A Review on Deep Learning-Based Automatic Lipreading

Carlos Santos[1], António Cunha[2,3], and Paulo Coelho[1,4(✉)]

[1] School of Technology and Management, Polytechnic of Leiria, 2411-901 Leiria, Portugal
2180284@my.ipleiria.pt

[2] Escola de Ciências e Tecnologias, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5001-801 Vila Real, Portugal
atcunha@utad.pt

[3] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal

[4] Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra), DEEC, Pólo II, 3030-290 Coimbra, Portugal
paulo.coelho@ipleiria.pt

Abstract. Automatic Lip-Reading (ALR), also known as Visual Speech Recognition (VSR), is the technological process to extract and recognize speech content, based solely on the visual recognition of the speaker's lip movements. Besides hearing-impaired people, regular hearing people also resort to visual cues for word disambiguation, every time one is in a noisy environment. Due to the increasingly interest in developing ALR systems, a considerable number of research articles are being published. This article selects, analyses, and summarizes the main papers from 2018 to early 2022, from traditional methods with handcrafted feature extraction algorithms to end-to-end deep learning based ALR which fully take advantage of learning the best features, and of the evergrowing publicly available databases. By providing a recent state-of-the-art overview, identifying trends, and presenting a conclusion on what is to be expected in future work, this article becomes an efficient way to update on the most relevant ALR techniques.

Keywords: Automatic Lip-reading · Deep Learning · Audio-visual Automatic Speech Recognition

1 Introduction

Human speech is one of the most important forms of communication, through which we rapidly convey information, concepts, and ideas, therefore being a key driver of Human Evolution. It also allows to enhance the performance of teaching/learning and, as we are social beings, gossip, which makes up a large portion of our human-to-human interactions.

A. Cunha et al. (Eds.): MobiHealth 2022, LNICST 484, pp. 180–195, 2023.
https://doi.org/10.1007/978-3-031-32029-3_17

Phoneticians call Phonemes the basic speech structures (or minimal units of speech), that enable to distinguish one word from another (e.g.: the phoneme /p/ is what distinguishes the word pat from bat) [1]. Visemes are a shortened version of the phrase visual phonemes and refer to any individual and contrastive visually perceived unit [2].

Lip reading or visual speech recognition is a procedure used to recognize and interpret the speech by analyzing the lip's movement. It is used to complete relayed information by people with hearing difficulties, either by being congenitally deaf or just by a significant decrease in speech-to-noise ratio, i.e., by augmenting the noise level and maintaining the speech level or by maintaining the noise level but diminishing the speech level [1]. Bauman's study [3] reported that hearing-impaired people understood 21% of speech, just using residual hearing, 64% if they combined residual hearing with either a hearing aid or with speechreading, and 90% if they used their residual hearing, hearing aids, and speechreading.

Video Speech Recognition, may be considered lip-reading based on artificial intelligence techniques, and also contributes to: speech synthesizing; multi-view mouth rendering; silent passwords; audio-less videos transcriber; speech recognition under noisy conditions; isolation of individual speakers; forensic study on surveillance videos; and face liveness detection.

Within approximately a single human generation, lip-reading evolved from a 4-grey layered template recognition model as the first Audio-Visual Automatic Speech Recognition (AV-ASR) [4], to recent, powerful, and more complex end-to-end deep neural network models [5–9], where linear considerations try to answer a non-linear question. ALR techniques vary in ways to find Regions of Interest (RoI) from frames, extract features, transform them, and to classifying in simple utterances up to full sentences, in real-time or for forensic analysis. In this time span, research on ALR has evolved from a limited number of researchers, to multinational, multicultural, ever-growing, and diversifying teams of researchers, producing an equally growing number of papers [10–14].

As a survey on the specific field of ALR, this document mainly aims and contributes to: (1) select a considerable number of recent relevant literature, as state-of-the-art; (2) summarize the selected literature; (3) identify patterns and tendencies; (4) establish a review that serves as a critical basis for the most recent methods, strategies and results that allow the development of works related to this area of knowledge.

Section 2 of this paper presents the criteria and methods used to select the reviewed works, Section 3 presents the state of the art in automatic lip-reading, along with a brief author's discussion, and Section 4 presents the conclusions and what can be done in terms of improvements in future works.

2 Materials and Methods

2.1 Research Questions

The systematic review that follows is based on a set of research questions: RQ1 - Which methods are more suitable for visual clues only automatic lip-reading?; RQ2 - Which methods are mainly used to support the analysis of lip-reading data? RQ3 - Which methods are specifically studied with the available datasets?; RQ4 - What challenges are still open for lip-reading solutions?

2.2 Inclusion Criteria

The research included only studies that fitted the following cumulative criteria: (1) ALR, VSR or AV-ASR related; (2) published from 2018 onwards; (3) traditional or end-to-end methods applied to ALR or VSR; (4) presented original research; (5) documents written in English; (6) publicly available; (7) for corpora in the English language; (8) for single view video corpora. Few exceptions were made, to include either historic articles, eminent authors, or articles referenced by previous or posterior readings that did not appear in the search results. The only exclusion criteria were applications and devices-oriented research, for being outside of the scope of this paper.

