electronics

MDPI

Review

Data-Driven Advancements in Lip Motion Analysis: A Review

Shad Torrie, Andrew Sumsion, Dah-Jye Lee * and Zheng Sun

Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT 84602, USA; st392@student.byu.edu (S.T.); andreww9@student.byu.edu (A.S.); zsun2@student.byu.edu (Z.S.)
* Correspondence: djlee@byu.edu

Abstract: This work reviews the dataset-driven advancements that have occurred in the area of lip motion analysis, particularly visual lip-reading and visual lip motion authentication, in the deep learning era. We provide an analysis of datasets and their usage, creation, and associated challenges. Future research can utilize this work as a guide for selecting appropriate datasets and as a source of insights for creating new and innovative datasets. Large and varied datasets are vital to a successful deep learning system. There have been many incredible advancements made in these fields due to larger datasets. There are indications that even larger, more varied datasets would result in further improvement upon existing systems. We highlight the datasets that brought about the progression in lip-reading systems from digit- to word-level lip-reading, and then from word- to sentence-level lip-reading. Through an in-depth analysis of lip-reading system results, we show that datasets with large amounts of diversity increase results immensely. We then discuss the next step for lip-reading systems to move from sentence- to dialogue-level lip-reading and emphasize that new datasets are required to make this transition possible. We then explore lip motion authentication datasets. While lip motion authentication has been well researched, it is not very unified on a particular implementation, and there is no benchmark dataset to compare the various methods. As was seen in the lip-reading analysis, large, diverse datasets are required to evaluate the robustness and accuracy of new methods attempted by researchers. These large datasets have pushed the work in the visual lip-reading realm. Due to the lack of large, diverse, and publicly accessible datasets, visual lip motion authentication research has struggled to validate results and real-world applications. A new benchmark dataset is required to unify the studies in this area such that they can be compared to previous methods as well as validate new methods more effectively.

Keywords: lip reading; machine vision; biometrics; datasets; deep learning

Citation: Torrie, S.; Sumsion, A.; Lee, D.-J.; Sun, Z. Data-Driven Advancements in Lip Motion Analysis: A Review. *Electronics* **2023**, *12*, 4698. https://doi.org/10.3390/electronics12224698

Academic Editor: Luca Mesin

Received: 11 October 2023
Revised: 16 November 2023
Accepted: 16 November 2023
Published: 18 November 2023

1. Introduction

Deep learning has revolutionized the computer vision and natural language processing (NLP) worlds. Advancements such as AlexNet [1], VGG [2], and ResNet [3] in the early 2010s ushered in a new era of computer vision. The introduction of attention [4] and the transformer architecture [5] brought about huge leaps in the NLP world. The two worlds of computer vision and natural language processing are joined together in vision-based lip analysis. There have been vast amounts of research dedicated to applying these network architectures to lip-reading and lip authentication tasks. While reviewing these two emerging fields, we recognized the crucial role that datasets play in their potential, progression, and advancements. Thus, in this work, we focus on the data-driven advancements in lip motion analysis. Our comprehensive analysis of data-driven advancements in these fields aims to support future research in the selection, usage, and creation of datasets.

Automated lip-reading (ALR) is an exciting area of research. These systems take videos of individuals speaking and, without audio, attempt to predict what was spoken. There are many promising future applications of ALR systems, such as computer text input, accessibility tools, and speech therapy. This work will highlight the progress that has occurred in this field with a large emphasis on the datasets that have enabled this progress.

Particularly, we emphasize the progress from digit-/letter- to word- and sentence-level ALR systems and the datasets that enabled these large steps. We propose that the natural next frontier for ALR systems is dialogue-level ALR. This is the natural step forward to improve accuracy and real-world applicability.

Lip motion analysis has proven to be a highly successful authentication method compared to other biometric authenticators. This entails authenticating a user based on the motion their lips make in a video. This adds a large benefit over other biometric authentication methods because it acts as a liveness detection as well as an awareness detection layer of security. However, there is a dataset issue that will be discussed that is holding back the lip motion authentication research community.

