 10 short sentences of daily-use in English, where each utterance was repeated by the same speaker up to 5 times. The LILiR [74], MIRACL-VC [79], UNMC-VIER [76] and Austalk [80] databases contain 12, 15, 123 and 1000 speakers, respectively. However, MIRACL-VC and UNMC-VIER contain rather few sentences (10 and 12), while LILiR and Austalk contain 200 and

Table 3
Multi-view audio-visual databases, in chronological order.

Name	Year	Cites	Language	Task	Speakers	Classes	Utterances	View (°)
CUAVE [64]	2004	248	English	Digits	36	10	7000	−90, 0, 90
AVICAR [57]	2004	150	English	Sentences	100	1317 [a]	59,000	Variable (4 views)
CMU AVPFV [88]	2007	62	English	Words	10	150	15,000	0, 90
IBMSR [66]	2008	15	English	Digits	38	10	1661	−90, 0, 90
HIT-AVDB-II [89]	2008	4	Multiple (2)	Sentences	30	11	1980	0, 30, 60, 90
QuLips [71]	2010	11	English	Digits	2	10	3600	0, 10, 20,…, 90
LILiR [74]	2010	49	English	Sentences	12	200	2400	0, 30, 45, 60, 90
LTS5 [90]	2011	5	French	Digits	20	10	180	0, 30, 60, 90
OuluVS2 [82]	2015	17	English	Sentences	53	540	2120	0, 30, 45, 60, 90
TCD-TIMIT [84]	2015	20	English	Sentences	62	6913	13,826	0, 30
MV-LRS [25]	2017	1	English	Sentences	3783	14,960[a]	74,564	From 0 to 90
AV Digits [86]	2018	0	English	Digits	53	10	795	0, 45, 90
				Phrases	39		5850	

[a] Number of words.

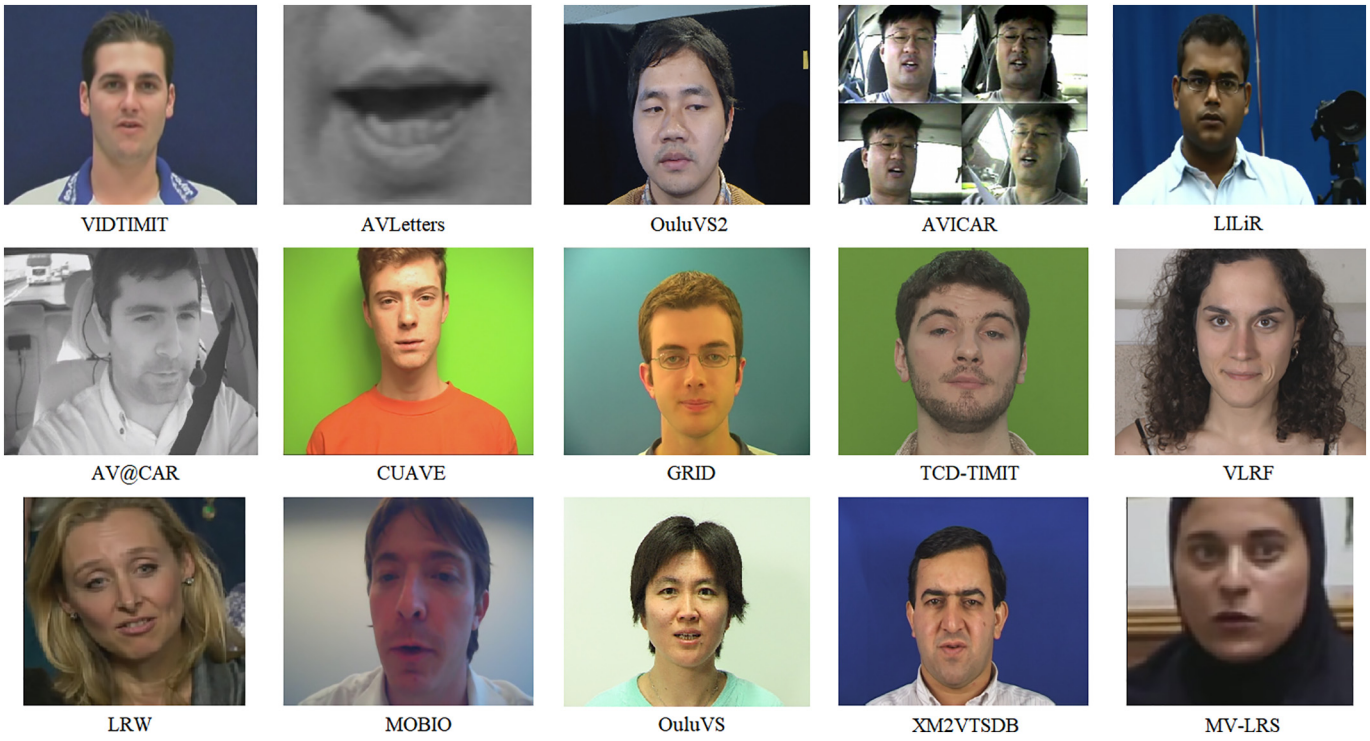

Fig. 2. Example shots of audio-visual speech databases.

59 different sentences, respectively. Yet within English corpora, we also find the MOBIO database [77]. Differently from those previously mentioned, the MOBIO database was designed for evaluating automatic face and speaker recognition on a mobile phone. It contains videos from 150 speakers answering short and free-speech questions and reading pre-defined texts, always recorded with a mobile phone held by themselves.

Audio-visual databases recorded in other languages are much less frequent than those in English. For example, in French language we find the IV^2 [67] and BL [75] databases; the first one provides a large number of speakers (300) uttering 15 sentences, while BL provides just 17 speakers but 238 sentences each. Other examples include the UWB-07-ICAVR database [68], which provides 10,000 utterances from 50 subjects in Czech, the NDUTAVSC database [72], with 66 German speakers, the AV@CAR database [63], in Spanish (already described above) and the VLRF database [37], also in Spanish, providing 1507 utterances from 24 speakers. In Table 2 we show examples of sentences of some of these AV-databases.

More recently, other databases have been published. Among them we find the single-speaker RM-3000 corpus [54] which contains a vocabulary of 1000 different words and 3000 utterances. In contrast, we find several multi-speaker databases, namely OuluVS2 [82], TCD-TIMIT [84], HAVRUS [85], IBM AV-ASR [83], VLRF [37] and AV Digits [86], which contain 53, 62, 20, 262, 24 and 53 subjects, respectively. OuluVS2 contains recordings of speakers uttering phrases and sentences; each speaker repeated three times a set of 10 daily-use phrases (similar to OuluVS) and read 10 TIMIT sentences randomly chosen from a total of 530 sentences. On the other hand, the TCD-TIMIT dataset contains more than 6900 different sentences and nearly 14,000 utterances while the HAVRUYS database [85], in Russian, provides 4000 utterances from 20 speakers. The IBM AV-ASR database is a large corpus whose sentences contain more than 10,000 words, but unfortunately it is not publicly available. The VLRF database, in Spanish, contains 24 speakers repeating up to three times sets of 25 sentences selected from a pool of 500 phonetically balanced sentences (10,000+ word utterances). Interestingly, this corpus includes participants with different hearing capabilities: 15 were normal-hearing and 9 were hearing-impaired subjects, who also performed lip-reading on the recorded videos. The transcriptions of the human lip-reading are also provided, allowing for a direct comparison between human and ALR. Finally, the very recent AV Digits database contains videos of 39 speakers uttering 10 daily-use phrases (similar to OuluVS and OuluVS2). Each phrase is repeated five times in three different speech modes: normal, whispered and silent.

Another key element to consider is the widespread use of Deep Neural Networks (DNNs) in the last few years, which has produced important advances in many aspects of computer vision, including of course lip-reading systems. While these networks have demonstrated considerable improvements on classification performance, this is only possible if appropriate data are available for training. In other words, DNNs are characterized by the need for big amounts of training data. Even though we have mentioned numerous audio-visual databases suitable for ALR, most of them do not contain a sufficient number of samples or do not cover enough vocabulary to train DNNs that generalize well. Thus, early attempts of ALR systems based on DL faced a shortage of data and, among the available corpora, those with larger number of utterances per class became more popular. For example, the GRID corpus [53] was introduced in 2006 but its use has considerably increased in the last few years. This corpus contains data collected from 34 speakers uttering 1000 constrained sentences, each fitting into a 3-second time window. Each speaker produced all combinations of "color", "digit" and "letter" by following the fixed sentence structure <command> + <color> + <preposition> + <digit> + <letter> + <adverb> It contains 34,000 utterances of very similar sentences with a vocabulary that covers 51 words. There exist also other databases that follow a similar sentence structure such as WAPUSK20 [73] or MODALITY [81]. These corpora provide rather large number of instances per class, which is adequate for training DNNs, but cannot generalize outside of the rather small set of words that they cover.

Therefore, new databases have been recently recorded with the aim of providing both large numbers of utterances and a wider vocabulary. Among these, most relevant efforts include the LRW [19],

LRS [16] and MV-LRS [25] databases. The Lip Reading Words (LRW) and Lip Reading Sentences (LRS) databases are based on recordings from BBC programs between 2010 and 2016. LRW contains sentences from more than 1000 speakers and a vocabulary of 500 words that occur at least 800 times each (~400,000 utterances in total). LRS contains 17,428 different words combined in 118,116 utterances along with the corresponding facetrack. Finally, the Multi-View-LRS (MV-LRS) database was also recorded from BBC programs but, while LRW and LRS contain only frontal face shots, MV-LRS includes shots from any viewing angle between 0 and 90°.

2.3. Multi-view databases

ALR systems have been usually based on visual speech understanding from frontal-view recordings. However, in a practical system it is not always possible to ensure that the input images will be exclusively from frontal shots. For example, in the case of imaging multiple speakers in a conversation with a single camera, we will need to work with images from different angles for each speaker. Thus, practical ALR systems should tackle multi-view lip-reading to be able to understand speech in realistic application scenarios. Furthermore, studies with human lip-readers have found that perfectly frontal shots are not necessarily the best ones to perform lip-reading. Indeed, angles slightly departing from frontal view have shown to be beneficial because lip protrusion and rounding can be better observed [87]. Then, in this section we review datasets that provide speaker recordings from different viewpoints (Table 3).

There is a considerable variability in the recording setups that have been used to capture multi-view databases for audio-visual research. Some of them contain only frontal and full-profile views, while others contain several slots between 0 and 90°. On the other hand, there are datasets which have been recorded by multiple cameras simultaneously capturing the speaker at different angles, while others have used a single camera to record different views of the speaker sequentially, at different time instants.

The AVICAR database, described in Section 2.1, was recorded in a moving automobile using an array of four cameras and eight microphones. The cameras were placed on the dashboard of the car and recorded simultaneously 4 near-frontal views of the driver. Other databases contain recordings from frontal and profile views such as the CUAVE, the CMU AVPFV [88] and the IBMSR databases. CUAVE contains single-camera recordings from people uttering sequences of digits in frontal views and in both profiles (further details in Section 2.1). In contrast, the CMU AVPFV database [88] consists of simultaneously-recorded profile and frontal views. It contains data from 10 subjects, with each subject repeating 150 possible word 10 times. Similarly, the IBMSR database, consists of recordings of three cameras simultaneously capturing frontal and two side views from 38 subjects while uttering digits sequences, but unfortunately it is not publicly available.

More recently, several databases have been presented with views between 0 and 90°. For digit recognition, we find the QuLips database [71] and the LTS5 database [90]. QuLips contains recordings from two cameras capturing each speaker while uttering sequences of digits in English (2 speakers in total). The first camera was always kept at the initial position while the subject and the second camera were allowed to rotate, so that different angles at 10° steps could be captured two at a time. In contrast, LTS5 consists of recordings of 20 native French speakers uttering digit sequences. The recordings involve one frontal camera plus one camera rotating to 30°, 60° and 90° relative to the speaker in order to obtain two simultaneous views of each sequence. For each possible position of the second camera, the speaker repeated three times the same digit sequence.

Several multi-view databases have been presented for sentence recognition in English: LILiR [74], OuluVS2 [82], TCD-TIMIT [84], MV-LRS [25], AV Digits [86] and HIT-AVDB-II [89]. Most of them have been recorded by multiple cameras, so that the different views are synchronized. For instance, LILiR contains recordings of 5 cameras located at 0°, 30°, 45°, 60° and 90° while OuluVS2 contains recordings from the same positions as LILiR but using 2 cameras with different resolution for frontal views. Similarly, TCD-TIMIT and HIT-AVDB-II contain recordings with two cameras, one fixed at frontal view and the other one fixed at 30° for TCD-TIMIT or rotating at 30°, 60° and 90° for HIT-AVDB-II. Interestingly, HIT-AVDB-II provides various types of utterances in English and Chinese. AV Digits contains high resolution recordings with three cameras, one fixed at frontal view, another one fixed at 45° and the last one fixed at full-profile view. Finally, MV-LRS is based on a selection from a wide range of BBC programs where people engage in conversations with one another, and are therefore more likely to be captured from lateral views. Thus, it contains recordings of people captured at variable views from 0 to 90°; although this dataset does not provide the viewing angle between the speaker and the camera.