2.3 Search Strategy

The data used for this study were collected between 14 and 18 May 2022, using Science Direct and Scopus search engines, enabling easy access to abstract reading for exclusion purposes, classifying papers in subject areas, and showing a relevant bulk of papers. Google Scholar and IEEE were also firstly included, however, these comparably underperformed. The terms used in the research were “lip-reading” and “audiovisual automatic speech recognition”. There was a total of 43 screened studies and after more meticulous selection, 23 studies were selected for the final analysis.

3 Results

Figure 1 summarizes as a flow diagram the steps for the gathering and analysis of the source papers for the review. Initially, were Identified 49 studies from the selected sources, and 8 of these papers were duplicated. Two additional records were added to the results, gathered from different queries to the databases. After analyzing each research article’s metadata, namely the title, the keywords, and the abstract, 9 of such studies were eliminated from the analysis due to the lack of relation to computer-based lip-reading. The full text of the remaining 34 articles was evaluated considering the aforementioned inclusion criteria, and therefore, 13 articles were also dismissed. The remaining 21 papers were fully assessed.

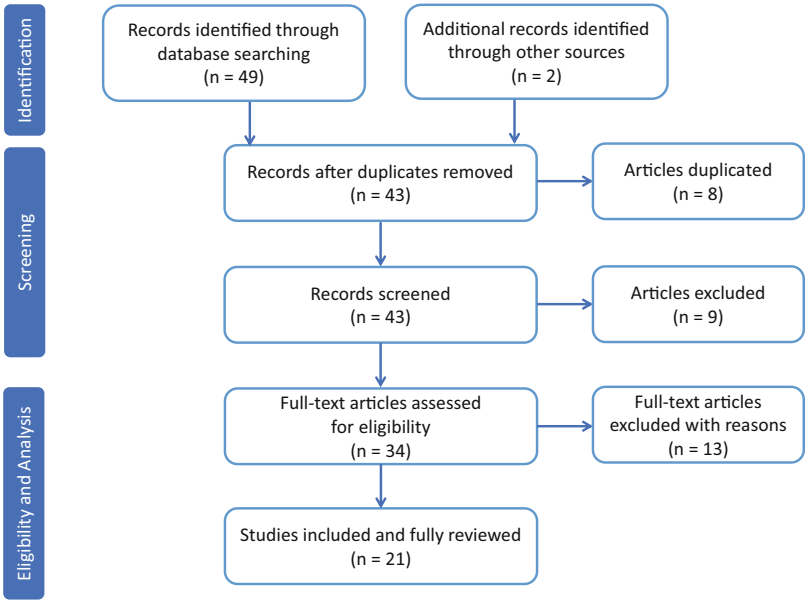

Fig. 1. Flow diagram depicting the selection of the papers.

Nowadays, with the deep learning techniques and the vast databases publicly available, some works are showing promising results. This section presents the current state of the art from the most recent studies and methods on ALR. As a criteria for the presentation of the studies, these were grouped in terms of focus of the methods, that is, these were grouped considering the characteristics that are intended to be obtained with each specific method (lip detection, features, homophemes, visemes), followed by which the specific objective (extraction of words, words and sentences, characters and sentences, speech).

Also, each of the following papers will have a description of their relevant characteristics, followed by a critical analysis/discussion from the authors.

Lopez et al. [15] aimed to study the upper bound of visual-only speech recognition in controlled conditions. Since the literature is not clear on who are the better lip-readers, the authors compared the lip-reading abilities of 15 normal-hearing and 9 hearing-impaired people. A database was constructed, and the speakers were instructed to facilitate lip-reading. Another study was to compare the performances of human and VSR systems, under optimal and directly comparable conditions. In the authors' tests, hearing-impaired participants just nearly outperformed normal-hearing participants. When comparing humans' performance to visual-only automatic systems, a 44% to 20% spoken message decoding decrease gap was observed. However, similar performances were obtained in terms of phonemes, suggesting that the difference between automatic and human speech-reading could be more influenced by the context than the ability to interpret mouth appearance.

Rethinking the RoI for ALR is the aim of Zhang et al. [16]. This paper questions the standard RoI for ALR papers, as human lip-readers do not just look at the lips, during a conversation. [15] states that facial expressions help to decode the spoken message, and context framing the speech (e.g.: a sad expression augments the probability of sad-related words/sentences). Using state-of-the-art VSR models at word level and sentence level, a thorough study is presented assessing the effects of extraoral information, including the upper part and the entire face, the cheeks, and the mouth. The author's proposed model [16] was trained on large-scale "in-the-wild" VSR datasets, depicting many real-world variables, such as variable lighting conditions, cluttering backgrounds, different human poses, expressions, accents, speaking manners, etc. According to the study, using crop augmentation (similar to drop pixels) with face-aligned inputs can produce stronger features, improving recognition by also making the model learn fewer evident extraoral cues from the data.

Using only the mouth entrances makes VSR an isolated problem, as the process will not consider other parts of the human face. Therefore, there has been no consensus on choosing RoIs, and intuition still plays a large role in RoI cropping.