This work contributes the following:

1. A comprehensive overview, analysis, and comparison of ALR methods and datasets and results on said datasets. We show that larger and more comprehensive datasets result in immense improvements in performance.
2. We propose the next frontier for ALR systems is to move toward dialogue-level ALR. To our knowledge, no works have attempted or mentioned dialogue-level ALR systems or datasets. We propose that this will be the next step to increase accuracy and approach real-world applicability. This transition must be preceded by datasets to be used for training and testing.
3. We provide a comprehensive overview, analysis, and comparison of lip motion authentication methods and datasets. We show that the datasets in the literature for lip motion authentication are severely lacking in size, variation, and accessibility.
4. We identify the need for a large open access lip motion authentication dataset to be used as a benchmark dataset. When compared with the large open access dataset in the ALR research, the lip motion authentication research field is severely lacking. There is no large benchmark dataset to compare these methods, which prohibits comparison and growth. This type of dataset is vital for further improvements in this area.

2. Background

There are many areas of research that stem from the connections between facial image(s), authentication, audio, and text, as can be seen in Figure 1. On the conjunction between text and audio, we see NLP tasks, speech-to-text systems, and text-to-speech systems, which are seen commonly in human computer interfaces today. These are areas that have already found commonplace uses in our daily lives with digital assistants and voice-to-text dictation on our smart devices. Sequence-to-sequence models, such as the LSTM [6] and transformer [5] architectures, have been found to be very successful in these areas.

On the intersections between lip images and text, we have lip-reading and text-based lip synthesis. Lip-reading takes sequential images of an individual speaking and attempts to interpret what the person spoke without the audio signal. We will be covering lip-reading in depth in this work. Text-based lip synthesis research focuses on taking text and an image or images of an individual and generating images that make the individual appear to utter the text given [7–15]. Similar to text-based lip synthesis is audio-based lip synthesis. Systems take as input an audio stream of a person speaking and image(s) of the individual that they would like to appear to speak the given audio track [16–19]. These applications can give interesting insight when determining future directions for other lip-reading and authentication systems but will not be covered in this work.

Analysis of lip motion can also be utilized to improve or even create audio signals of the individual speaking. This research is found on the conjunction from lip images to audio in Figure 1. Adeel et al. utilized an LSTM [6] lip-reading model in parallel with an enhanced, visually-derived Wiener filter (EVWF) [20] to enhance the audio of a speaker, such as removing background noise and fortifying the speaker's voice [21]. They found that utilizing lip-reading results in cleaner audio compared to conventional audio enhancers

that do not utilize visual lip-reading methods. Creating audio of individuals speaking is an even more difficult task [22–27] that also gives interesting insights into lip-reading systems.

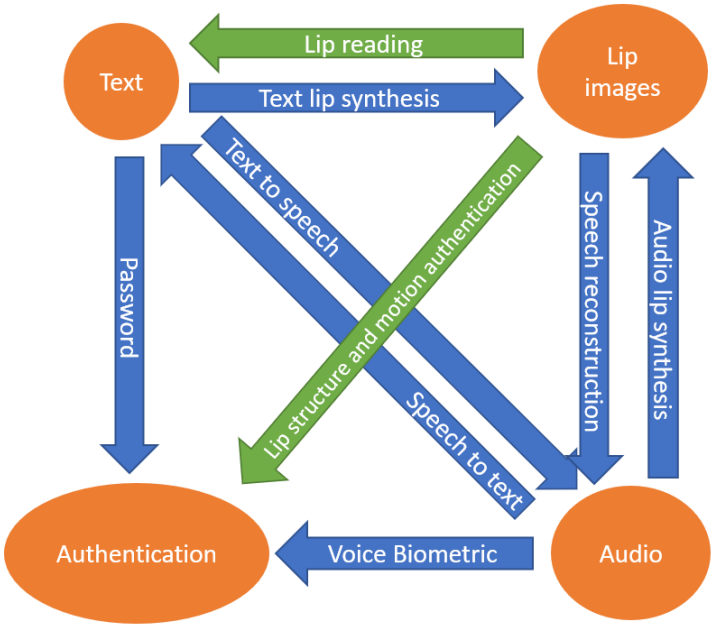

Figure 1. The many research areas connecting the modalities of text, authentication, lip images, and audio. This work provides an in-depth analysis of the two junctions highlighted in green: lip motion authentication and visual lip-reading. The other junctions between these modalities are not covered in this work and are mentioned only briefly.