3. Automatic lip-reading systems

In this section we review the research on ALR systems published between 2007 and 2017. Fig. 1 provides a quick view of the growth of the field in this period of time, by showing the cumulative number of papers that were published per year. We can observe a significant increase of the number of papers published in the last few years that, as we shall see, coincides with the growing development of DL architectures and the availability of large-scale databases.

Tables 4–6 summarize the main characteristics of the ALR systems considered in Fig. 1. Specifically, we show the publication year, number of citations, the proposed architecture (in terms of features and classifiers), the database used, the recognition task that was targeted and the accuracy that was reported. Whenever possible, we provide the accuracy in terms of Word Recognition Rates (WRR); otherwise we provide other metrics indicative of ALR performance as provided in the corresponding publications (e.g. phoneme or viseme accuracy and correctness).

An interesting aspect that emerges from the above tables is the shift of ALR systems toward architectures based on DL, which is especially noticeable in 2016 and 2017. Thus, we analyze in separate subsections the approaches previous to DL (which we refer to as traditional) and those that employ DL architectures. In all cases we focus on the aspects specific to lip-reading and skip other pre-processing stages more related to face analysis applications in general. Specifically, in Fig. 3 we show the schematic diagram of a typical ALR system, which consists of three main blocks: 1) Lips localization, 2) Extraction of visual features, and 3) Classification into sequences. The first block, focused on face detection and lips localization, will not be covered in this survey; the interested reader is referred to works on face localization and landmarking [91-100]. The goal of the feature extraction block is to parametrize the visual information observable at a given time instant or window and the classification block aims to map the visual features into speech units while incorporating temporal constraints to ensure that the decoded message is coherent. The latter provides robustness against noisy or imperfect estimates from the visual cues and helps to disambiguate between visually similar speech units. The rest of the section will focus on the last two blocks: feature extraction and classification.

We review traditional ALR systems in Section 3.1 and DL systems in Section 3.2. In both cases, we address a quantitative analysis of the different systems by organizing them in terms of the task that they target (e.g. recognition of letters or digits and words or sentences) and comparing their reported performance in the most commonly used datasets. This is important for a fair comparison, given that results are usually reported in different databases, for different recognition tasks, with a variable number of speakers, vocabularies, language and

Table 4
ALR systems from 2007 to 2017 — part I.

Year	Reference	# cit	Model		Database	Recognition task	WRR (%)
			Features	Classifier			
2007	Fu et al. [132]	21	LDG	HMM	AVICAR	Digits	37.87%
2007	Kumar et al. [88]	62	Mouth geometry	HMM	CMU AVPFV	Words	32.39%[b]
2007	Lucey et al. [108]	27	DCT + LDA	HMM	IBMSR	Digits	68.58%
2007	Marcheret et al. [133]	15	DCT + LDA + MLLT	HMM	IBMIH	Digits	63.00%
2008	Cox et al. [56]	50	Sieve + PCA	HMM	AVLetters2	Alphabet	83.00%
			AAM	HMM	AVLetters2	Alphabet	85.00%
2008	Lucey et al. [114]	14	DCT + LDA	HMM	CUAVE	Digits	53.12%
2008	Lucey et al. [66]	15	DCT + PCA	HMM	IBMSR	Digits	66.21%
2008	Pachoud et al. [115]	11	MCM-ST	Prob. seq. matching	CUAVE	Digits	80.00%
2008	Papandreou et al. [112]	3	AAM	HMM	CUAVE	Digits	75.70%
2008	Seymour et al. [9]	45	DCT	HMM	XM2VTS	Digits	87.89%
			PCA	HMM	XM2VTS	Digits	86.57%
			FDCT	HMM	XM2VTS	Digits	85.36%
			LDA	HMM	XM2VTS	Digits	86.35%
2008	Shao and Barker [134]	32	DCT	HMM	GRID	Phrases	58.40%
2008	Wang et al. [113]	20	ASM	RDA	Own data	Digits	88.32%
			ASM	HMM	Own data	Digits	91.27%
2009	Gurban and Thiran [109]	51	DCT + LDA	HMM	CUAVE	Digits	60.00%
2009	Hilder et al. [7]	17	AAM	HMM	AVLetters2	Alphabet	75.24%
2009	Kolossa et al. [120]	28	DCT	HMM	GRID	Phrases	57.00%
2009	Lan et al. [119]	45	Sieve	HMM	GRID	Phrases	40.00%
			DCT	HMM	GRID	Phrases	40.00%
			Eigenlips	HMM	GRID	Phrases	52.00%
			AAM	HMM	GRID	Phrases	65.00%
2009	Papandreou et al. [111]	66	AAM	HMM	CUAVE	Digits	83.00%
2009	Zhao et al. [69]	172	LBP-TOP	SVM	AVLetters	Alphabet	62.80%
			LBP-TOP	SVM	OuluVS	Phrases	62.40%
2010	Pass et al. [71]	11	DCT	HMM	QuLips	Digits	98.00%
2010	Saitoh and Konishi [135]	8	L2 between key points	HMM	Own data	Words	68.93%
2010	Zhou et al. [121]	11	Graph embedding		OuluVS	Phrases	90.60%[b]
2011	Cappelletta and Harte [24]	16	Optical flow	HMM	VIDTIMIT	Sentences	57.00%[a] V
			PCA	HMM	VIDTIMIT	Sentences	60.10%[a] V
2011	Navarathna et al. [136]	4	DCT + PCA	HMM	AVICAR	Digits	25.00%
2011	Ngiam et al. [137]	892	ST-PCA	Autoencoder	AVLetters	Alphabet	64.40%
2011	Ong and Bowden [123]	14	Binary feature	TGD-Boosting	OuluVS	Phrases	65.60%
2011	Ong and Bowden [124]	23	Binary feature	SP-Boosting	OuluVS	Phrases	86.20%
2011	Zhou et al. [122]	64	LBP-TOP	SVM	OuluVS	Phrases	81.30%
2012	Chiţu and Rothkrantz [4]	7	Mouth geometry	HMM	NDUTAVSC	Digits	84.24%
2012	Estellers et al. [117]	33	DCT	HMM	CUAVE	Digits	60.40%
2012	Estellers and Thiran [138]	18	DCT + LDA	HMM	Own data	Digits	71.00%
2012	Lan et al. [87]	19	AAM	HMM	LILiR	Sentences	33.00%[a] V
2012	Lan et al. [131]	19	AAM + LDA	HMM	LILiR	Sentences	14.08%
2013	Bowden et al. [107]	15	AAM	HMM	LILiR	Sentences	30.20%[b]
2013	Huang and Kingsbury [110]	69	DCT + LDA	HMM	Own data	Digits	35.20%
			DCT + LDA	DBN	Own data	Digits	35.70%
2013	Pei et al. [118]	50	RFMA		AVLetters	Alphabet	69.60%
			RFMA		AVLetters2	Alphabet	91.80%
			RFMA		OuluVS	Phrases	89.70%
2014	Bear et al. [28]	12	AAM	HMM	AVLetters	Alphabet	35.00%[a] C[b]
2014	Noda et al. [139]	23	CNN	MS-HMM	ATR	Words	37.00%
2014	Stewart et al. [39]	27	DCT	MS-HMM	XM2VTS	Digits	70.00%
2014	Zhou et al. [125]	27	Latent variables	Cross correlation	OuluVS	Phrases	74.00%
2015	Bear et al. [140]	6	AAM	HMM	AVLetters2	Alphabet	38.00%[a] C[b]
2015	Bear et al. [141]	5	AAM	HMM	LILiR	Sentences	61.80%[a] C[b]
2015	Biswas et al. [142]	5	AAM	HMM	AVICAR	Sentences	28.23%

[a] V: Viseme accuracy, P: Phoneme accuracy, C: Correctness.
[b] Speaker-dependent.

so on. Furthermore, we discuss the most popular DNN architecture for ALR systems and compare several variations that follow this baseline structure. In addition, we comment other DNNs used for lip-reading that explore alternatives from the baseline architecture and provide supplementary figures with block diagrams of the most representative end-to-end ALR systems up to 2017.

3.1. Traditional ALR systems

ALR systems start by detecting the face and extracting the region that comprises the mouth and its surrounding area. Leaving aside these pre-processing step, once the speaker's lips are located, feature extraction techniques are applied. However, for visual speech recognition, there is no consensus on which is the best feature extraction technique and there are discrepancies, for example, on whether there is more information in the position of the lips or in their movement [24,26,101]. Thus, many researchers have proposed ALR systems with different visual features based on image transforms (e.g. DCT), motion (e.g. Optical flow), geometry (e.g. width and height of the mouth) or statistical models (e.g. AAM) [13,21,102-104,105-107]. In contrast, most traditional ALR systems use HMMs to classify the visual features into speech units because they help to disambiguate between visually similar speech units while they give linguistic consistency to the output message.

Table 5
ALR systems from 2007 to 2017 — part II.

Year	Reference	# cit	Model		Database	Recognition task	WRR (%)
			Features	Classifier			
2015	Moon et al. [143]	9	DBN		AVLetters	Alphabet	55.30%
2015	Mroueh et al. [83]	39	Scattering coeffs + LDA	Feed-forward	IBM AV-ASR	Sentences	30.64%[a] P
2015	Ninomiya et al. [144]	13	DBN	MS-HMM	CENSREC-1-AV	Digits	39.30%
2015	Noda et al. [51]	72	CNN	MS-HMM	ATR	Words	22.50%
2015	Sui et al. [15]	12	DBM + DCT + LDA	HMM	AusTalk	Digits	69.10%
2015	Thangthai et al. [129]	11	AAM	CI-HMM	RM-3000	Sentences	33.32%
			AAM	CD-HMM	RM-3000	Sentences	47.48%
			AAM	Feed-forward	RM-3000	Sentences	77.49%
			HiLDA	Feed-forward	RM-3000	Sentences	84.67%
2016	Almajai et al. [18]	10	LDA	HMM	LILiR	Sentences	23.00%
			LDA + MLLT	HMM	LILiR	Sentences	25.00%
			LDA + MLLT + SAT	HMM	LILiR	Sentences	43.00%
			LDA + MLLT + SAT	Feed-forward	LILiR	Phrases	53.00%
2016	Assael et al. [34]	19	3D-CNN	Bi-GRU	GRID	Phrases	93.40%
2016	Bear and Harvey [145]	13	AAM	HMM-bigram net	LILiR	Sentences	23.00%[a] C
2016	Chung and Zisserman [146]	14	VGG-M	LSTM	OuluVS2	Phrases	31.90%
			SyncNet	LSTM	OuluVS2	Phrases	94.10%
2016	Chung and Zisserman [19]	30	CNN		LRW	Words	61.10%
			CNN		OuluVS	Phrases	91.40%
			CNN		OuluVS2	Phrases	93.20%
2016	Howell et al. [130]	4	AAM	CD-HMM	RM-3000	Sentences	75.58%
2016	Hu et al. [147]	17	RTMRBM	SVM	AVLetters	Alphabet	64.63%
			RTMRBM	SVM	AVLetters2	Alphabet	31.21%
2016	Lee et al. [128]	5	DCT + PCA	HMM	OuluVS2	Phrases	63.00%
			RAW	PLVM	OuluVS2	Phrases	73.00%
			DCT + HiLDA	HMM	OuluVS2	Phrases	74.00%
			CNN	LSTM	OuluVS2	Phrases	81.10%
2016	Petridis and Pantic [17]	18	DBNF + DCT	LSTM	AVLetters	Alphabet	58.10%
			DBNF + DCT	LSTM	OuluVS	Phrases	81.80%
2016	Rekik et al. [116]	4	HOG + MBH	SVM	CUAVE	Digits	70.10%
			HOG + MBH	K-NN	MIRACL-VC	Phrases	58.10%
			HOG + MBH	SVM	OuluVS	Phrases	68.30%
			HOG + MBH	HMM	MIRACL-VC	Phrases	69.60%
			HOG + MBH	SVM	MIRACL-VC	Phrases	79.20%
2016	Saitoh et al. [148]	5	CFI + NIN		OuluVS2	Phrases	81.10%
			CFI + AlexNet		OuluVS2	Phrases	82.80%
			CFI + GoogLeNet		OuluVS2	Phrases	85.60%
2016	Takashima et al. [149]	4	CBN	HMM	ATR	Words	51.00%
2016	Wand et al. [20]	35	Eigenlips	SVM	GRID	Phrases	69.50%[b]
			HOG	SVM	GRID	Phrases	71.20%[b]
			Feed-forward	LSTM	GRID	Phrases	79.50%[b]
2016	Wu et al. [127]	3	SDF + STLF	SVM	OuluVS2	Phrases	87.55%
2016	Zimmermann et al. [150]	4	PCA_{NN} + LSTM	HMM	OuluVS2	Phrases	73.00%
2017	Bear and Harvey [151]	1	AMM	HMM	AVLetters2	Alphabet	36.53%[a] C[b]
			AMM	HMM	LILiR	Sentences	41.53%[a] C[b]
2017	Chung and Zisserman [25]	1	CNN	LSTM + attention	OuluVS2	Phrases	91.10%
			CNN	LSTM + attention	MV-LRS	Sentences	43.60%
2017	Chung et al. [16]	39	CNN	LSTM + attention	LRW	Words	76.20%
			CNN	LSTM + attention	GRID	Phrases	97.00%
			CNN	LSTM + attention	LRS	Sentences	49.80%
2017	Fernandez-Lopez et al. [37]	1	DCT + SIFT + LDA	HMM	VLRF	Sentences	20.00%
2017	Fernandez-Lopez and Sukno [31]	2	DCT + SIFT + LDA	HMM	AV@CAR	Sentences	23.00%
2017	Petridis et al. [152]	8	Autoencoder	LSTM	OuluVS2	Phrases	84.50%
2017	Petridis et al. [153]	0	Autoencoder	Bi-LSTM	OuluVS2	Phrases	91.80%
2017	Petridis et al. [154]	0	Autoencoder	Bi-LSTM	OuluVS2	Phrases	94.70%

[a] V: Viseme accuracy, P: Phoneme accuracy, C: Correctness.
[b] Speaker-dependent.