Lu et al. [17] also propose a lip-segmentation method, but now in the framework of the maximum a posteriori Markov random field (MAP-MRF), a statistical segmentation method that considers the spatial relations between image pixels into account. The proposed method sets up a multi-layer hierarchical model, in which each pixel of each layer corresponds to the four nodes in a quad-tree structure (QTS). The probability of a branch node can be derived from the probability of the previous one, throughout the tree structure. Then a Markov random field derived from the model is obtained, so the unsupervised segmentation is formulated as a labeling optimization problem. The method also proposes a variable weight segmentation approach, to improve the robustness of over-segmentation.

Results show that the proposed method has better performance than the related methods, however, it runs between 3 and 4 s, therefore it is not suitable for real-time applications. Lip segmentation accuracy plays an important role in automatic lip-reading and can directly affect the recognition rate.

Lu and Liu [18] propose a localized active contour model-based method, using two initial contours in combined color space: a rhombus as the initial contour of a closed mouth; a combined semi-ellipse as the initial contours of both outer and inner lip boundaries for an open mouth. The method first applies illumination equalization to RGB images to reduce interference of uneven illumination, then adopts a combined color space, which involves the U component in the CIE-LUV color space and the sum components of the Discrete Hartley Transform (DHT). Finally, the shape of initial contours is determined, due to the positions of four key points in the combined color space.

The method improves segmentation results and gets more similar to the true lip boundary, compared with using a circle as the initial contour to segment grey images and images in combined color space.

Das et al. [2] show a refinement in automatic lip contour extraction, using pixel-based segmentation. This embodies an alternative to pixels classification of different color planes, a potential difficulty to lip contours detection in adverse conditions like variations in illumination and clothing. The mouth region is extracted, by k-means clustering binary classification based on Red/Green ratio thresholding. To avoid false detection, a big connected region around the center of the cropped image is considered to be the RoI. Next, the upper and lower lip areas are detected by k-means clustering algorithm binary classification, of Green plane and to weighted RGB plane respectively. The combined lip area is further processed to detect the centrally located big connected region. By finding the centrally located big connected region, rather than the biggest connected region in the whole binary classified image, variations in illumination and clothing effects are overcome and RoI is restricted around the mouth region. For smooth edges, piece-wise polynomial fitting is employed with a higher degree for the upper lip and a lower degree for the lower lip.

The proposed method works well, even for images with varying illumination and clothing effects. A future aim is to use this algorithm to obtain the best possible lip contour.

Visual Speech Recognition (VSR) is highly influenced by the selection of visual features, which can be categorized into static (geometrically based) and dynamic (motion-based). Radha et al. [19] propose a three-viseme model study, one as the control group and two considering both categories, one fused at the features level and the other fused at the model level. For dynamic-motion features extraction, Motion History Image (MHI) is calculated from all the visemes, from which Discrete Cosine Transform (DCT), Wavelet, and Zernike coefficients are extracted. For static-geometric features extraction, an Active Shape Model (ASM) is used. Fusion models are individually built by Gaussian Mixture Model Left-to-Right Hidden Markov Model (GMM L-R HMM).

The results show an improvement in performance due to the fusion, and the presence of complementary cues in the motion-based and geometric-based features, and that geometric cues provide better discrimination of visemes.

Weng and Kitani [20] experiment on word-level visual lipreading from video input with no audio, by replacing shallow 3DCNN + a deep 2DCNN with deep 3DCNN two-stream Inflated Convolution Networks (I3D), and evaluating different combinations of front-end and back-end modules, with the greyscale video and optical flow inputs on the LRW dataset, as presented in Fig. 2. 3D convolution networks can capture the short-term dynamics and be advantageous in visual lipreading, even when RNNs are deployed for the back-end. However, due to the huge number of parameters introduced by the 3D kernels, state-of-the-art methods in lipreading have only explored the shallow (under 3 layers) 3DCNNs. To explore these networks to their maximum, the authors present the first word-level lipreading pipeline using deep (over 3 layers) 3DCNNs.

The experiments show that: compared to the shallow 3D CNNs + deep 2D CNNs front-end, the deep 3D CNNs front-end with two-round pre-training on the large-scale image and video datasets can improve the classification accuracy;

using the optical flow input alone can achieve comparable performance as using the greyscale video as input; the two-stream network using both the greyscale video and optical flow inputs can further improve the performance.

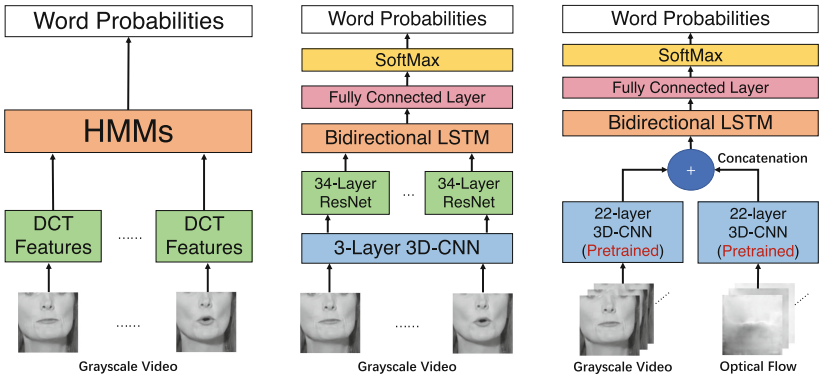

Fig. 2. The architecture of the framework proposed by [20].