Dating back to 1976, researchers found that humans utilize vision to aid understanding while talking with each other, even in easy-to-hear circumstances [28]. This is an area that is researched and has proven useful. Audio–visual speech recognition (AVSR) is a thoroughly researched area as well. Many of the datasets and research that we reviewed can be employed for both VSR and AVSR systems. Adding the visual aspect has slightly improved audio speech recognition systems [29]. This area has been well researched, and because of the already high accuracy that audio-only speech recognition systems achieve (being as high as 98.7% [29]), visual-only speech recognition is the more challenging problem and thus where larger improvements can occur. Therefore, this is where we will focus our review.

An automated visual Russian sign language recognition pipeline was created [30]. This research discovered that adding a lip-reading module to analyze the lip movements of deaf individuals improved accuracy. This indicates that lip motion is a viable way to extract information from the speaker.

Lip-reading is not a new idea. Deaf individuals have been utilizing the visual aspects of speech for centuries to understand spoken language. It has been determined that deaf individuals can only distinguish 20–40% of speech [31–35]. A lip-reading experiment was conducted on eighty-four normal hearing undergraduate students [36]. These students were tested on 25 sentences. The results were very interesting. The average sentence success rate was 12.41%, with two outliers achieving around 30%. This is disheartening for those that hope to teach a neural network to lip-read when average people cannot even lip read very well.

Human lip-reading accuracy was compared to machine lip-reading accuracy on word, phoneme (audibly distinct sounds), and viseme (visually distinct sounds) levels [37]. On the word level, the machine performed worse than a person, with 3.75% compared to 18.42%. The machine, however, performed significantly better on phonemes and visemes. On phonemes, humans achieved on average 31.6%, and the machine achieved 80.27%. On

visemes, humans achieved 35.4%, and machines achieved 91.6%. Ref. [37] attributes the machine's deficiency in word-level accuracy to the fact that it lacks previous knowledge of language and words. Since their study took place in 2009, their machine model was not a neural network. Modern neural networks are able to learn the previous knowledge and expectations that humans have and thus lead to better results on the word level. Their results, however, show that automated lip-reading systems have the potential to match and outperform humans.

There are numerous automatic lip-reading (ALR) review articles in the literature. Sooraj et al. conducted a review of the general structure of ALR systems and how they work [38]. Oghbaie similarly reviewed the general structure of ALR systems and determined areas for future advancements in the systems' structures [39]. Agrawal et al. conducted a literature review that focused on the methods by which the networks are trained, particularly comparing the prepossessing and pretraining methods that benefit ALR systems [40]. These reviews did not focus much attention on datasets and their role in ALR systems. Hao et al. similarly conducted a review of ALR technology [41]. They highlighted current ALR systems and specifically pointed out the difficulties due to visual factors such as lighting, background, and variation of speakers' appearances. They concluded that larger, more varied datasets are required to advance ALR systems.

Fernandez et al. analyzed the advances over a 20 year period in automatic lip-reading (ALR) [42]. They found that since the introduction of deep learning, there has been an increased interest in ALR systems as seen by the number of publications on the subject over time, shown in Figure 2. They additionally determined that deep learning models performed very similarly to traditional approaches in simpler lip-reading tasks, such as single digits and letters, but that deep learning approaches widely outperformed traditional approaches on more complex tasks, such as word- and sentence-level lip-reading. Deep learning ALR systems were able to master digit- and letter-level lip-reading quickly, thus overtaking the conventional methods. The next task was word-level lip-reading, which, as can be seen in Table 1, has seen steady improvement over the years and is seeing very good results. Following word-level lip-reading, the next step is sentence-level lip-reading, which is much more challenging, however, steady progress has been made in sentence-level lip-reading research. The improvements and the limitations in the sentence-level research give insight into the following steps for the natural progression of ALR datasets and systems.