3.1.1. Digit and letter recognition

There are 23 ALR architectures targeting digit or alphabet recognition since 2007. Looking at Tables 4–6 we observe that most traditional systems use feature techniques based on image transforms [9,66,108-110] or shape and appearance models [7,56,111-113]. In Fig. 4 we show i) the number of times that each feature technique has been integrated into ALR systems addressing digit or letter recognition; ii) the same for each classification method. On the left-side of the figure, we observe that the most used visual features have been AAMs, DCT or combinations of DCT with other transforms such as LDA or PCA. On the other hand, in the right-side of the figure, a single HMM for each digit or letter is the most used classification method, being also the most used in audio speech recognition. Other methods such as Support Vector Machines (SVMs) or Regularized Discriminant Analysis (RDA) have less been frequently explored.

Given the variety of methods addressing digit or letter recognition, it is interesting to compare them in terms of performance. This can be directly done by comparing the methods evaluated in the same databases. Thus, we will compare the methods evaluated in the most commonly used databases for digit or alphabet recognition, which are CUAVE, XM2VTS or AVLetters2.

Architectures presented in [114-117,112,109,111] have been evaluated using the CUAVE database. These methods reported WRR between 53.12% and 83.00%. For the 5 architectures using HMMs

Table 6
ALR systems from 2007 to 2017 — part III.

Year	Reference	# cit	Model		Database	Recognition task	WRR (%)
			Features	Classifier			
2017	Rahmani and Almasganj [155]	0	ASM	HMM	CUAVE	Digits	56.30% [a] P
			DBNF	HMM	CUAVE	Digits	63.40% [a] P
			ASM	DNN-HMM	CUAVE	Digits	58.90% [a] P
			DBNF	DNN-HMM	CUAVE	Digits	64.90% [a] P
2017	Stafylakis and Tzimiropoulos [156]	4	3D-CNN + ResNet	Bi-LSTM	LRW	Words	83.00%
2017	Sterpu and Harte [157]	0	DCT	HMM	TCD-TIMIT	Sentences	31.59% [a] V [b]
2017	Sui et al. [126]	1	CHAVF	SVM	OuluVS	Phrases	68.90% [b]
			CHAVF	HMM	AusTalk	Digits	69.18%
2017	Thangthai and Harvey [158]	2	PCA + LDA + MLLT	DNN-HMM	TCD-TIMIT	Sentences	43.61%
2017	Thangthai et al. [159]	0	Eigenlips	DNN-HMM	TCD-TIMIT	Sentences	42.97%
2017	Wand and Schmidhuber [160]	2	Feed-forward	LSTM	GRID	Phrases	42.40%
2018	Afouras et al. [161]	0	3D-CNN + ResNet	Bi-LSTM + LM	LRS	Sentences	37.80%
				Depthwise CNN			45.00%
				Attention encoder + LM			50.00%
2018	Fung and Mak [162]	0	3D-CNN	Bi-LSTM	OuluVS2	Phrases	87.60%
2018	Petridis et al. [163]	3	3D-CNN + ResNet	Bi-GRU	LRW	Words	82.00% [b]
2018	Petridis et al. [86]	0	Autoencoder	Bi-LSTM	AV Digits	Phrases	69.70%
						Digits	68.00%
2018	Wand et al. [164]	0	Feed-forward	LSTM	GRID	Phrases	84.70%
2018	Xu et al. [165]	1	3D-CNN + highway	Bi-GRU + attention	GRID	Phrases	97.10%

[a] V: Viseme accuracy, P: Phoneme accuracy, C: Correctness.
[b] Speaker-dependent.

as classification method, two of them used DCT [114] and LDA [109] features, reporting 53.12% WRR and 60.00% WRR, respectively. Similarly, the system presented by Estellers et al. [117] used DCT features and obtained 60.40% WRR. In contrast, both architectures presented by Papandreou et al. [111,112] used AMM models and reported 75.70% WRR and 83.00% WRR, respectively. The latter is the best WRR reported in this database. Nevertheless, the ALR system proposed by Pachoud et al. [115] based on probabilistic sequence matching classification of macro-cuboids using spatio-temporal SIFT descriptors and local displacements (named MCM-ST features) reported a similar performance (80% WRR). Finally, there is an ALR system presented in 2016 by Rekik et al. [116] that used a combination of Histogram of Oriented Gradients (HOG) and Motion Boundary Histograms (MBH) features and SVM classifiers reporting a performance of 70.10% WRR.

For the XM2VTS database, Seymour et al. [9] presented experiments comparing different image transforms (DCT, PCA, LDA, and FDCT) combined with HMMs and obtained WRR between 85.36% and 87.89%. On the other hand, the ALR system presented by Stewart et al. [39] presented a conventional system based on DCT features and HMMs, reporting 70.00% WRR. The best-performing architecture for XM2VTS used DCT features and HMMs classifiers and reported 87.89% WRR [9].

Finally, for alphabet recognition, AVLetters2 has been one of the most used databases. Several traditional architectures have been proposed with WRR up to 91,80% [7,56,118]. For the HMM-based systems, feature extraction techniques such as Sieve filters combined with PCA [56] and AAM [7,56] have been used. However, the best WRR was reported by the system presented by Pei et al. [118] that consists of an end-to-end system based on Random Forest Manifold Alignment (RFMA), which obtained 91,80% WRR followed by the 75,24% WRR obtained by Hilder et al. [7].

Therefore, even though DCT has been the most implemented feature in ALR systems tackling digit or alphabet recognition, AMM features in combination with HMMs have produced the highest reported WRR.

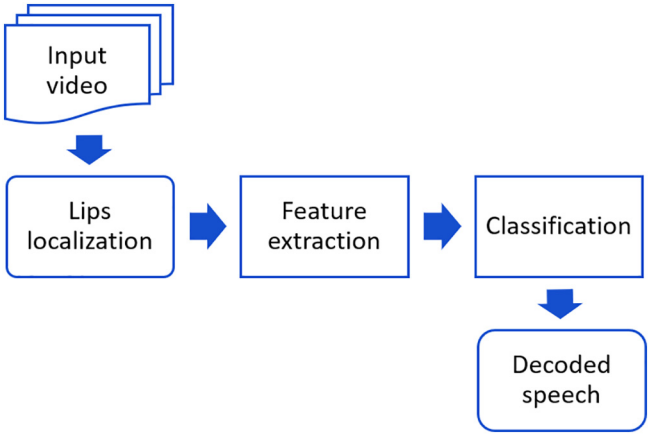

Fig. 3. The main processing blocks of an ALR system.

3.1.2. *Word and sentence recognition*

Digit and letter recognition have been very popular, but the resulting models cannot be extrapolated to more complex tasks such as word or sentence recognition and hence are of limited applicability. In Fig. 5 we show the number of ALR architectures targeting *digit or alphabet* and *word or sentence* recognition from 2007 to 2017. In the figure, we can observe a clear tendency from early systems trying to solve easier recognition tasks in controlled vocabularies (e.g. digits) toward systems dealing with more complex tasks such as word or sentence recognition. In this section we compare the 33 traditional systems presented in Tables 4–6 that target word or sentence lip-reading. Similarly to Section 3.1.1, we firstly explain the architecture's components and then compare systems in terms of performance.

In Figs. 6 and 7 we show, respectively, the number of times that each feature or classification technique has been integrated into ALR systems targeting word or sentence recognition. In Fig. 6 we observe that the most used visual features are similar to those used in digit or alphabet recognition, namely PCA, DCT, and AAM. Notice that even though these features do not have the highest usage frequencies by themselves, they appear multiple times combined with others. Compared to digit or letter recognition there is a bigger pull of features,

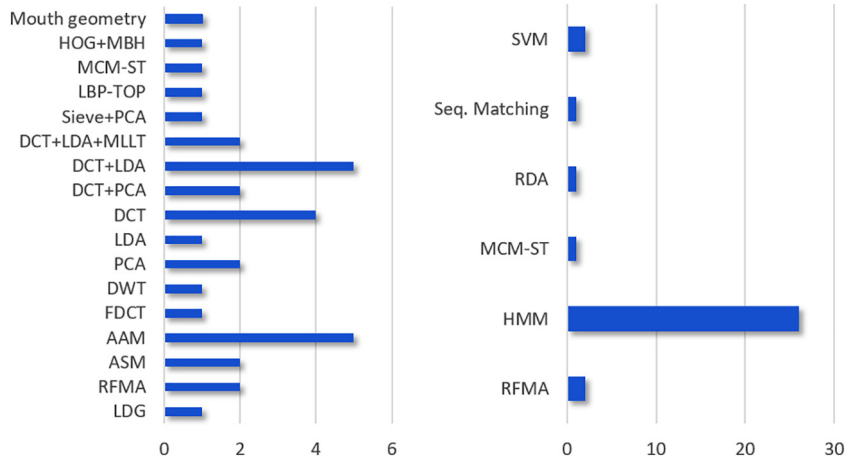

Fig. 4. Digit and alphabet recognition. Left-side: number of times that each feature technique has been used from 2007 to 2017; right-side: number of times that each classification method has been used from 2007 to 2017.

e.g. Local Binary Patterns extracted from Three Orthogonal Planes (LBP-TOP), Shape Difference Feature (SDF) or Spatio-Temporal Lip Feature (STLF). In terms of classifiers (Fig. 7), we also observe a similar tendency to digit or letter recognition, where HMMs are the most used classification method, although there is also an increment of systems using alternative classifiers, especially SVMs.

In terms of performance evaluation, the most used databases for word or sentence recognition have been GRID, OuluVS, OuluVS2 and RM-3000.

For the GRID corpus, Lan et al. [119] used a subset of 15 speakers and centered their experiments in comparing different features such as DCT, Sieve, PCA and AAM. They used one HMM per word for decoding the message, 52 HMMs in total (51 words plus silence). They obtained WRR between 40.00% and 65.00%, being AAM the most successful feature. In contrast, Kolossa et al. [120] proposed a similar model composed of DCT features and one HMM per word and reported 57.00% WRR in experiments using the full set of speakers. More recently, Wand et al. [20] compared PCA and HOG using SVM as classifier. They obtained WRR of 69.50% for PCA features and 71.20% for HOG on speaker-dependent experiments over a subset of 20 subjects. Speaker-dependent experiments means that training and testing data for the classifiers are always taken from the same speaker and the results are averaged over all the speakers.

For the OuluVS database, 9 different architectures have been presented [69,116,118,121-126]. For the ALR systems evaluated in this database, a varied set of features has been used, but most works used SVMs as classifiers. Rekik et al. [116] used a combination of spatio-temporal HOG and MBH features with SVMs and obtained WRR of 68.30%. Sui et al. [126] presented a feature extraction technique named Cascade Hybrid Appearance Visual Feature (CHAVF), which is based on LBP-TOP and DCT features and combined them with SVMs, achieving WRR of 68.90% for speaker-dependent experiments. In contrast, both Zhao et al. [69] and Zhou et al. [122] used LBP-TOP features combined with SVMs and reported 62.40% and 81.30% WRR, respectively. This big difference (~20%) are because Zhou et al. [122] introduced a process of curve matching that normalizes the video signal by mapping the original video onto a curve which is then re-sampled to produce video sequences with the same number of frames. In contrast, Ong and Bowden [123,124] proposed two systems based on binary features combined with Temporal Gradient Descend Boosting (TGD-Boosting) [123] or with Sequential Pattern Boosting (SP-Boosting) classifiers [124], reporting 65.60% and 86.20% WRR, respectively. Pei et al. [118] presented an end-to-end system based on RFMA and reported 89.70% WRR, which is the highest performance achieved so far in this database. Other alternative systems were presented by Zhou et al. [121,125]. The first one [121] uses graph embedding to capture video dynamics and the second one [125] used latent variable (LV) models to generate the representation of a sequence of images. For leave-one-utterance-out cross-validation in [121] they obtained 90.60% WRR, while for leave-one-speaker-out cross-validation in [125] they obtained 74.00% WRR.