Lu and Yan [21] propose a CNN and bidirectional LSTM (BLSTM) that uses hybrid neural network architecture for an automatic lip-reading system. The method first extracts key frames from each isolated video clip, uses five key points to locate the mouth region, extracts features from raw mouth images using an eight-layered CNN, and uses BLSTM to capture the correlation of sequential information among frame features in both directions in time, and uses the softmax layer to predict final recognition result. The limited number of key points reduces redundant information in consecutive frames, therefore the complexity of computation and processing. The CNN copes with image deformation, by translation, rotation, and distortion, hence strengthening the robustness and fault-tolerant capability, and a fully connected layer is used to get static features of a single mouth image. BLSTM improves both finding and exploiting long-time dependencies, from sequential data, so the relationship of the features among frames is built and strengthened.

The results show that the proposed DNN can effectively predict words from the mouth area, on a self-made database (6 speakers, 9 digits), compared to traditional algorithms that combine handcrafted features with a classification model.

Mesbah et al. [22] aimed to the development of a visual-only speech recognition system, proposing Hahn Convolutional Neural Network (HCNN), seizing their ability to represent images with less redundancy, and to be parameterized to retain the global or local characteristics of the image in the lowest orders. The proposed architecture consists of Hahn moments as a filter in the first layer, with its ability to hold and extract the most useful information in images effectively, and the performance of the CNNs in learning patterns and image classification.

The results show a reduction in processing time, normal to spatio-temporal modeling features, and visual features extraction with 3D CNN, ResNet, and Bidirectional LSTM.

Ma et al. [23] focus specifically on lip feature extraction under variant lighting conditions, since research has been mainly conducted for ideal conditions, therefore ideal lighting. The method consists of a pre-processing chain of illumination normalization and also improved local binary patterns (LBP) features. The first is applied to remove the influence of external illumination noise before the lip feature extraction in four steps: median filtering, gamma correction, multi-scale Retinex filtering, and contrast equalization. LBP is an illumination invariant descriptor of edges, which improves the recognition rate of lip-reading under variant lighting conditions.

Experiments show that the proposed algorithm has a lower recognition rate in natural than traditional pixel-based feature extraction method, but higher under variant lighting conditions.

Jeon et al. [24] address the homophemes as word ambiguity enablers, and words under 0,02s as "a", "an", "eight", and "bin", as they do not provide sufficient visual information to learn from. A novel lipreading architecture is presented, combining three different CNNs: 3D CNN, to efficiently extract features from consecutive frames; densely connected 3D CNN - to fully utilize the features; and multi-layer feature fusion 3D CNN with a pixel dropout layer and spatial dropout layer - to avoid overfitting and to extract shapes with strong spatial correlations with fine movements, while exploring the context information both in temporal and spatial domains. Then follows a two-layer bi-directional gated recurrent unit (GRU). The network was trained using connectionist temporal classification (CTC).

The results of the proposed architecture show character (5,681%) and word (11,282%) error rate reductions, for the unseen-speaker dataset, even when visual ambiguity arises.

Wang [25] also addresses the homophemes question, accumulating diverse lip appearances and motion patterns among the speakers, by capturing the nuances between words and different speakers' different styles respectively. As for the front-end, the method utilizes 2D (spatial only) and 3D (spatio-temporal) ConvNets to extract both frame-wise spatial fine-grained and short-term medium-grained spatio-temporal features, to capture both grained patterns of each word and various conditions in speaker identity, lighting conditions, and so on. Then fuses the different granularity features with an adaptive mask (bidirectional ConvLSTM, augmented with temporal attention, which aggregates spatio-temporal information in the entire input sequence), to obtain discriminative representations for words with similar phonemes, as a multi-grained spatio-temporal novel modeling of the speaking process, as depicted in Fig. 3.

The proposed model demonstrates state-of-the-art performance on two challenging lip-reading datasets. In future work, the authors propose to simplify the front-end and extract multi-grained features with a more lightweight structure.

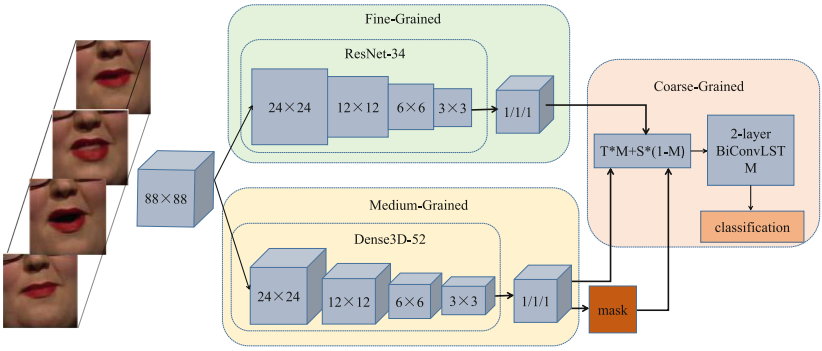

Fig. 3. The architecture of the framework proposed by [25].