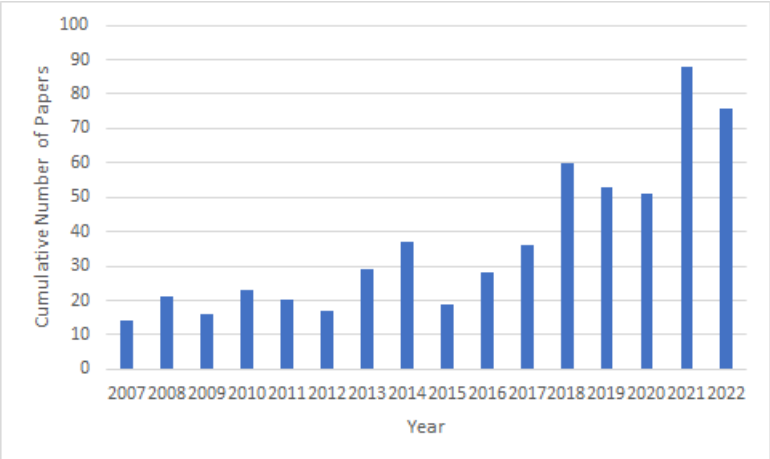

Figure 2. The cumulative number of papers on ALR systems published between 2007 and 2022. This indicates the increase in research and interest in the area of visual lip-reading systems over the years and validates the necessity for this work as a review of the progress made in this field due to dataset availability. The conceptual foundation for our diagram is rooted in the work of Fernandez et al. [42]. Inspired by their approach, we developed an updated version of the figure, drawing on more recent data obtained from Google Scholar. This updated diagram aims to provide a contemporary perspective, reflecting the latest trends in the field.

Table 1. Progression of state-of-the-art lip-reading on the Lip-Reading in the Wild (LRW) dataset [43]. Illustrates the steady progression made on large benchmark datasets. Neural network architectures are listed for convenience and are not discussed at length in this work. A large portion of the table is provided by [44].

Method			Data		LRW
Author's Name (Year)	**Frontend**	**Backend**	**Input Image Size**	**Input**	**Accuracy (%)**
Chung et al. (2016) [43]	3D and VGG M	-	112 × 112	Mouth	61.10
Son et al. (2017) [45]	3D and VGG M version	LSTM & Attention	120 × 120	Mouth	76.20
Petridis et al. (2018) [46]	3D and ResNet-34 M	Bi-GRU	96 × 96	Mouth	82.00
Stafylakis et al. (2017) [47]	3D and ResNet-34	Bi-LSTM	112 × 112	Mouth	83.00
Cheng et al. (2020) [48]	3D and ResNet-18	Bi-GRU	88 × 88	Mouth and 3D augmentations	83.20
Wang et al. (2019) [49]	ResNet-34 and DenseNet3D-52	Bi-LSTM	88 × 88	Mouth	83.34
Courtney et al. (2019) [50]	Alternating ResNet Bi-LSTM	Alternating ResNet Bi-LSTM	64 × 64	Mouth	83.40
Luo et al. (2020) [51]	3D and 2-Stream ResNet-18	Bi-GRU	88 × 88	Mouth and gradient policy	83.50
Weng et al. (2019) [52]	3D and 2-Stream ResNet-18	Bi-LSTM	112 × 112	Mouth and optical flow	84.07
Xiao et al. (2020) [53]	3D and 2-Stream ResNet-18	Bi-GRU	88 × 88	Mouth and deformation flow	84.13
Zhao et al. (2020) [54]	3D and ResNet-18	Bi-GRU	88 × 88	Mouth and mutual information	84.41
Zhang et al. (2020) [55]	3D and ResNet-18	Bi-GRU	112 × 112	Mouth (aligned)	85.02
Feng et al. (2020) [56]	3D and SE ResNet-18	Bi-GRU	88 × 88	Mouth (aligned) & augmentations	85.00
Martinez et al. (2020) [57]	3D and ResNet-18	MS-TCN	88 × 88	Mouth (aligned)	85.30
Ren et al. (2021) [58]	ResNet-18	TM-Seq2Seq&KD [59]	N/A	Mouth (aligned)	85.7
Tsourounis et al. (2021) [44]	ALSOS and ResNet-18 blocks	MS-TCN	88 × 88	Mouth (aligned)	87.01
Peng et al. (2022) [60]	3D and ResNet-18 blocks	MS-TCN	88 × 88	Mouth (aligned)	87.7
Ma et al. (2021) [61]	3D and ResNet-18	MS-TCN	88 × 88	Mouth (aligned)	88.50
Koumparoulis et al. (2022) [62]	EfficientNetV2-L	TCN & Transformer	88 × 88	Mouth (aligned)	89.52
Ma et al. (2022) [63]	3D and ResNet-18	DC-TCN	88 × 88	Mouth (aligned)	91.6

Fenghour et al. similarly reviewed the past 24 years and determined that deep learning models performed better on word- and sentence-level lip-reading tasks compared to conventional methods [64]. They particularly pointed out the issue of classification for ALR systems due to the large possible lexicon, meaning that the dataset the network is trained on cannot cover every possible word in the language; thus, the network must learn how to recognize and dictate words that it has never encountered. They determine that this is a large hurdle that lip-reading systems need to overcome to reach higher accuracies.