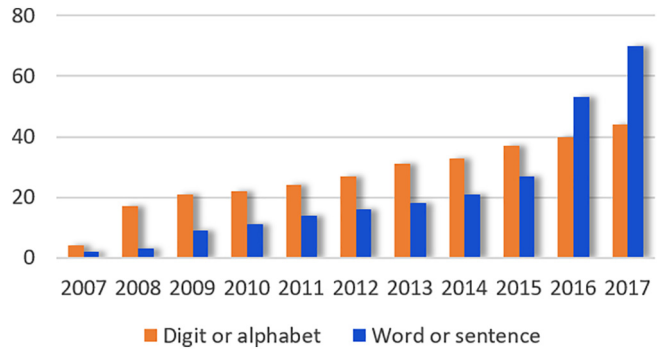

Fig. 5. Cumulative number of ALR systems targeting *digit or alphabet* and *word or sentence* recognition from 2007 to 2017.

For the OuluVS2 database, Wu et al. [127] presented a feature extraction technique based on SDF and STLF features and SVM classifiers to decode the spoken message, obtaining 55.00% WRR. In contrast, Lee et al. [128] presented three different systems. HMM-based systems were based on DCT-PCA and DCT-HiLDA features and reported 63.00% and 74.00% WRR, respectively, while the third system was based on LV models combined with raw pixel values as features and reported 73.00% WRR.

For the single-speaker RM-3000 dataset with 1000 different words, Thangthai et al. [129] and Howell et al. [130] proposed similar ALR systems using AAM features and HMM classifiers. Thangthai et al. [129] trained Context-Independent HMMs (CI-HMM) and Context-Dependent HMMs (CD-HMM). Instead of directly constructing word models they defined phoneme models. Then, they joined the corresponding phonemes of each word to form word models (model of models). The CI-HMM consisted of monophone models with 3 states per phoneme (45 phonemes in English), while the CD-HMM models distinguished between phonemes with different previous and posterior phonemes. They obtained 33.32% WRR for CI-HMMs and 47.48%

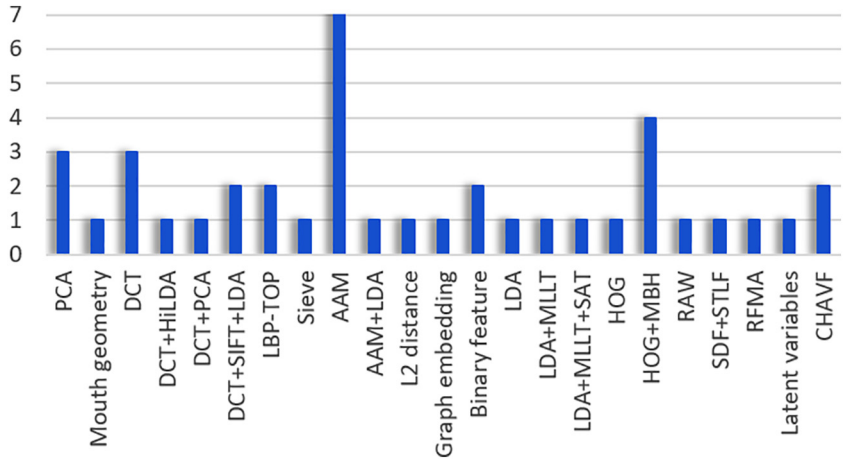

Fig. 6. Word and sentence recognition. Number of times that each feature technique has been used from 2007 to 2017.

WRR for CD-HMMs. Similarly, Howell et al. [130] presented an ALR system based on AAMs and triphoneme word decoders, and reported a WRR of 75.58%. As we can observe, for databases covering large vocabularies it seems useful to train phoneme or triphoneme models instead of just training words, because this increases the number of samples per class available for training.

For the LILiR database, Bowden et al. [107] proposed a system based on the combination of AAM features and HMM classifiers and obtained 30.20% WRR for one-speaker experiments. Lan et al. [131] used Fisher phoneme-to-viseme mapping [23] and proposed an ALR system that combines AMM + LDA features with HMMs trained on viseme classes, obtaining 14.08% WRR. Almajai et al. [18] also used Fisher phoneme-to-viseme mapping and proposed several CI-HMM and CD-HMM systems. Specifically, they proposed a CI-HMM based on monophone and monoviseme models using first- and second-order derivative features and CD-HMMs based on triphone and triviseme models with LDA, LDA + MLLT and LDA + MLLT + SAT features. In their experiments, they found that when phoneme models are used instead of viseme models, the WRR increases significantly, up to 8%, reaching up to 43.00% WRR for the whole database. Interestingly, the opposite result was reported in [31] for the Spanish database AV@CAR, where a phoneme-to-viseme mapping with an appropriate vocabulary length provided the highest WRR. Thus, there is not a general consensus on whether using visemes is advantageous or disadvantageous for ALR.

Summarizing the systems targeting word or sentence recognition, we have seen that different architectures have been evaluated for each database, both in terms of features and classifiers. In contrast to the case of digit and letter recognition systems, the disparity of features evaluated in each database makes it difficult to conclude which might be the best-performing ones. Something similar occurs in terms of classifiers: HMMs reported the best performance for the GRID database, SVMs for the OuluVS database and LV models for the OuluVS2 database. However, no system based on HMMs or LV models was tested in the OuluVS dataset and, although some HMM systems were used for OuluVS2, their features did not match those from the best-performing system. Thus, it is difficult to produce a fair comparison beyond the frequency with which the different features and classifiers have been used.

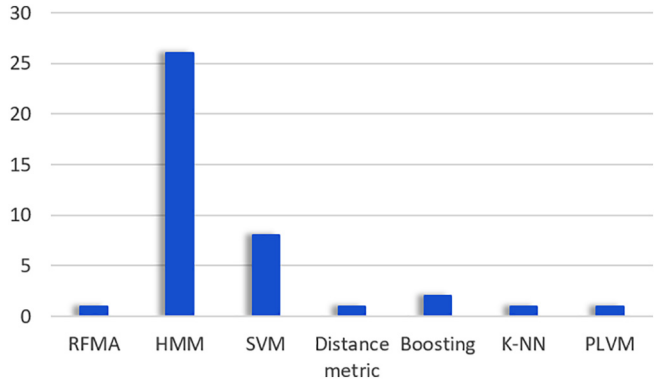

Fig. 7. Word and sentence recognition. Number of times that each classification method has been used from 2007 to 2017.

3.2. DNN-based ALR systems

While there is an extensive literature dedicated to hand-crafted methods (Section 3.1), there has been a significant improvement in the performance of ALR systems in the last years thanks to the advances in deep neural networks and the availability of large-scale databases.

There is a strong parallelism in the way that DNNs have been adopted by audio-based and video-based speech recognition systems. Initially, hybrid ASR systems combining traditional blocks with DNNs were proposed. More precisely, neural networks were first considered as feature extractors, mainly in combination with HMM-based classifiers. Afterwards, recurrent networks, e.g. Long-Short Term Memory (LSTM) networks [166], were introduced as a suitable replacement for HMMs. More recently, end-to-end DNNs have been used to fully replace all building blocks of ASR systems by neural networks, achieving considerably higher performance than traditional systems [167-169].

A similar progression is observed for video-based systems. In Tables 4–6 we see that hybrid ALR systems, firstly proposed in 2011, consist of combinations of traditional features or classifiers with neural networks [17,15,137,139,51]. In subsequent years, there has been a tendency toward ALR systems based purely on DL, known as end-to-end DNN architectures.

In this section, the DNN-based systems presented in Tables 4–6 are analyzed. Similarly to Section 3.1, we firstly explain the architectures' components and then compare the different systems in terms of performance.

3.2.1. Configuration of DNN architectures

ALR systems based on end-to-end DNNs follow a similar pipeline to traditional ones (shown in Section 3.1-Fig. 3). Similarly to the previous section, we will compare systems in terms of feature extraction

and classification stages. Block diagrams of the most representative end-to-end ALR systems up to 2017 are provided in the Supplementary materials.

We start by showing in Figs. 8 and 9 how frequently the different types of DNNs have been integrated into ALR systems as a feature or classification technique. In Fig. 8 we observe that Convolution Neural Networks (CNN) have been the most used networks to extract features, but other DNNs such as Feed-forward networks or Deep Belief Networks (DBN) have also been used. In terms of classifiers, in Fig. 9 we can see a predominance of LSTMs, although CNNs, Feed-forward DNNs and DBNs have also been used.

Looking at Tables 4–6 we observe that there are 24 end-to-end DL architectures, from which 11 consist of combinations of CNNs and RNNs (LSTMs or GRUs). Thus, this combination stands out as the most used DL architecture for ALR and we will analyze it in more detail. In Fig. 10 we show a CNN-LSTM baseline system where a sequence of video frames are processed by a convolutional network followed by a recurrent network. CNNs have been established as a powerful model to extract visual features for image recognition and classification tasks [170,171] and consist of alternating convolutional layers and pooling layers. The convolutional layers compute the inner product between linear filter and the receptive field and then they are followed by a non-linear activation function (e.g. sigmoid, tanh, ReLu). On the other hand, LSTMs are recurrent neural networks (RNN) useful for modeling sequences due to their cyclic connections that form a temporal memory [172,173]. LSTMs have been widely used because they solve the vanishing and exploding gradient problem [174] that appears in conventional RNNs. In contrast to RNNs, LSTMs have a cell unit that is regulated by 3 gates, known as input, output and forget gates, which use additive and multiplicative connections to ensure constant error flow, thus retaining short- and long-context information.

3.2.2. Architectures based on CNNs and LSTMs

Several authors have proposed CNN-LSTM networks that follow the baseline in Fig. 10. For instance, Chung and Zisserman [146] proposed a network that performs sentence-level classification. Notice that "sentence-level classification" means that the system's output is restricted to a finite number of possible sentences, which therefore act as the classes of a classification problem. The architecture inputs gray-scale images into a convolutional network, named SyncNet, which consists of five convolutional layers followed by two fully connected layers. For each frame, the output of the last CNN layer is the input to a single LSTM layer that accumulates the contribution of each frame and returns the estimated class at the end of the sequence. The block diagram of this architecture is provided in Suppl. Fig. S1. Still within the same work [146], Chung et al. compare the proposed CNN with a pre-trained network, known as VGG-M

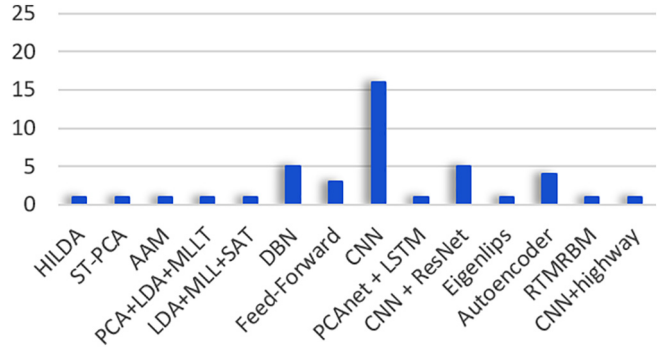

Fig. 8. DNN-based systems. Number of times that each feature technique has been used from 2007 to 2018.

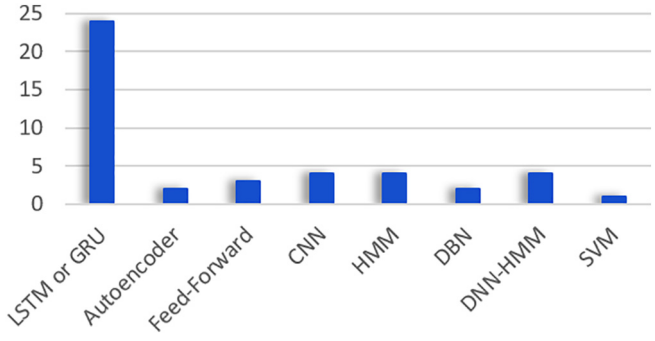

Fig. 9. DNN-based systems. Number of times that each classification method has been used from 2007 to 2018.

(Suppl. Fig. S2). VGG-M consists of five convolutional layers followed by three fully connected layers pre-trained in the ImageNet database [170]. The VGG-M output is the input to a single LSTM layer that performs the classification at the end of the sequence, similarly to SyncNet. As we will see in Section 3.2.4, in spite of having an additional fully connected layer, the pre-trained VGG-M did not perform as good as SyncNet given that the training of the latter was much more specific to the lip-reading task.

Lee et al. [128] proposed a DNN architecture that performs sentence-level classification (Suppl. Fig. S3). Their system inputs RGB normalized images that are processed by a CNN with two convolutional layers and one fully connected layer. They also define a temporal model based on two LSTM layers that receive the CNN features and accumulate the contribution of each frame until the end of the sequence, which is finally processed by a fully connected layer that returns the classification of the whole sequence into a phrase.

Assael et al. [34] proposed LIPNET, an end-to-end DL architecture that also performs sentence-level classification (Suppl. Fig. S4). The model's input is a fixed-length sequence of RGB normalized images that are processed by three spatio-temporal convolutional layers. The output features of the CNN are fed to two Bidirectional Gated Recurrent Network (GRU) layers that are finally followed by a linear transformation at each time step and a softmax over the vocabulary (which in this case is a character-based representation). This end-to-end model is trained with a Connectionist Temporal Classification (CTC) [167] network that has a softmax output layer with as many units as the number of labels in the vocabulary plus one unit for the blank character "␣". The CTC computes the probability of all possible combinations of a string. For example, if the sequence length is fixed to 3, the CTC defines the probability of a string "*am*" as $p(aam) + p(amm) + p(\text{␣}am) + p(a\text{␣}m) + p(am\text{␣})$. The model predicts frame labels and finds the optimal alignment between the predictions and the output sequence (which is a full sentence within the possible pre-defined classes).