Viseme-based lip-reading systems do not require pre-trained lexicons and can be used to classify both unknown words and different languages. Fenghour et al. [26] explore this fact to classify visemes in continuous speech, uses visemes as a classification schema for reading sentences, and use perplexity analysis for visemes to word conversion, stating that all contributions improve sentence-level lip reading. The proposed method uses visemes as a very limited number of classes, a unique deep learning model for classification, and perplexity analysis for recognized visemes to possible word conversion, resorting to purely visual cues from the LRS2 dataset and being robust to varying levels of lighting.

Results demonstrate a significant improvement in the classification accuracy of words compared to state-of-the-art works. For future research, the authors hint towards a more suitable architecture to further enhance the generalization capability and a higher training/test number of samples ratio.

Fenghour et al. [27] focus on viseme-based lipreading systems has been well suited to decoding videos of people uttering entire sentences. As the paper points out, the high classification accuracy of visemes (e.g., over 90%) contrasts with a comparatively low classification accuracy of words (e.g., just over 60%), due to the homovisemes phenomenon which leads to a one-to-many problem (e.g., "I Love You" = "Olive Juice" = "Elephant Shoes"). Aiming for a more efficient viseme-to-word conversion method to tackle this accuracy decline, the authors developed a DNN model with an Attention-based Gated Recurrent Unit and compared it against three other approaches (Perplexity-Iterator, Feed-Forward Neural Network, and Hidden Markov Model) through the LRS2 and LRS3 corpora.

Results show that the proposed model is effective at discriminating between words sharing visemes that are either semantically or syntactically different, and at modeling long and short-term dependencies, therefore being robust to incorrectly classified visemes.

Intending to learn strong models that recognize speech in silent videos, Prajwal et al. [9] focus on challenges in lip reading and propose tailored solutions, contributing to lip movement representations aggregation, robustness improve-

ment to ambiguity by sub-word units based modeling, and to a Visual Speech Detection (VSD) model proposal. The paper proposes an end-to-end trainable attention-based pooling mechanism that learns to track and aggregate the lip movement representations, a sub-word (word-pieces) tokenization that not only matches with multiple adjacent frames but also with those which are semantically meaningful for learning a language easily, therefore greatly reducing the run-time and memory requirements, and a model for VSD trained on top of the lip-reading network since there is no automated procedure for cropping out the clips where the person is speaking.

The results show state-of-the-art Word Error Rate (WER), outperforming work trained on public data, even industrial models trained on orders of magnitude more data. Also, the designed Visual Speech Detection obtains state-of-the-art results, on this task and even outperforms audio-visual baselines.

Martinez et al. [28] address the limitations of the Bidirectional Gated Recurrent Unit (BGRU) and propose corresponding improvement proposals. First, the mouth region was extracted, and DCT was used to feature transform and then fed to HMM for modeling of the temporal dynamics. To address the limitations of the model and the authors proposed: that to improve the overall performance, BGRU layers are replaced with Temporal Convolutional Networks (TCN); to reduce training time (from 3 to 1-week GPU-time), and avoid relying on a cumbersome 3-stage sequential training, a cosine scheduler was adopted; to improve the generalization capabilities, variable-length augmentation was proposed. As each TCN receptive field is defined by kernel and stride sizes, several temporal convolutional blocks are achieved and stacked sequentially to act as a deep feature sequence encoder. Next, a dense layer is applied to each time-indexed feature vector, and a simple averaging consensus function is used. With different-sized kernels and multiple temporal scales, long and short-term information can be mixed up during the feature encoding.

Results on the largest publicly available datasets for isolated word recognition in English and Mandarin, LRW and LRW1000, show that a new state-of-the-art performance was achieved.

Huang et al. [5] propose a novel lip reading model using a transformer network, to achieve higher accuracy. The method makes use of the pre-trained neural network VGG16 to extract the lip features from the GRID corpus, adopts dimensionality reduction towards the originally high dimensions extracted features, and processes the features through the author's proposed Transformer network for training. The transformer adopts a self-attention mechanism instead of CNN and RNN, as is commonly used in deep learning. RNN tends to be slow in some sequential processing tasks. On the other hand, transformers' parallel processing greatly improves training speed.

The experiment shows a significant reduction in training costs, without compromising the enhancement of the lip-reading accuracy of the model.

In [29], the authors tackle the difficulty of meeting the requirements of practical applications for ALR, due to the complexity of image processing, hard-to-train classification, and long-term recognition processes, in three steps. Firstly

they extract keyframes from their own established independent database. Secondly, they use the Visual Geometry Group of Oxford University and the Google DeepMind (VGG) network to extract the lip image features. Then, as an attention-based RNN, they compare two lip-reading models: a fusion model with an attention mechanism; and a fusion model of two networks.

The results of the proposed hybrid neural network architecture of CNN and attention-based LSTM, show an increase of 3.3% to the general CNN-RNN. The authors manifested the future intention to train the model on datasets of real-time broadcast videos.