Pu et al. very recently conducted a literature review of ALR systems [65]. They determined a few advancements that need to be made to improve ALR systems. It was determined that a standard for viseme classification would help standardize the results found by these systems and help direct future research. They recognized that the exploration of useful lip regions has led to improved results and impressed a need for further research in

that area. It was determined that larger unconstrained datasets are needed, especially for the Mandarin Chinese language.

In this paper, we will discuss the need for larger, more varied datasets similar to [41,65], with the added evaluation of the advancement of scope that ALR systems are run in. We propose that ALR systems are ready for an expanded scope from sentence-level lip-reading to dialogue-level lip-reading. This expanded scope will be brought about by new datasets made for this previously unexplored level of lip-reading.

With the breakthroughs in ALR systems thanks to deep learning, another area of research was poised to benefit from these improvements: that of lip motion authentication. There are many authentication methods in our modern world. Conventional text passwords have been the default authentication method for years, but biometrics have gained recent popularity with the rise of mobile electronics. Facial [66], fingerprint, and voice authentication have particularly gained attention in recent years. Research has found that many facial authentication systems can be easily fooled by images, projections of images onto 3D heads, and 3D silicone masks [67]. Lip motion authentication comes in two forms to solve this problem. One is to identify an individual based on their unique way of moving their lips, and the other is to have the individual enroll a lip motion passphrase.

Facial, fingerprint, and voice biometric systems are heavily researched, and therefore, there are immense benchmark datasets to evaluate new methods created [68,69]. Lip motion authentication is not quite as popular and thus lacks large varied benchmark datasets to evaluate systems on. This results in a difficult problem when attempting to compare various methods, as well as determine future steps to improve on what has been done, which we attempt to do in this survey.

Chowdhury et al. conducted a survey on lip biometrics [70]. They covered static lip biometrics as well as lip motion biometrics. They came to the conclusion that most solutions do not address unconstrained scenarios, which limits the applications for this technology. We reiterate this point and add that there is a large dataset issue when it comes to comparing various techniques to perform lip motion authentication.

3. Automated Lip-Reading

Automated lip-reading (ALR), also known as visual speech recognition (VSR), is a challenging problem to solve. The datasets consist of silent video clips of people talking. The goal of these systems is to determine what is being said by the person by just analyzing their facial/lip movements. This has a wide range of challenges to overcome, as well as a wide range of applications. The limitation to visual input alone in ALR results in an issue when it comes to visemes, which are speech sounds that are distinguishable audibly yet visually identical. As one would expect, these cause many difficulties for vision-only ALR systems. Word-level datasets [43,71] tend to sidestep this problem by avoiding words that are visemes of each other, which yields good results, as seen in Table 1. This "solution", however, weakens the implications of the results and indicates the lack of robustness of these systems.

Another solution to this issue is to do sentence- or phrase-level ALR rather than word-level ALR. This enables the network to be able to distinguish between visemes based on the context of the words before and after it. This is a preferable method to address the viseme issue because it correlates much more to the real-world applications of ALR technology and makes for much more robust systems. Sentence-level lip-reading, however, brings on a whole new host of challenges for networks, namely determining the beginning and end of a word, as well as the timing between them.

3.1. Word-Level ALR

There has been much research in the scope of word-level vision-based automated lip-reading. There are many datasets and networks that have been used to show competent lip-reading solutions. Table 2 lists the more popular word-level ALR datasets and their various statistics.

3.1.1. CMU AVPFV

The CMU AVPFV dataset [71] contains videos collected from 10 subjects, each reciting the same 150 words 10 times each. The videos were collected from two angles, frontal and profile. Utilizing a hidden Markov model (HMM), they were able to achieve ∼45% accuracy when analyzing the profile view, ∼32% when analyzing the frontal view, and ∼49% by analyzing the profile and frontal views combined. These results are not too impressive when compared to the state-of-the-art accuracies achieved on more varied datasets, such as LRW with deep learning, but the results are informative to camera positioning for future datasets and solutions such as OuluVS2. This dataset is no longer used as a metric for ALR systems due to the small amounts of videos and limited variation as compared to newer, diverse datasets.