On the other hand, Stafylakis and Tzimiropoulos [156] proposed a system that performs word-level classification (Suppl. Fig. S5). In their model, the inputs are video sequences of gray-scale normalized images, with a fixed duration of 1 s. The proposed architecture is based on a spatio-temporal convolutional layer followed by a residual network (ResNet [175]). The residual network consists of 34 layers (including convolutional, pooling and fully connected layers) that progressively reduce the spatial dimensionality with max pooling layers, until the output becomes a single dimensional vector per time step. Then, these vectors are used as input features to two bidirectional LSTMs (Bi-LSTM) [173] (two in each direction) which are concatenated at each time step for classification. Differently from previous works, the classification is not performed at the last time step of the LSTM output, once all the sequence has been encoded by

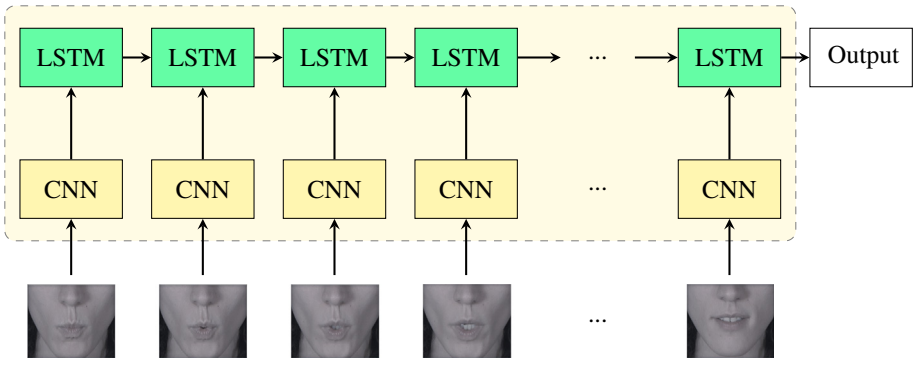

Fig. 10. Baseline DL architecture for lip-reading, consisting of combinations of CNNs and LSTMs.

the LSTM, but the softmax is applied at each time step. Hence, the overall loss is defined as the aggregated loss over all time steps.

Notice that these two last systems [34,156] used Bi-LSTMs or Bi-GRUs for their ability to produce outputs conditioned on past and future contexts, as opposed to the standard LSTMs that work only in one direction. Other very recent works have also explored the use of these bidirectional networks. On one hand, Petridis et al. [163] proposed a model very similar to [156], where the main difference between both lip-reading architectures is that [163] used Bi-GRU networks with a bigger number hidden units instead of the Bi-LSTMs networks used in [156]. On the other hand, Fung and Mak [162] used Bi-LSTMs for sentence-level classification. Their network consists of 8 spatio-temporal convolutional layers followed by a maxout activation function without pooling layer that is fed to the Bi-LSTM layer. The final output is obtained with a softmax layer at the last time step of the sequence.

Chung et al. proposed a system for AV-ASR [16] and another one for ALR [25] (Suppl. Figs. S6–S8). For the AV-ASR system, they proposed an end-to-end network based on four main modules, named *Watch*, *Listen*, *Attend* and *Spell*, that learned to predict characters from spoken sentences. The *Watch* module receives video input and consists of five 3D convolutional layers followed by one fully connected layer and then three LSTM layers stacked one behind the other to catch different levels of abstraction. A similar network is employed for *Listen* to process audio. The *Spell* module consists of three LSTMs, two attention mechanisms (for the audio and visual contexts provided by *Watch* and *Listen*) and a multi-layer perceptron (MLP). Thus, *Spell* LSTMs use: the previous character, the previous LSTM state and the concatenation of the last time step of *Watch* and *Listen* LSTMs. Next, two context vectors are computed in the *Attend* module, from audio and visual contexts. These context vectors are computed at each time step by the attention mechanisms. The attention mechanisms use the output produced by the *Watch* or *Listen* LSTMs at each time step and the current outputs of *Spell* LSTMs. Finally, the probability distribution of the output character is generated by the MLP with a softmax layer over the output. The authors emphasize that gray-scale image sequences are processed in reverse time-order, as this was found to improve results. They also explain that attention is crucial for the system because without it the model forgets the input signal, and produces an output sequence that does not correlate with the input beyond the first word or two (which the model gets correct, as these are the last words to be seen by the encoder). In addition, unidirectional encoders for the *Watch* and *Listen* modules were compared with bidirectional encoders, but the latter networks took significantly longer to train, while providing no obvious performance improvement. For the ALR system proposed in [25], where audio information is not available, the same architecture was proposed except that there were no audio attention nor *Listen* blocks.

As the last example of the CNN-LSTM architecture, Xu et al. [165] presented a network named LCANet that performs character-level classification. The video encoder of LCANet has three components: 3D convolutions, a highway network, and Bi-GRU networks. LCANet feeds 3 consecutive frames into a 3D convolutional neural network to encode both visual and short temporal information. Then, they stack two layers of highway networks [176] on top of the 3D-CNN. The highway network module has a pair of transform gate and carries a gate that allows the deep neural network to carry some input information directly to the output. These networks have been enabled to encode much richer semantic features. At the end of the video encoding, Bi-GRU networks are feed after the highway networks to encode long-term temporal information. To capture information explicitly from longer context, LCANet feeds the encoded spatio-temporal features into a cascaded attention-CTC decoder. Attention mechanism debilitates the constraint of the conditional independence assumption in CTC loss, but it improves the modeling capability on the lip-reading problem and can give better predictions on visually similar visemes.

3.2.3. Other DL architectures

Some authors have also proposed end-to-end architectures that do not follow the CNN-LSTM baseline from Fig. 10. For instance, Wand et al. proposed three DNN architectures [20,160,164] that perform word-level classification. The system proposed in [20] (Suppl. Fig. S9) consists of one feed-forward layer followed by two LSTMs and a softmax layer to perform classification within a set of pre-defined classes. Similarly, the system proposed in [160] (Suppl. Fig. S10) consists of three feed-forward layers followed by one LSTM layer and a softmax layer to perform classification within the set of words. In order to mitigate the discrepancy between known and unknown speakers, it incorporates domain adversarial training, by means of an intermediate layer driven to learn a domain-agnostic representation of the input data. Specifically, at the second feed-forward layer, a supplementary network consisting of two feed-forward layers and a softmax layer is integrated to perform speaker classification. The incorporation of the adversarial network is supposed to be beneficial because by feeding its inverted gradient into the main network, the system is prevented from learning speaker-dependent features. Finally, the system proposed in [164] consists of three feed-forward layers followed by one LSTM layer and a softmax layer that performs word classification at the end of the sequence. In this architecture all layers, including the LSTM, have the same number of neurons.

Chung and Zisserman [19] also proposed a DNN architecture that performs word-level classification (Suppl. Fig. S11). The method preprocesses each input frame with a first convolutional layer whose outputs are concatenated so that the whole sequence is sent to a second convolutional layer. The output of the second layer is fed into

the following layers, which have a similar structure to VGG-M: three additional convolutional layers, three fully connected layers and one softmax layer.

Saitoh et al. [148] proposed an end-to-end system for sentence-level classification that instead of processing the sequence frame by frame, constructs a macro image by concatenating a subset of the whole video sequence, which they call concatenated frame image (CFI). They testes the CFI in combination with three pre-trained CNNs: Networks in Networks (NIN) [177], AlexNet [170] and GoogLeNet [171]. NIN is a novel network that replaces the usual linear convolutional layers by MLP-Convolutional layers (mlpconv). Specifically, Saitoh et al. used four mlpconv followed by a spatial max pooling layer. AlexNet consists of five convolutional layers followed by three fully connected layers, and GoogleLeNet is a twenty-two layer deep network that uses a sparsely connected architecture (inception modules) to avoid computational bottlenecks. Despite the different architectures of the three networks, their performance in the ALR tests reported by Saitoh et al. [148] were fairly similar, with differences that did not exceed 5% WRR between them.

Petridis et al. [152,153,154,86] proposed four end-to-end systems for sentence-level classification. Firstly in [152] (Suppl. Fig. S12), they proposed a system based on two independent streams; the first one extracts features directly from single-images, while the second one extracts features from the difference between two consecutive frames. Both streams follow a bottleneck architecture with three hidden layers and one linear layer. At the end of the bottleneck architecture, the first and second derivatives are computed and appended to the bottleneck layer. The output of the bottleneck network of each stream is fed into an LSTM layer. Finally, the LSTM outputs of both streams are concatenated and fed into a Bi-LSTM in order to fuse their information. The output layer is a softmax layer that performs the classification using the last time step of the Bi-LSTM output, once all the sequence has been encoded. On the other hand, the system proposed by Petridis et al. in [153] is a very similar network that also incorporates audio input. Specifically, the frame difference data is replaced by audio features, so that one stream per modality is used. They also replace the LSTM networks at the end of each stream by Bi-LSTMs. The third system presented by Petridis et al. [154] tackled multi-view lip-reading for sentence-level classification. It consists of three identical streams which extract features from three images captured from different view angles. The streams follow the same architecture from [153] and their outputs are concatenated and fed into a Bi-LSTM and a softmax layer that performs the classification similarly to the two architectures previously described. Finally, the fourth system [86] was proposed as a modification of [153]. The key difference is that the new system used only a single stream (corresponding to the video frames) instead of the use of two streams proposed previously.

A transfer DL framework was presented by Moon et al. [143] for alphabet recognition. The system uses audio and visual information independently to learn abstract representations of the data using a standard deep belief network (DBN) with multiple Restricted Boltzmann machines (RBMs). This allows for semantic-level transfer between the source and target modules. Both DBNs, for audio and visual information are built with the same number of intermediate layers, and then inter-modal embeddings are learned for each layer. Then, the learned mappings between the source and target are used to fine-tune the network with the transferred data and categorize each sequence into a letter.

More recently, Afouras et al. [161] proposed three systems that perform character-level classification. The visual front-end is common across the three systems and consists of a 3D-CNN on the input image sequence, with a filter width of five frames, followed by a ResNet which gradually decreases the spatial dimensions as depth increases. In contrast, the temporal back-end that receives the frame feature vectors and outputs a sentence character by character, is different for each system. The first one consists of three stacked Bi-LSTMs trained with CTC loss and decoding is performed with a beam search that incorporates prior information from an external language model. The second system uses depth-wise separable convolution layers, which consist of a separate convolution along the time dimension for every channel followed by a projection along the channel dimensions. The network contains 15 convolutional layers that were trained with a CTC loss and decoding is performed as described in the same way as the previous system. Finally, the last system has an encoder-decoder structure based on multi-head attention layers. It uses a base model with 6 encoder and decoder layers and 8 attention heads. This system has been trained with cross-entropy loss instead of CTC, hence it would be expected to implicitly learn an internal language model. Nevertheless, authors report that integrating an external language model in the decoding process improved their results.

3.2.4. Performance comparison

In this section we compare the performance of both hybrid and end-to-end DNN-based architectures. We compare the methods from Tables 4–6 that have been evaluated in the most common databases, being them AVLetters, GRID, LRW and OuluVS2.

For alphabet recognition, we find four DNN-based systems evaluated in the well known AVLetters database [137,143,17,147]. The first one was presented by Ngiam et al. [137] and consists of PCA features followed by a deep autoencoder, obtaining a classification accuracy of 64.40% WRR. In contrast, Moon et al. [143] proposed a method to obtain abstract representations of the raw data using a standard DBN. They fine-tune the video model with additional information transferred from audio data, obtaining 55.30% WRR. Petridis and Pantic [17] proposed to first train a deep autoencoder to compress the high dimensional image data into a low dimensional representation (named bottleneck features). Next, DCT features are computed to complement bottleneck ones and fed to an LSTM network to model the temporal dynamics, obtaining 58.10% WRR. Finally, Hu et al. [147] proposed a system based on multimodal RBMs (MRBMs), named Recurrent Temporal Multimodal Restricted Boltzmann Machines (RTMRBMs), which have the ability to extract semantic information from multi-sensory data and learn a joint representation across audio-visual modalities. They reported 64.63% WRR. Interestingly, these results are below those obtained by some traditional systems, e.g. the RFMA-based system presented in [118] obtained 69.60% WRR. Thus, for letter recognition in datasets such as AVLetters, traditional systems still outperform DL systems. The reason for this seems related to the dataset size, which is not large enough to train robust DL systems.