"No Data, no Deep Learning", is a common hearing among AI researchers. Petridis et al. [7], focus on lip-reading for isolated word recognition training on small-scale datasets. The proposed method consists of two streams (each consisting of an encoder and a BLSTM): one stream encodes static information, using raw mouth RoIs as input; the other stream encodes local temporal dynamics, taking as input the difference between two consecutive frames. Each stream's temporal dynamics are modeled by a BLSTM, and stream fusion is done by another BLSTM. Four benchmark datasets were used, before the usage of very large lip-reading datasets.

The proposed method learns simultaneously to extract features and perform classification using LSTM networks. Results demonstrate that the proposed model achieves state-of-the-art performance, outperforming all other approaches reported in the literature, on all datasets.

Afouras et al. [30] aim to boost lip reading performance, by training strong models learning from ASR strong models, and not requiring human-annotated ground truth data. The proposed method distills (transfers knowledge/weights from a large model to a smaller one) from an ASR model, trained on a large-scale audio-only unlabelled corpus, with a teacher-student approach (the teacher's prediction is used to train the student). The cross-modal distillation combines CTC with a frame-wise cross-entropy loss, minimizing the KL-divergence between the student and teacher posterior distributions. The method and paper's contributions show that: ground truth transcriptions are not essential to train a lip-reading system; arbitrary amounts of unlabelled video data can be leveraged to improve performance; distillation significantly speeds up training; state-of-the-art results on (publicly available) LRS2 and LRS3 datasets can be obtained.

Results demonstrate effectiveness in training strong models for VSR by distilling knowledge from a pre-trained ASR model, and more generally from any available video of talking heads, e.g. from YouTube, therefore from any arbitrarily large amount of data.

Deep Learning methods have been used for developing ALR systems. As DL is vulnerable to adversarial attacks, so will ALR DL-based systems. Gupta et al. [31] proposed Fooling AuTomAtic Lip Reading (FATALRead), a method to perform adversarial attacks on state-of-the-art word-level ALR systems, conducted on a publicly available dataset, in view of making model design more robust and resilient against engineered attacks. Adversarial attacks toward video classification consist of adding a well-crafted minimal and imperceptible perturbation to the input, such that its classification is incorrect. The proposed model aims to

replace the target output for another, by adding perturbations that alter the classification prediction.

FATALRead attacked successfully fools state-of-the-art ALR systems based on sequential and temporal convolutional architectures. The results show the vulnerability of the sequential and temporal convolutional network (TCN) architectures, to an adversarial attack in the domain of ALR.

Table 1 presents a resumed overview of the previously mentioned methods.

Table 1. Summary of the study analysis for Automatic Lipreading.

Study	Year of publication	Location	Focus	Method
Lopez et al. [15]	2017	Pompeu University, Barcelona, Spain	Study upper limit in Speech Recognition	Constructed database, compared hearing-impaired and non-hearing-impaired performances.
Zhang et al. [16]	2020	UCAS, Beijing, China	Rethinking the RoI - extraoral relevance	Study done with word-level and sentence-level VSR models, including extraoral parts, trained on in-the-wild dataset.
Lu et al. [17]	2019	NCUT, Beijing, China	Lip segmentation improvement	Each pixel of each layer is QTS structured, the probability of a branch is derived, a MAP-MRF is obtained, then the unsupervised segmentation turns in labelling optimization.
Lu and Liu [18]	2018	NCUT, Beijing, China	Lip segmentation improvement	Active contour model based, with a rhombus and a semi-ellipse as initial contours. Illumination equalization to RGB images, then combination of U (CIE-LUV) and DHT, resulting on 4 key points to adjust initial shape.
Das et al. [2]	2017	NIT, Nagaland, India	Lip contour extraction refinement	Pixel-based segmentation, mouth region k-means extracted on R/G ratio threshold, and refinement by binary classification of G and RGB planes, and by processing of combined lip partial areas.
Radha et al. [19]	2020	Chennai, India	Static and dynamic (s&d) visual features selection	Three viseme models comparison: one as control; one fusing s&d at features level; one fusing s&d at model level. MHI with DCT, Wavelet, and Zernike. Fusion models built by GMM L-R HMM.
Weng and Kitani [20]	2019	Carnegie Mellon, Pittsburgh, USA	Visual clues only word-level lip-reading	Deep 3DCNN two-stream I3D, instead of a shallow 3DCNN and a deep 2DCNN. Different combinations of front-end and back-end, greyscale, optical flow on LRW dataset.
Lu and Yan [21]	2020	NCUT, Beijing, China	Comparing DNN with traditional algorithms	Extraction of 5 key frames, then 5 key points of mouth region, CNN to extract features and BLSTM to extract both directions correlations.
Mesbah et al. [22]	2019	USMBA, Fez, Morocco	Visual-only speech recognition system	Hahn moments as a filter, and CNN as classifier.
Ma et al. [23]	2017	HIT, Shenzhen, China	Feature extraction under variant lighting	Pre-processing chain of illumination normalization, and improved LBP features.