3.1.2. NDUTAVSC

Chitu et al. created a Dutch dataset (which many papers incorrectly label as German) called NDUTAVSC [72]. The data were collected in a lab with set lighting, camera position, and head position for all individuals. It contains both word and sentence videos. This dataset and the solutions evaluated on it predated the deep learning revolution. Rothkrantz et al. achieved the state-of-the-art result on the NDUTAVSC dataset with an accuracy of 84.27% [73]. This dataset is not used as a metric for many ALR models due to the fact that it is in Dutch and due to the lack of variation within the dataset.

3.1.3. AVAS

The AVAS dataset [74] is an Arabic lip-reading dataset collected in a lab, but the subjects were required to record their utterances on two separate days. This requirement introduced variation due to changes in mood, make-up, levels of attentiveness, etc. They also varied light in the lab, as well as the head position. The dataset contains 35 words and 12 phrases in the Arabic language that each individual repeated multiple times. The state-of-the-art results were achieved by the creators of the dataset by utilizing the k-nearest neighbor algorithm with an accuracy of 85% [75]. Due to this dataset's less commonly spoken language, its limited variation, and its creation before deep learning was utilized, it has not become a benchmark dataset for new ALR systems.

3.1.4. MIRACL-VC1

The MIRACL-VC1 [76,77] dataset was collected in a controlled environment. This dataset contains 10 words and 10 phrases. This limitation to words and phrases allows for an easy way to compare subject-dependent and subject-independent methods of ALR. The dataset was collected with a Microsoft Kinect sensor [78], thus allowing the collection of RGB-D data. The extra depth information was proven as beneficial when comparing subject-independent lip-reading, which is likely because different people have different facial structures; thus, the extra depth information enabled the SVM (support vector machine) model to normalize better. Using the SVM model, they were able to achieve respectable results of 96.4% (79.2% on phrases and 63.1% on words). These results are reasonable, especially in the pre-deep-learning era. Using an LSTM network, Parekh et al. were able to achieve 98% accuracy on the MIRACL-VC1 dataset [79]. This demonstrates the growth and improvements that deep learning systems bring over conventional systems.

3.1.5. AusTalk

The AusTalk [80] dataset was collected in a controlled environment. Each subject was given a script to read for half of the videos, and the other half are spontaneous speech. This dataset actually contains digits, words, and sentences. It contains the largest amount of videos compared to other word-level datasets, as indicated in Table 2. Despite it being the largest word-level dataset that we highlight, it is limited due to the controlled environment. and it also has added complexity due to it containing digits, words, and sentences. It

thus has not become a benchmark dataset for new research methods. The state-of-the-art accuracy on the AusTalk dataset is 69.18%, achieved by [81].

Table 2. List of commonly used word-level ALR datasets, the number of speakers, classes, and video counts, and current state-of-the-art results. It is important to acknowledge that the comparison of state-of-the-art (SOTA) accuracies between datasets can be challenging. The accuracies presented are intended to illustrate potential areas for improvement in the field rather than be used for direct comparison between datasets. This perspective recognizes the inherent variability and specific conditions of each test environment, emphasizing that these metrics are indicative benchmarks rather than definitive comparisons. Data retrieved from [41,42,64] and individually cited papers. * Accuracy percent from combined dataset with words and phrases.

Dataset	Year	Language	Speakers	Classes	Total Videos	Environment	State of the Art Accuracy
CMU AVPFV [71]	2007	English	10	150	15,000	Lab	∼49% [71]
NDUTAVSC [72]	2010	Dutch	66	6907	6907	Lab	84.27% [73]
AVAS [74]	2013	Arabic	50	48	13,850	Varied Lab	85% [74]
MIRACL-VC1 [76]	2014	English	15	10	1500	Lab	98% [79] *
AusTalk [80]	2014	English	1000	966	966,000	Lab	69.18% [81]
LRW [43]	2016	English	1000	500	400,000	Wild	91.6% [63]
LRW-1000 [82]	2019	Mandarin	2000	1000	718,018	Wild	57.5% [83]