For word or sentence recognition, the most used databases have been GRID, LRW and OuluVS2. For the GRID corpus, we found six different architectures. Wand et al. presented three models for this database: the first one [20] consists of one Feed-forward layer followed by two recurrent LSTM layers and reported 79.50% WRR, while the second and third systems [160,164] combine three Feed-forward layers with an LSTM layer and reported 83.30% and 84.70% WRR for speaker-dependent experiments and 42.40% WRR in [160] for experiments in which the test speakers were unknown to the system. In contrast, Assael et al. [34] proposed a spatio-temporal CNN in combination with Bi-LSTMs and obtained a higher recognition rate of 93.40% WRR. Chung et al. [16] obtained 97.00% WRR with a system based on CNN and LSTM networks combined with attention mechanisms. Finally, Xu et al. [165] outperformed previous methods with a system that combines 3D-CNNs, highway networks, Bi-GRUs and attention mechanisms, obtaining slightly higher performance than [16] with 97.10% WRR. There is a considerable improvement in performance with respect to traditional systems, where the highest accuracy was 57.00% WRR reported by [120].

For the LRW database, Chung and Zisserman [19] presented an end-to-end architecture based on CNNs, reporting 61.10% WRR.

Stafylakis and Tzimiropoulos [156] presented a system based on 3D-CNN, residual networks and Bi-LSTMs and reported more than 20% improvement (83.00% WRR). Similarly, Petridis et al. [163] presented a system based on 3D-CNN, residual networks and Bi-GRU networks and reported 82.00% WRR. In yet another contribution, Chung et al. [16] proposed a system based on CNN and LSTM networks combined with attention mechanisms and obtained the best results reported so far, with 84.50% WRR.

For the OuluVS2 dataset, 13 architectures have been presented. Saitoh et al. [148] and Chung and Zisserman [19] presented several end-to-end systems mainly based on CNNs. The three systems proposed by Saitoh et al. reported recognition rates between 81.10% and 86.50% WRR, while Chung et al. reported 94.10% WRR. The main difference between these two works is that the networks in [148] used CFIs as input while [19] used directly a single image. In addition, Saitoh et al. used three well known pre-trained models based on CNNs: NIN [177], AlexNet [170] and GoogLeNet [171], while Chung et al. trained the network from scratch for the specific task of lip-reading. Several architectures were also proposed with LSTMs or Bi-LSTMs as classifiers. For these systems, different models to extract features were applied: CNNs in [128,25], VGG-M and SyncNet in [146], autoencoders in [152-154], 3D-CNN in [162] and PCA-NN in [150]. The latter one, in addition, used HMMs to model the temporal dynamics. For these architectures, the reported recognition rates were between 31.90% and 94.70% WRR. The lowest recognition rate corresponds to the system using VGG-M [146]. This comparatively low accuracy can be explained because VGG-M was pre-trained on ImageNet, a large database for object recognition and classification tasks, but not specific for lip-reading. In contrast, Petridis et al. [154] presented a system based on encoded features that reported the highest performance of 94.70% WRR, nearly followed by Chung and Zisserman [19] with 94.10% WRR. Nevertheless, compared to traditional architectures, there is a significant improvement of at least a 20% with respect to the highest performing traditional system, achieving 74.00% WRR in [128].

From the above paragraphs we can see that DNNs brought substantial accuracy improvements to ALR systems on databases such as GRID or OuluVS2, which focus on word- or sentence-classification tasks. These improvements have encouraged researchers to address more realistic settings and propose systems that target continuous lip-reading. Such settings are considerably more challenging than those found in word- or sentence-classification tasks, because each sentence has an unknown structure and can contain an arbitrary number of words whose time-boundaries are not known beforehand. For these reasons, when targeting continuous lip-reading it is convenient to predict smaller structures that approach the minimum distinguishable language units. Recent advances in end-to-end DL architectures have indeed focused on ALR systems that try to predict phonemes [149,51,139,83] or characters [16,25,161,165], instead of full words or pre-defined sentences. For example, Mroueh et al. [83] proposed Feed-forward DNNs to predict phonemes using the IBM AV-ASR database, a large-scale non-public AV database. Other architectures using CNNs and HMMs were presented by Noda et al. [51,139] and by Takashima et al. [149]. They tried to recognize Japanese phonemes using the ATR Japanese corpus [178] and obtained 22.50% WRR, 37.00% WRR and 51.00% WRR, respectively. Another architecture evaluated in the highly used GRID corpus has been recently presented by Xu et al. [165] for character-based classification. This very deep network combines 3D-CNNs, highway networks, Bi-GRUs and attention mechanisms and reported 97.10% WRR. In contrast, Chung et al. [16,25] presented an architecture based on CNN and LSTM networks combined with attention mechanisms. They evaluated their system in recently recorded large-scale databases such as MV-LRS and LRS, obtaining for character-based recognition 43.60% WRR and 49.80% WRR, respectively for each dataset. More recently, Afouras et al. [161] presented a comparison of three architectures dealing with character-based recognition evaluated on the LRS dataset. The architectures share the same visual features and only differ in the sequence classification; they obtained 37.80% WRR for the model using Bi-LSTMs, 45.00% WRR for the one using depth-wise convolutional layers and 50.00% WRR for the one using encoder-decoder with multi-head attention layers.

Thus, most recent DNN-based architectures report WRRs that, despite the different experimental settings, nearly double the performance reported by traditional systems, with WRRs of about 20% [31,37,131]. While this constitutes a great step forward in continuous lip-reading, it is worth noting that these results are still far from a system that can fully decode visual speech. Indeed, in real-world scenarios, the top-performing ALR systems currently approach WRRs of 50%, which means that we cannot recognize about half of the message. Thus, DNN-based systems and large-scale databases have significantly advanced the field but continuous ALR remains still an open problem.

4. Summary and conclusions

In this survey, we review the progression of ALR systems from 2007 to 2017 which highlights the technology shift from traditional architectures, typically consisting of image features in combination with HMMs, toward end-to-end DNN architectures, currently dominated by CNN features in combination with LSTMs.

In both the traditional and the DNN-based systems, we can conceptually identify two major blocks specific to ALR whose objectives are i) to parametrize the visual information observable at a given time instant or window, and ii) to map the visual features into speech units while incorporating temporal context, i.e. constraints to ensure that the decoded message is coherent. The latter provides robustness against noisy or imperfect estimates from the visual cues and helps to disambiguate between visually similar speech units.

Traditional ALR systems mainly consist of features based on appearance or image transforms in combination with HMMs that model the temporal dynamics of the spoken sequence using short term context information. While HMMs can be considered the de-facto standard for modeling context, a variety of features have been explored with the goal to find the best descriptor for visual speech. As shown in Section 3.1, the most widely used features in visual speech systems have been DCT and AAMs, but there is no agreement on which feature would be optimal.

In the last years, we observe how DNN-based systems have quickly started to replace all the blocks from traditional systems by end-to-end DNNs. In this survey, we discuss the most popular DNN architectures for ALR systems and compare several variations that follow the same baseline structure (i.e. combinations of CNNs and LSTMs). In particular, variants on the feature side include different types of data used to feed the CNNs (e.g. RGB or gray-scale images, 3D or 2D structures), and network specifications (e.g. number of convolutional and fully connected layers). In terms of classification, ALR researchers have explored LSTM networks that differ in how the output is decoded (e.g. step by step or at the end of the sequence), the network's direction (forward, backward or bidirectional), and the number of layers (which relates to the context scale that is considered). In addition, we comment other DNNs used for lip-reading that explore alternatives to the CNN-LSTM baseline, such as Feed-forward networks, DBN, or CNNs.

Comparing traditional systems with DL architectures we observe that the latter provide a significant improvement in terms of performance. For instance, for the GRID corpus, several DL architectures considerably outperformed the best traditional system with up to a 40% improvement, e.g. Assael et al. [34], Chung et al. [16] and Xu et al. [165] proposed end-to-end architectures that achieved up to 97% WRR, compared to the 57% WRR obtained by Kolossa et al. [120].

Similarly, in the OuluVS2 database, DNN-systems [148,146] reported more than 20% improvement with respect to the best-performing traditional system, which achieved 74% WRR [128].

Nevertheless, the remarkable results of end-to-end DL architectures addressing word or sentence recognition in databases such as GRID or OuluVS2, cannot be directly extrapolated to more realistic settings that target continuous lip-reading. In word or sentence recognition tasks, the output of the system is restricted to a pre-defined number of possible classes, in contrast to continuous lip-reading where the target is natural speech. In this way, continuous lip-reading systems must be able to decode any word of the dictionary and process sentences that contain an arbitrary number of words with unknown time-boundaries. Thus, recent attempts to produce continuous lip-reading systems have focused on elementary language structures such as characters or phonemes. For instance, hybrid architectures for continuous speech recognition in Japanese [51,139,149] have targeted phonemes achieving between 22% and 51% WRR, while Chung et al. [16] and Afouras et al. [161] achieved near 50% WRR targeting characters for a large-scale dataset in English.

Despite the recognition rates for continuous lip-reading may appear modest in comparison to those achieved by audio-based systems, the field has undeniably made a significant step forward. Interestingly, an analogous effect can be observed when humans try to decode speech: given sufficiently clean signals, most people can effortlessly decode the audio channel, but would struggle to perform lip-reading, since the ambiguity of the visual cues makes it necessary the use of further context to decode the message. Thus, it is not surprising that the main challenges in ALR systems regard to the robustness to visual ambiguities through the modeling of context information.

Most recent works suggest that the optimal modeling of temporal sequences is still an open problem, which is currently been tackled by means of recurrent neural networks. Specifically, LSTMs have been widely used for modeling sequences because of their ability to retain both short- and long-term context information in their cell structures, although it is not clear how to take full advantage of such ability. For instance, several authors have tried to model different scales of context by adding multiple LSTM layers, aiming to introduce constraints related to bigger speech structures such as connected phonemes, syllables, words or sentences. Other authors have used bidirectional networks (widely used in audio speech recognition because of their ability to model past and future context), which should be helpful for dealing with visual ambiguities that are related to previous and posterior mouth positions (i.e. a similar idea to that from triphoneme models). However, bidirectional networks involve a higher computational cost that unidirectional ones and require that the whole signal is available beforehand, not allowing for real-time decoding. Finally, attention models have also been recently explored because they help to highlight the most relevant pieces of information from the large amount of data potentially available. Thus, current efforts tend toward techniques that allow a more comprehensive modeling and interpretability of the retained context.

Conflict of interest statement

There is no conflict of interest.

Acknowledgments

This work is partly supported by the Spanish Ministry of Economy and Competitiveness, Spain under project grant TIN2017-90124-P, the Ramon y Cajal Programme, the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), and the Kristina project funded by the European Union Horizon 2020 - Research and Innovation Framework Programme under grant agreement No. 645012.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2018.07.002.

References

[1] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.
[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (9) (2003) 1306–1326.
[3] G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-visual automatic speech recognition: an overview, Issues Vis. Audio-Vis. Speech Process. 22 (2004) 23–61.
[4] A. Chiţu, L.J. Rothkrantz, Automatic visual speech recognition, Speech Enhanc. Model. Recognit. Algoritm. Appl. (2012) 95–120.
[5] N.P. Erber, Auditory-visual perception of speech, J. Speech Hear. Disord. 40 (4) (1975) 481–492.
[6] W.H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am. 26 (2) (1954) 212–215.
[7] S. Hilder, R. Harvey, B.-J. Theobald, Comparison of human and machine-based lip-reading., Proc. International Conference on Auditory-Visual Speech Processing, 2009. pp. 86–89.
[8] R.E. Ronquest, S.V. Levi, D.B. Pisoni, Language identification from visual-only speech signals, Atten. Percept. Psychophys. 72 (6) (2010) 1601–1613.
[9] R. Seymour, D. Stewart, J. Ming, Comparison of image transform-based features for visual speech recognition in clean and corrupted videos, J. Signal Image Video Process. (2008) 14–22.
[10] E. Antonakos, A. Roussos, S. Zafeiriou, A survey on mouth modeling and analysis for sign language recognition, Proc. International Conference on Automatic Face and Gesture Recognition, vol. 1, 2015. pp. 1–7.
[11] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[12] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, K. Murphy, A coupled HMM for audio-visual speech recognition, Proc. International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2002. pp. 2013–2016.
[13] Z. Zhou, G. Zhao, X. Hong, M. Pietikäinen, A review of recent advances in visual speech decoding, Image Vis. Comput. 32 (9) (2014) 590–605.
[14] W.C. Yau, D.K. Kumar, H. Weghorn, Visual speech recognition using motion features and hidden Markov models, Proc. International Conference on Computer Analysis of Images and Patterns, 2007. pp. 832–839.
[15] C. Sui, M. Bennamoun, R. Togneri, Listening with your eyes: towards a practical visual speech recognition system using deep Boltzmann machines, Proc. International Conference on Computer Vision, 2015. pp. 154–162.
[16] J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Lip reading sentences in the wild, Proc. Conference on Computer Vision and Pattern Recognition, 2017. pp. 3444–3453.
[17] S. Petridis, M. Pantic, Deep complementary bottleneck features for visual speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing, 2016. pp. 2304–2308.
[18] I. Almajai, S. Cox, R. Harvey, Y. Lan, Improved speaker independent lip reading using speaker adaptive training and deep neural networks, Proc. International Conference on Acoustics, Speech and Signal Processing, 2016. pp. 2722–2726.
[19] J.S. Chung, A. Zisserman, Lip reading in the wild, Proc. Asian Conference on Computer Vision, 2016. pp. 87–103.
[20] M. Wand, J. Koutník, J. Schmidhuber, Lipreading with long short-term memory, Proc. International Conference on Acoustics, Speech and Signal Processing, 2016. pp. 6115–6119.
[21] A.V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, Dynamic Bayesian networks for audio-visual speech recognition, EURASIP J. Adv. Signal Process. 2002 (11) (2002) 1–15.
[22] W.F. Twaddell, On defining the phoneme, Language 11 (1) (1935) 5–62.
[23] C.G. Fisher, Confusions among visually perceived consonants, J. Speech Lang. Hear. Res. 11 (4) (1968) 796–804.
[24] L. Cappelletta, N. Harte, Viseme definitions comparison for visual-only speech recognition, Proc. European Conference on Signal Processing, 2011. pp. 2109–2113.
[25] J.S. Chung, A. Zisserman, Lip reading in profile, Proc. British Machine Vision Conference, 2017.
[26] V. Sahu, M. Sharma, Result based analysis of various lip tracking systems, Proc. International Conference on Green High Performance Computing, 2013. pp. 1–7.
[27] K.L. Moll, R.G. Daniloff, Investigation of the timing of velar movements during speech, J. Acoust. Soc. Am. 50 (2B) (1971) 678–684.
[28] H.L. Bear, R.W. Harvey, B.-J. Theobald, Y. Lan, Which phoneme-to-viseme maps best improve visual-only computer lip-reading? Proc. International Symposium on Visual Computing, 2014. pp. 230–239.
[29] T.J. Hazen, K. Saenko, C.-H. La, J.R. Glass, A segment-based audio-visual speech recognizer: data collection, development, and initial experiments, Proc. International Conference on Multimodal Interfaces, 2004. pp. 235–242.
[30] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, Audio visual speech recognition, Tech. rep., IDIAP. 2000.