(*continued*)

Table 1. (*continued*)

Study	Year of publication	Location	Focus	Method
Jeon et al. [24]	2021	GIST, Korea	Homophemes disambiguation; Character and word recognition	New architecture combining 3DCNN, densely connected 3DCNN, multilayer feature fusion 3DCNN, then GRU and CTC. Pixel dropout layer and spatial dropout layer.
Wang [25]	2019	UCAS, Beijing, China	Homophemes disambiguation capturing multi-grained spatiotemporal features	2D and 3D CNNs as front-end, for frame-wise fine spatial and short-term medium spatio-temporal features. Then fuses by Bi-ConvLSTM, with temporal attention.
Fenghour et al. [26]	2020	LSBU, London, UK	Visemes for unknown words in speech	Lexicon-free visemes as a very-limited classification schema and perplexity analysis for word conversion, using LSR2 and visual clues only.
Fenghour et al. [27]	2021	LSBU, London, UK	Viseme-to-word conversion	DNN with Attention-based GRU, compared to Perplexity-iterator, FFN and HMM, using LSR2 and LSR3.
Prajwal et al. [9]	2021	Oxford, UK	Learn strong models of ALR in silent videos	End-to-end trainable attention-based pooling mechanism, sub-word tokenization, and propose a visual speech detection mechanism.
Martinez et al. [27]	2020	SAIRS, Cambridge, UK	Limitations of BGRU	BGRU layers are replaced with multi-scale TCN, reducing 3 weeks to 1 week GPU-time, on LRW and LRW1000 datasets.
Huang et al. [28]	2022	SWUN, Chengdu, China	End-to-end model	Transformer network adopting a self-attention mechanism instead of CNN and RNN, on GRID corpora.
Lu and Li [29]	2019	NCUT, Beijing, China	Requirements of practical ALR applications	Extract key frames, use VGG for features extraction and compare a fusion model with attention mechanism to a fusion model of two networks. Apply in own independent database.
Petridis et al. [7]	2019	Imperial College London, UK	Isolated word recognition, with small-scale datasets	Two streams consisting of an encoder and a BLSTM, one encoding static information, the other local temporal dynamics. Four datasets are used.
Afouras et al. [30]	2020	Oxford, UK	Dismissing human-labelled data	Transfer of knowledge (weights) from large model to a smaller one, teacher-student approach. Cross-modal distillation, combining CTC with a frame-wise cross-entropy loss function.
Gupta et al. [31]	2021	IIT Indore, India	Robustness to adversarial attacks	Adversarial attacks on sequential and temporal convolutional architectures based ALR systems, and on publicly available datasets.

4 Conclusions

In many of the referenced papers, statements reinforce the idea of a lack of consensus, which only acts as a catalyst for more research and more work. The evo-

lution has been so swift that decade-old methods are most likely to be obsolete, and derivations are so diversified that it is as unpredictable what will happen in the next decade as the movement of a simple magnetic pendulum subjected just to 3 magnetic fields beside the gravitational one.

Although uncertain, in the author's point of view the future will be based on: continuing the application of newer methods, followed by the simplification of the same; simplifying parts of the process, as attention narrows inputs' processing target; complicating parts of the process, weighing other data as visual or emotional contextualization.

Answering the Research Questions: RQ1 - End-to-end deep learning, resorting to Attention-based LSTM or Transformers appear to be more suitable for visual clues for automatic lip-reading.; RQ2 - The same answer as in RQ1.; RQ3 - No specific methods are studied with the available datasets; Although some datasets are more commonly explored in the presented papers, namely LRS2; RQ4 - As mentioned, many challenges are still open, e.g., how effective is recurrent training, resorting to parallel and in-series training?

Acknowledgements. This work is funded by FCT/MEC through national funds and, when applicable, co-funded by the FEDER-PT2020 partnership agreement under the project UIDB/00308/2020.

References

1. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, Upper Saddle River (2001)
2. Das, S.K., Nandakishor, S., Pati, D.: Automatic lip contour extraction using pixel-based segmentation and piece-wise polynomial fitting. In: 2017 14th IEEE India Council International Conference (INDICON), Roorkee. IEEE, pp. 1–5 (2017). https://ieeexplore.ieee.org/document/8487538/
3. Bauman, N.: Speechreading (Lip-Reading) (2011). https://hearinglosshelp.com/blog/speechreading-lip-reading/
4. Petajan, E.D.: Automatic lipreading to enhance speech recognition. In: Degree of Doctor of Philosophy in Electrica l Engineering, University of Illinois, Urbana-Champaign (1984)
5. Huang, H., et al.: A novel machine lip reading model. Procedia Comput. Sci. **199**, 1432–1437 (2022). https://linkinghub.elsevier.com/retrieve/pii/S187705092200182X
6. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: LipNet: end-to-end sentence-level lipreading (2016). arXiv:1611.01599
7. Petridis, S., Wang, Y., Ma, P., Li, Z., Pantic, M.: End-to-end visual speech recognition for small-scale datasets (2019). arXiv Version Number: 4. https://arxiv.org/abs/1904.01954
8. Fung, I., Mak, B.: End-to-end low-resource lip-reading with maxout Cnn and Lstm. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB. IEEE, pp. 2511–2515 (2018). https://ieeexplore.ieee.org/document/8462280/