3.1.6. LRW

Chung et al. collected and labeled a dataset named lip-reading in the wild (LRW) from BBC television network recordings [43]. This is one of the first lip-reading datasets that utilizes a large public media database to allow for much more real-world variation in the data. This type of data is called "wild" data because it was not collected under constrained conditions. This idea of collecting datasets from large television and other media databases is very important to lip-reading datasets. To do so, they developed a pipeline for large-scale data collection from TV broadcasts. They also created a CNN architecture trained on LRW. When collecting their dataset, they aligned the text with the video timestamp using a Penn Phonetics Lab Forced Aligner and IBM's Watson speech-to-text service. Then, to determine which face is speaking in the frame, they ran a landmark tracker, found the distance between the center top point of the lips and the center bottom point of the lips, and took the Fourier transform of this distance over time data. A linear SVM classifier was trained on the frequency spectrum that distinguishes between a face that is speaking and one that is not. These methods were utilized to extract a word-level dataset that contains the words that occurred the most in the TV broadcasts. These words are between 5 to 10 syllables.

This dataset is frequently used as a benchmark for researchers to evaluate lip-reading solutions. The creators of this dataset went on to create the LRS family of datasets, which are arguably the best sentence-level datasets currently publicly available. The current state-of-the-art result is 91.6% [63]. Even higher accuracy has been achieved by supplementing the pre-training dataset with the LRS2-BBC, LRS3, and AVspeech datasets (discussed in Section 3.2), resulting in an accuracy of 92.1%. The creators even further improved upon this by supplying their model with the word-boundary data supplied with the LRW dataset. This resulted in an accuracy of 94.1%. This is not listed as the state-of-the-art result in Tables 1 and 2 because of the addition of data, which detracts from a fair comparison of models. This, however, is indicative that future advancements can be made by supplementing the training datasets. This dataset is one of the largest datasets collected for word-level ALR, as well as one of the most varied; for this reason, it is a very popular choice as a benchmark of word-level ALR systems.

Because LRW is the most commonly used benchmark dataset for word-level ALR systems, the history of the state-of-the-art results on this dataset are very indicative of the overall growth of the word-level lip-reading research as a whole. Table 1 shows the progression that has occurred across the LRW since its creation and initial evaluation in

2016. As we can see, initially, there was a large 15.1% jump in accuracy with the addition of the LSTM and attention methods by [84], and another large 5.8% jump with the utilization of the bi-directional GRU [85] model by [46]. After this advancement, the accuracy grows at a much slower rate, as is expected when so many methods have been tried and tested. The increase in results here is very promising and indicates that there is much progress yet to be made in lip-reading for sentence-level solutions and beyond.

3.1.7. LRW-1000

A large Mandarin Chinese lip-reading dataset named LRW-1000 was generated in 2019 [82]. The data were extracted from TV programs in China. The creators of this dataset used iFLYREC to retrieve time-aligned sentences and separate out different speakers [86]. They manually annotated who was speaking, and then used the landmark SeetaFaceEngine2 detector [87,88] to locate the full face and compare it to manually annotated coordinates. A kernelized correlation filter was used to ensure that the sequence retrieved the same face. Audio-to-video synchronization was carried out with the SyncNet described in [8]. This dataset contains one of the largest vocabularies of any of the word-level lip-reading datasets; this and the fact that it is a very recent dataset has resulted in it being a good benchmark to evaluate ALR models on. The state-of-the-art result is 57.5% accuracy [83]; the same model resulted in a much higher accuracy on the LRW dataset at 88.5% [43]. This indicates that the LRW-1000 dataset is much more difficult and could possibly be a good metric for how lip-reading systems perform. The language difference is likely also a large factor in the difficulty difference.

3.1.8. Methods

There are various methods used for the collection of said datasets. This is a crucial difference in datasets. Many use static, controlled environments to collect data, e.g., individuals will go into a well-lit room and repeat 100 words 10 times each. Datasets in such controlled environments were necessary for the early development of ALR systems to ensure first that it was possible to distinguish motions in controlled environments. The CMU AVPFV [71], NDUTAVSC [72], AVAS [74], AusTalk [80], and MIRACL-VC1 [76,77] datasets are such datasets collected in controlled environments but were useful for initial proof-of-concepts. The more real-world method of collecting a dataset is to extract image sequences from large collections of videos, such as TV programs. This results in more variation in head poses, lighting, resolution of the lip area, etc. The LRW [43] and LRW-1000 [82] datasets are examples of in-the-wild datasets collected in similar manners. Because of their increased difficulty, as well as increased alignment with real-world systems, in-the-wild datasets drive developments much more than controlled datasets.