[31] A. Fernandez-Lopez, F.M. Sukno, Automatic viseme vocabulary construction to enhance continuous lip-reading, Proc. Int. Conf. Comput. Vis. Theory and Appl. 5 (2017) 52–63.
[32] J. Jeffers, M. Barley, Speechreading (lipreading), Thomas. 1971.
[33] E. Bozkurt, C.E. Erdem, E. Erzin, T. Erdem, M. Ozkan, Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation, Proc. International Conference on Signal Processing and Communications Applications, 2007. pp. 1–4.
[34] Y.M. Assael, B. Shillingford, S. Whiteson, N de Freitas, LipNet: sentence-level lipreading, Proc. GPU Technology Conference, 2017.
[35] J.N. Buchan, M. Paré, K.G. Munhall, Spatial statistics of gaze fixations during dynamic face processing, Soc. Neurosci. 2 (1) (2007) 1–13.
[36] I.d.l.R.R. Ortiz, Lipreading in the prelingually deaf: what makes a skilled speechreader? Span. J. Psychol. 11 (02) (2008) 488–502.
[37] A. Fernandez-Lopez, O. Martinez, F.M. Sukno, Towards estimating the upper bound of visual-speech recognition: the visual lip-reading feasibility database, Proc. International Conference on Automatic Face and Gesture Recognition, 2017. pp. 208–215.
[38] A. Gabbay, A. Ephrat, T. Halperin, S. Peleg, Seeing through noise: speaker separation and enhancement using visually-derived speech, Proc. International Workshop on Computer Vision for Audio-Visual Media, 2017.
[39] D. Stewart, R. Seymour, A. Pass, J. Ming, Robust audio-visual speech recognition under noisy audio-video conditions, IEEE Trans. Cybern. 44 (2) (2014) 175–184.
[40] F.S. Lesani, F.F. Ghazvini, R. Dianat, Mobile phone security using automatic lip reading, Proc. International Conference on E-Commerce in Developing Countries: With Focus on e-Business, 2015. pp. 1–5.
[41] S. Mathulaprangsan, C.-Y. Wang, A.Z. Kusum, T.-C. Tai, J.-C. Wang, A survey of visual lip reading and lip-password verification, Proc. International Conference on Orange Technologies, 2015. pp. 22–25.
[42] S. Sengupta, A. Bhattacharya, P. Desai, A. Gupta, Automated lip reading technique for password authentication, Int. J. Appl. Inf. Syst. (2012)2249-0868.
[43] C. Georgakis, S. Petridis, M. Pantic, Visual-only discrimination between native and non-native speech, Proc. International Conference on Acoustics, Speech and Signal Processing, IEEE. 2014, pp. 4828–4832.
[44] C. Georgakis, S. Petridis, M. Pantic, Discriminating native from non-native speech using fusion of visual cues, Proc. International Conference on Multimedia, 2014. pp. 1177–1180.
[45] C. Georgakis, S. Petridis, M. Pantic, Discrimination between native and non–native speech using visual features only, IEEE Trans. Cybern. 46 (12) (2016) 2758–2771.
[46] A. Ephrat, T. Halperin, S. Peleg, Improved speech reconstruction from silent video, Proc. International Workshop on Computer Vision for Audio-Visual Media, 2017.
[47] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, B. Denby, An articulatory-based singing voice synthesis using tongue and lips imaging, Proceedings of Interspeech, 2016. pp. 1467–1471.
[48] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, B. Yvert, Real-time control of an articulatory-based speech synthesizer for brain computer interfaces, PLoS Comput. Biol. 12 (11). (2016)e1005119.
[49] A. Gabbay, A. Shamir, S. Peleg, Visual speech enhancement, Proceedings of Interspeech (in press), 2018.
[50] A. Britto Mattos, D.A. Borges Oliveira, Multi-view mouth renderization for assisting lip-reading, Proc. International Conference on the Web for All (in press), 2018.
[51] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Audio-visual speech recognition using deep learning, Appl. Intell. 42 (4) (2015) 722–737.
[52] T. Afouras, J.S. Chung, A. Zisserman, The conversation: deep audio-visual speech enhancement, Proceedings of Interspeech (in press), 2018.
[53] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2006) 2421–2424.
[54] D.L. Howell, Confusion Modelling for Lip-reading, University of East Anglia. 2015.Ph.D. Thesis.
[55] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[56] S.J. Cox, R. Harvey, Y. Lan, J.L. Newman, B.-J. Theobald, The challenge of multi-speaker lip-reading, Proc. International Conference on Auditory-Visual Speech Processing, 2008. pp. 179–184.
[57] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T.S. Huang, AVICAR: audio-visual speech corpus in a car environment., Proceedings of Interspeech, 2004.
[58] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Proc. International Conference on Audio and Video-based Biometric Person Authentication, vol. 964, 1999. pp. 965–966.
[59] C. Sanderson, The VidTIMIT database, Tech. Rep., IDIAP. 2002.
[60] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, et al. The BANCA database and evaluation protocol, Proc. International Conference on Audio- and Video-based Biometric Person Authentication, 2003. pp. 625–638.
[61] J. Huang, G. Potamianos, J. Connell, C. Neti, Audio-visual speech recognition using an infrared headset, Speech Comm. 44 (1) (2004) 83–96.
[62] R. Goecke, J.B. Millar, The audio-video Australian English speech data corpus AVOZES, Proc. International Conference on Spoken Language Processing, 2004. pp. 2525–2528.
[63] A. Ortega, F. Sukno, E. Lleida, A.F. Frangi, A. Miguel, L. Buera, E. Zacur, AV@CAR: a Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition., Proc. International Conference on Language Resources and Evaluation, 2004. pp. 763–767.
[64] E.K. Patterson, S. Gurbuz, Z. Tufekci, J.N. Gowdy, CUAVE: a new audio-visual database for multimodal human-computer interface research, Proc. International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2002. pp. 2017–2020.
[65] N.A. Fox, B.A. OMullane, R.B. Reilly, VALID: a new practical audio-visual database, and comparative results, Proc. International Conference on Audio- and Video-based Biometric Person Authentication, 2005. pp. 777–786.
[66] P.J. Lucey, G. Potamianos, S. Sridharan, Patch-based analysis of visual speech from multiple views, Proc. International Conference on Auditory-Visual Speech Processing, 2008.
[67] D. Petrovska-Delacrétaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, M. Ardabilian, E. Krichen, M.-A. Mellakh, A. Chaari, S. Guerfi, et al. The IV 2 multimodal biometric database (including iris, 2D, 3D, stereoscopic, and talking face data), and the IV 2-2007 evaluation campaign, Proc. International Conference on Biometrics: Theory, Applications and Systems, 2008. pp. 1–7.
[68] J. Trojanová, M. Hrúz, P. Campr, M. železnỳ, Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition, Proc. International Conference on Language Resources and Evaluation, 2008.
[69] G. Zhao, M. Barnard, M. Pietikainen, Lipreading with local spatiotemporal descriptors, IEEE Trans. Multimedia 11 (7) (2009) 1254–1265.
[70] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, et al. CENSREC-1-AV: an audio-visual corpus for noisy bimodal speech recognition, Proc. International Conference on Auditory-Visual Speech Processing, 2010.
[71] A. Pass, J. Zhang, D. Stewart, An investigation into features for multi-view lipreading, Proc. International Conference on Image Processing, 2010. pp. 2417–2420.
[72] A.G. Chitu, K. Driel, L.J. Rothkrantz, Automatic lip reading in the Dutch language using active appearance models on high speed recordings, Proc. International Conference on Text, Speech and Dialogue, 2010. pp. 259–266.
[73] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, R. Orglmeister, WAPUSK20 — a database for robust audiovisual speech recognition., Proc. International Conference on Language Resources and Evaluation, 2010.
[74] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, R. Bowden, Improving visual features for lip-reading, Proc. International Conference on Auditory-Visual Speech Processing, 2010.
[75] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, F. Bimbot, BL–Database: A French Audiovisual Database for Speech Driven Lip Animation Systems, INRIA. 2011.Ph.D. Thesis.
[76] Y.W. Wong, S.I. Chng, K.P. Seng, L.-M. Ang, S.W. Chin, W.J. Chew, K.H. Lim, A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities, Pattern Recogn. Lett. 32 (13) (2011) 1503–1510.
[77] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data, Proc. International Workshop on Multimedia and Expo, 2012. pp. 635–640.
[78] M. Igras, B. Ziółko, T. Jadczyk, Audiovisual database of Polish speech recordings, Studia Informatica 33 (2B) (2012) 163–172.
[79] A. Rekik, A. Ben-Hamadou, W. Mahdi, A new visual speech recognition approach for RGB-D cameras, Proc. International Conference on Image Analysis and Recognition, 2014. pp. 21–28.
[80] D. Estival, S. Cassidy, F. Cox, D. Burnham, AusTalk: an audio-visual corpus of Australian English, Proc. International Conference on Language Resources and Evaluation, 2014.
[81] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, M. Szykulski, An audio-visual corpus for multimodal automatic speech recognition, J. Intell. Inf. Syst. (2017) 1–26.
[82] I. Anina, Z. Zhou, G. Zhao, M. Pietikäinen, OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis, Proc. International Conference on Automatic Face and Gesture Recognition, vol. 1, 2015. pp. 1–5.
[83] Y. Mroueh, E. Marcheret, V. Goel, Deep multimodal learning for audio-visual speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing, 2015. pp. 2130–2134.
[84] N. Harte, E. Gillen, TCD-TIMIT: an audio-visual corpus of continuous speech, IEEE Trans. Multimedia 17 (5) (2015) 603–615.
[85] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, M. železnỳ, HAVRUS corpus: high-speed recordings of audio-visual Russian speech, Proc. International Conference on Speech and Computer, 2016. pp. 338–345.
[86] S. Petridis, J. Shen, D. Cetin, M. Pantic, Visual-only recognition of normal, whispered and silent speech, Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.
[87] Y. Lan, B.-J. Theobald, R. Harvey, View independent computer lip-reading, Proc. International Conference on Multimedia and Expo, 2012. pp. 432–437.
[88] K. Kumar, T. Chen, R.M. Stern, Profile view lip reading, Proc. International Conference on Acoustics, Speech and Signal Processing, vol. 4, 2007. pp. 429–432.
[89] X.L.H. Yao, X.H.Q. Wang, HIT-AVDB-II: a new multi-view and extreme feature cases contained audio-visual database for biometrics, Proc. Conference on Information Sciences, 2008.
[90] V. Estellers, J.-P. Thiran, Multipose audio-visual speech recognition, Proc. European Conference on Signal Processing, 2011. pp. 1065–1069.