9. Prajwal, K.R., Afouras, T., Zisserman, A.: Sub-word level lip reading with visual attention (2021). arXiv:2110.07603
10. Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P.: Deep learning-based automated lip-reading: a survey. IEEE Access, **9** 121184–121205 (2021). https://ieeexplore.ieee.org/document/9522117/
11. Hao, M., Mamut, M., Ubul, K.: A survey of lipreading methods based on deep learning. In: 2020 2nd International Conference on Image Processing and Machine Vision, Bangkok Thailand. ACM, pp. 31–39 (2020). https://dl.acm.org/doi/10.1145/3421558.3421563
12. Alam, M., Samad, M., Vidyaratne, L., Glandon, A., Iftekharuddin, K.: Survey on deep neural networks in speech and vision systems. Neurocomputing **417**, 302–321 (2020). https://linkinghub.elsevier.com/retrieve/pii/S0925231220311619
13. Bhaskar, S., Thasleema, T.M., Rajesh, R.: A survey on different visual speech recognition techniques. In: Nagabhushan, P., Guru, D.S., Shekar, B.H., Kumar, Y.H.S. (eds.) Data Analytics and Learning. LNNS, vol. 43, pp. 307–316. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-2514-4_26
14. Fernandez-Lopez, A., Sukno, F.M.: Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput. **78**, 53–72 (2018). https://linkinghub.elsevier.com/retrieve/pii/S0262885618301276
15. Fernandez-Lopez, A., Martinez, O., Sukno, F.M.: Towards estimating the upper bound of visual-speech recognition: the visual lip-reading feasibility database. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, USA. IEEE, pp. 208–215 (2017). http://ieeexplore.ieee.org/document/7961743/
16. Zhang, Y., Yang, S., Xiao, J., Shan, S., Chen, X.: Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition (2020). arXiv Version Number: 2. https://arxiv.org/abs/2003.03206
17. Lu, Y., Zhu, X., Xiao, K.: Unsupervised lip segmentation based on quad-tree MRF framework in wavelet domain. Measurement **141**, 95–101 (2019). https://linkinghub.elsevier.com/retrieve/pii/S0263224119302180
18. Lu, Y., Liu, Q.: Lip segmentation using automatic selected initial contours based on localized active contour model. EURASIP J. Image Video Process. **2018**(1), 7 (2018). https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-017-0243-9
19. Radha, N., Shahina, A., Khan, N.: Visual speech recognition using fusion of motion and geometric features. Procedia Comput. Sci. **171**, 924–933 (2020). https://linkinghub.elsevier.com/retrieve/pii/S1877050920310760
20. Weng, X., Kitani, K.: Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading (2019). arXiv:1905.02540. http://arxiv.org/abs/1905.02540
21. Lu, Y., Yan, J.: automatic lip reading using convolution neural network and bidirectional long short-term memory. Int. J. Pattern Recog. Artif. Intell. **34**(01), 2054003 (2020). https://www.worldscientific.com/doi/abs/10.1142/S0218001420540038
22. Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H., Daoudi, M.: Lip reading with Hahn convolutional neural networks. Image Vis. Comput. **88**, 76–83 (2019). https://linkinghub.elsevier.com/retrieve/pii/S0262885619300605
23. Ma, X., Zhang, H., Li, Y.: Feature extraction method for lip-reading under variant lighting conditions. In: Proceedings of the 9th International Conference on Machine Learning and Computing, Singapore. ACM, pp. 320–326 (2017). https://dl.acm.org/doi/10.1145/3055635.3056576

24. Jeon, S., Elsharkawy, A., Kim, M.S.: Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors **22**(1), 72 (2021). https://www.mdpi.com/1424-8220/22/1/72
25. Wang, C.: Multi-grained spatio-temporal modeling for lip-reading. arXiv Version Number: 2 (2019). https://arxiv.org/abs/1908.11618
26. Fenghour, S., Chen, D., Guo, K., Xiao, P.: Lip reading sentences using deep learning with only visual cues. IEEE Access, **8**, 215 516–215 530 (2020). https://ieeexplore.ieee.org/document/9272286/
27. Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P.: An effective conversion of visemes to words for high-performance automatic lipreading. Sensors **21**(23), 7890 (2021). https://www.mdpi.com/1424-8220/21/23/7890
28. Martinez, B., Ma, P., Petridis, S., Pantic, M.: Lipreading using temporal convolutional networks. arXiv Version Number: 1 (2020). https://arxiv.org/abs/2001.08702
29. Lu, Y., Li, H.: Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Appl. Sci. **9**(8), 1599 (2019). https://www.mdpi.com/2076-3417/9/8/1599
30. Afouras, T., Chung, J.S., Zisserman, A.: ASR is all you need: cross-modal distillation for lip reading (2020). arXiv:1911.12747 [cs, eess]. http://arxiv.org/abs/1911.12747
31. Gupta, A.K., Gupta, P., Rahtu, E.: FATALRead - fooling visual speech recognition models: put words on lips. Appl. Intell. (2021). https://link.springer.com/10.1007/s10489-021-02846-w