[91] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[92] S. Zafeiriou, C. Zhang, Z. Zhang, A survey on face detection in the wild: past, present and future, Comput. Vis. Image Underst. 138 (2015) 1–24.
[93] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, Proc. European Conference on Computer Vision, 2014. pp. 720–735.
[94] J. Orozco, B. Martinez, M. Pantic, Empirical analysis of cascade deformable models for multi-view face detection, Image Vis. Comput. 42 (2015) 47–61.
[95] J. Orozco, O. Rudovic, J. Gonzàlez, M. Pantic, Hierarchical on-line appearance-based tracking for 3D head pose, eyebrows, lips, eyelids and irises, Image Vis. Comput. 31 (4) (2013) 322–340.
[96] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, Proc. Conference on Computer Vision and Pattern Recognition, 2013. pp. 3444–3451.
[97] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: database and results, Image Vis. Comput. 47 (2016) 3–18.
[98] X. Yu, J. Huang, S. Zhang, D.N. Metaxas, Face landmark fitting via optimized part mixtures and cascaded deformable model, IEEE Trans. Pattern Anal. Mach. Intell. 38 (11) (2016) 2212–2226.
[99] G. Tzimiropoulos, M. Pantic, Fast algorithms for fitting active appearance models to unconstrained images, Int. J. Comput. Vis. 122 (1) (2017) 17–33.
[100] Y. Wu, T. Hassner, K. Kim, G. Medioni, P. Natarajan, Facial landmark detection with tweaked convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. (2017) 1-1.
[101] J. Luettin, N.A. Thacker, S.W. Beet, Visual speech recognition using active shape models and hidden Markov models, Proc. International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 1996. pp. 817–820.
[102] J.N. Gowdy, A. Subramanya, C. Bartels, J. Bilmes, DBN based multi-stream models for audio-visual speech recognition, Proc. International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 2004. pp. I–993.
[103] N. Eveno, A. Caplier, P.-Y. Coulon, Accurate and quasi-automatic lip tracking, Circuits Syst. Video Technol. 14 (5) (2004) 706–715.
[104] K. Mase, A. Pentland, Automatic lipreading by optical-flow analysis, Syst. Comput. Jpn. 22 (6) (1991) 67–76.
[105] X. Hong, H. Yao, Y. Wan, R. Chen, A PCA based visual DCT feature extraction method for lip-reading, Proc. International Conference on Intelligent Information Hiding and Multimedia, 2006. pp. 321–326.
[106] P. Lucey, G. Potamianos, Lipreading using profile versus frontal views, Proc. International Workshop on Multimedia Signal Processing, 2006. pp. 24–28.
[107] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, B.-J. Theobald, Recent developments in automated lip-reading, Optics and Photonics for Counterterrorism, Crime Fighting and Defence IX; and Optical Materials and Biomaterials in Security and Defence Systems Technology X, vol. 8901, International Society for Optics and Photonics. 2013, pp. 89010J.
[108] P.J. Lucey, G. Potamianos, S. Sridharan, A unified approach to multi-pose audio-visual ASR, Proceedings of Interspeech, 2007. pp. 650–653.
[109] M. Gurban, J.-P. Thiran, Information theoretic feature extraction for audio-visual speech recognition, Signal Process. 57 (12) (2009) 4765–4776.
[110] J. Huang, B. Kingsbury, Audio-visual deep learning for noise robust speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing, 2013. pp. 7596–7599.
[111] G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition, IEEE-ACM Trans. Audio Speech Lang. Process. 17 (3) (2009) 423–435.
[112] G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition, Proc. International Conference on Multimodal Processing and Interaction, 2008. pp. 1–15.
[113] S.-L. Wang, A.W.-C. Liew, W.H. Lau, S.H. Leung, An automatic lipreading system for spoken digits with limited training data, Circuits Syst. Video Technol. 18 (12) (2008) 1760–1765.
[114] P.J. Lucey, S. Sridharan, D.B. Dean, Continuous pose-invariant lipreading, Proceedings of Interspeech, 2008. pp. 2679–2682.
[115] S. Pachoud, S. Gong, A. Cavallaro, Macro-cuboid based probabilistic matching for lip-reading digits, Proc. Conference on Computer Vision and Pattern Recognition, 2008. pp. 1–8.
[116] A. Rekik, A. Ben-Hamadou, W. Mahdi, An adaptive approach for lip-reading using image and depth data, Multimedia Tools Appl. 75 (14) (2016) 8609–8636.
[117] V. Estellers, M. Gurban, J.-P. Thiran, On dynamic stream weighting for audio-visual speech recognition, IEEE-ACM Trans. Audio Speech Lang. Process. 20 (4) (2012) 1145–1157.
[118] Y. Pei, T.-K. Kim, H. Zha, Unsupervised random forest manifold alignment for lipreading, Proc. IEEE International Conference on Computer Vision, 2013. pp. 129–136.
[119] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, R. Bowden, Comparing visual features for lipreading, Proc. International Conference on Auditory-Visual Speech Processing, 2009. pp. 102–106.
[120] D. Kolossa, S. Zeiler, A. Vorwerk, R. Orglmeister, Audiovisual speech recognition with missing or unreliable data, Proc. International Conference on Auditory-Visual Speech Processing, 2009. pp. 117–122.
[121] Z. Zhou, G. Zhao, M. Pietikainen, Lipreading: a graph embedding approach, Proc. International Conference on Pattern Recognition, 2010. pp. 523–526.
[122] Z. Zhou, G. Zhao, M. Pietikäinen, Towards a practical lipreading system, Proc. Conference on Computer Vision and Pattern Recognition, 2011. pp. 137–144.
[123] E.-J. Ong, R. Bowden, Learning temporal signatures for lip reading, Proc. International Conference on Computer Vision Workshops, 2011. pp. 958–965.
[124] E.-J. Ong, R. Bowden, Learning sequential patterns for lipreading, Proc. British Machine Vision Conference, 2011.
[125] Z. Zhou, X. Hong, G. Zhao, M. Pietikäinen, A compact representation of visual speech data using latent variables, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1). (2014)
[126] C. Sui, R. Togneri, M. Bennamoun, A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition, Speech Comm. 90 (2017) 26–38.
[127] P. Wu, H. Liu, X. Li, T. Fan, X. Zhang, A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion, IEEE Trans. Multimedia 18 (3) (2016) 326–338.
[128] D. Lee, J. Lee, K.-E. Kim, Multi-view automatic lip-reading using neural network, Proc. Asian Conference on Computer Vision, 2016. pp. 290–302.
[129] K. Thangthai, R. Harvey, S. Cox, B.-J. Theobald, Improving lip-reading performance for robust audiovisual speech recognition using DNNs, Proc. International Conference on Auditory-Visual Speech Processing, 2015. pp. 127–131.
[130] D. Howell, S. Cox, B. Theobald, Visual units and confusion modelling for automatic lip-reading, Image Vis. Comput. 51 (2016) 1–12.
[131] Y. Lan, R. Harvey, B.-J. Theobald, Insights into machine lip reading, Proc. International Conference on Acoustics, Speech and Signal Processing, 2012. pp. 4825–4828.
[132] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, T.S. Huang, Lipreading by locality discriminant graph, Proc. International Conference on Image Processing, vol. 3, 2007. pp. 325–328.
[133] E. Marcheret, V. Libal, G. Potamianos, Dynamic stream weight modeling for audio-visual speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing, vol. 4, 2007. pp. 945–948.
[134] X. Shao, J. Barker, Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment, Speech Comm. 50 (4) (2008) 337–353.
[135] T. Saitoh, R. Konishi, Profile lip reading for vowel and word recognition, Proc. International Conference on Pattern Recognition, 2010. pp. 1356–1359.
[136] R. Navarathna, T. Kleinschmidt, D.B. Dean, S. Sridharan, P.J. Lucey, Can audio-visual speech recognition outperform acoustically enhanced speech recognition in automotive environment? Proceedings of Interspeech, 2011. pp. 2241–2244.
[137] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, Proc. International Conference on Machine Learning, 2011. pp. 689–696.
[138] V. Estellers, J.-P. Thiran, Multi-pose lipreading and audio-visual speech recognition, EURASIP J. Adv. Signal Process. 2012 (1) (2012) 51.
[139] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Lipreading using convolutional neural network, Proceedings of Interspeech, 2014. pp. 1149–1153.
[140] H.L. Bear, S.J. Cox, R.W. Harvey, Speaker-independent machine lip-reading with speaker-dependent viseme classifiers, Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing, 2015. pp. 190–195.
[141] H.L. Bear, R.W. Harvey, Y. Lan, Finding phonemes: improving machine lip-reading, Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing, 2017.
[142] A. Biswas, P.K. Sahu, M. Chandra, Multiple camera in car audio-visual speech recognition using phonetic and visemic information, Comput. Electr. Eng. 47 (2015) 35–50.
[143] S. Moon, S. Kim, H. Wang, Multimodal transfer deep learning with applications in audio-visual recognition, MMML Workshop at Neural Information Processing Systems, 2015.
[144] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, K. Takeda, Integration of deep bottleneck features for audio-visual speech recognition, Proceedings of Interspeech, 2015. pp. 563–567.
[145] H.L. Bear, R. Harvey, Decoding visemes: improving machine lip-reading, Proc. International Conference on Acoustics, Speech and Signal Processing, 2016. pp. 2009–2013.
[146] J.S. Chung, A. Zisserman, Out of time: automated lip sync in the wild, Proc. Asian Conference on Computer Vision, 2016. pp. 251–263.
[147] D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition, Proc. Conference on Computer Vision and Pattern Recognition, 2016. pp. 3574–3582.
[148] T. Saitoh, Z. Zhou, G. Zhao, M. Pietikäinen, Concatenated frame image based CNN for visual speech recognition, Proc. Asian Conference on Computer Vision, 2016. pp. 277–289.
[149] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono, Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss, Proceedings of Interspeech, 2016. pp. 277–281.
[150] M. Zimmermann, M.M. Ghazi, H.K. Ekenel, J.-P. Thiran, Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system, Proc. Asian Conference on Computer Vision, 2016. pp. 264–276.
[151] H.L. Bear, R. Harvey, Phoneme-to-viseme mappings: the good, the bad, and the ugly, Speech Comm. 95 (2017) 40–67.
[152] S. Petridis, Z. Li, M. Pantic, End-to-end visual speech recognition with LSTMs, Proc. International Conference on Acoustics, Speech and Signal Processing, 2017. pp. 2592–2596.
[153] S. Petridis, Y. Wang, Z. Li, M. Pantic, End-to-end audiovisual fusion with LSTMs, Proc. International Conference on Auditory-Visual Speech Processing, 2017.

[154] S. Petridis, Y. Wang, Z. Li, M. Pantic, End-to-end multi-view lipreading, Proc. British Machine Vision Conference, 2017.
[155] M.H. Rahmani, F. Almasganj, Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features, Proc. International Conference on Pattern Recognition and Image Analysis, 2017. pp. 195–199.
[156] T. Stafylakis, G. Tzimiropoulos, Combining residual networks with LSTMs for lipreading, Proceedings of Interspeech, 2017. pp. 3652–3656.
[157] G. Sterpu, N. Harte, Towards lipreading sentences using active appearance models, Proc. International Conference on Auditory-Visual Speech Processing, 2017.
[158] K. Thangthai, R. Harvey, Improving computer lipreading via DNN sequence discriminative training techniques, Proc. Interspeech (2017) 3657–3661.
[159] K. Thangthai, H.L. Bear, R. Harvey, Comparing phonemes and visemes with DNN-based lipreading, Proc. British Machine Vision Conference, 2017.
[160] M. Wand, J. Schmidhuber, Improving speaker-independent lipreading with domain-adversarial training, Proceedings of Interspeech, 2017. pp. 3662–3666.
[161] T. Afouras, J.S. Chung, A. Zisserman, Deep lip reading: a comparison of models and an online application, Proceedings of Interspeech (in press), 2018.
[162] H.L. Fung, B. Mak, End-to-end low-resource lip-reading with maxout CNN and LSTM, Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.
[163] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic, End-to-end audiovisual speech recognition, Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.
[164] M. Wand, N.T. Vu, J. Schmidhuber, Investigations on end-to-end Aaudiovisual fusion, Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.
[165] K. Xu, D. Li, N. Cassimatis, X. Wang, LCANet: end-to-end lipreading with cascaded attention-CTC, Proc. International Conference on Automatic Face and Gesture Recognition, 2018. pp. 548–555.
[166] F.A. Gers, J.A. Schmidhuber, F.A. Cummins, Learning to forget: continual prediction with LSTM, Neural Comput. 12 (10) (2000) 2451–2471.
[167] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, Proc. International Conference on Machine Learning, 2006. pp. 369–376.
[168] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks., Proc. International Conference on Machine Learning, vol. 14, 2014. pp. 1764–1772.
[169] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: scaling up end-to-end speech recognition, Proc. International Conference on Machine Learning, 2014.
[170] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Proc. Conference on Advances in Neural Information Processing Systems, 2012. pp. 1097–1105.
[171] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proc. Conference on Computer Vision and Pattern Recognition, 2015. pp. 1–9.
[172] F.J. Ordóñez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.
[173] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5) (2005) 602–610.
[174] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, Neural Netw. 5 (2) (1994) 157–166.
[175] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. Conference on Computer Vision and Pattern Recognition, 2016. pp. 770–778.
[176] R.K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, Advances in Neural Information Processing systems, 2015. pp. 2377–2385.
[177] M. Lin, Q. Chen, S. Yan, Network in network, Proc. International Conference on Learning Representations, 2014.
[178] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, T. Watanabe, Construction of a large-scale Japanese speech database and its management system, Proc. International Conference on Acoustics, Speech, and Signal Processing, 1989. pp. 560–563.