Word-level ALR datasets are generally set up in such a way that the networks trained on them need only classify which word a single sample is. This simplifies the training process and methodology to a simple classification problem. These large benchmark datasets mentioned thus unify the research work conducted to a common goal and method. The methods by which classification is carried out with the image sequence as inputs vary, as they should, but the actual method of lip-reading is unified. This is crucial for comparison and advancements of various neural network architectures and training methodologies. As this work's primary focus is data-driven advancements, we will not be analyzing neural network architectures or training methodologies.

3.1.9. Findings

Word-level lip-reading has progressed immensely since the deep learning revolution. The datasets and their corresponding ALR systems have proven that machines can distinguish between spoken words even better than humans can. The limitation of vocabulary that most of these datasets utilized resulted in an easier problem to solve in comparison with sentence-level ALR systems. Word-level ALR systems are stepping stones toward

sentence-level ALR systems. The lessons learned on these datasets have been applied to sentence-level lip-reading with encouraging results.

As can be seen in Table 2, the LRW dataset and LRW-1000 dataset are the only datasets collected in the wild. As such, these two datasets have become the typical benchmarks used when new word-level lip-reading systems are trained and tested. The LRW dataset tends to be used more often, which is likely due to the fact that the language spoken in the dataset corpus is English.

3.2. Sentence-Level ALR

Sentence-level lip-reading brings its own types of challenges and advantages over word-level lip-reading. For one, the context from the other words around a word makes it more easily distinguished from visemes and visually similar words. However, a large challenge presents itself when thinking about the differences in the structure of a sentence in text versus a sentence in a video. Text contains spaces that naturally indicate the separation of words, but there is no such indication in a video of a person talking. In the NLP world, a similar issue occurs in translation from one language to another due to the fact that not all translations are word-for-word translations. In the NLP world, this is overcome by utilizing bi-directional sequence-to-sequence networks, such as bi-GRUs and transformers, so that the new information can directly improve previous outputs. A similar approach is taken with sentence-level ALR systems. Table 3 describes some of the more commonly used sentence lip-reading datasets and their various details, as well as the latest state-of-the-art results.

Each sentence-level ALR dataset contributes something different to the research field. They each have the benefits as benchmark systems or as stepping stones for future datasets. Below, we will go into further detail about each dataset, its use, and its contribution to the ALR community.

Table 3. List of commonly used sentence-level ALR datasets, their number of speakers, classes, and video counts, and current state-of-the-art results. Data retrieved from [41,42,64,89] and individually cited papers. * Seems to have a different definition of utterances. ** Exact number not recorded. *** Used in training to increase accuracy on LRS3-TED.

Dataset	Year	Language	Availability	Speakers	Vocab	Utterances	Hours	Env.	SOTA Accuracy
GRID [90]	2006	English	Public	34	51	1 k	37.5	Lab	98.7% [91]
OuluVS2 [92]	2015	English	Public	52	550	1560	N/A	Lab	98.31% [93]
MODALITY [94]	2017	English	Public	35	182	231	31	Lab	54.00% [94]
LRS [45]	2017	English	Retired	1000+ **	17 K	118 K	246	Wild	49.8% [45]
MV-LRS [84]	2017	English	Retired	1000+ **	15 K	504 K	155	Wild	47.2% [84]
LRS2-BBC [95]	2018	English	Public	1000+ **	18 K	144 K	224	Wild	64.8% [96]
LRS3-TED [97]	2018	English	Public	∼10 K	17 K	165 K	437	Wild	63.7% [98]
GRID-Lombard [99]	2018	English	Public	55	51	5.4 k	N/A	Lab	N/A
LSVSR [100]	2018	English	Private	∼464 K	127 K	2.9 M	3.9 k	Wild	59.1% [100]
CMLR [101]	2019	Mandarin	Public	11	3.5 k	102 k	N/A	Wild	67.52% [101]
YTDEV18 [102]	2019	English	Private	N/A	N/A	20 k *	31 k	Wild	N/A ***
SynthVSR [103]	2023	English	Private	N/A	N/A	N/A	3k	Gen.	N/A ***

3.2.1. Grid

The GRID dataset [90] contains sentences in the following form: verb, color, preposition, letter, digit, and adverb. For example, "Lift blue to A 3 fast". This restriction of sentence structure and vocabulary makes the GRID dataset easier to learn than other more in-the-wild datasets [43,45,95,97,101]. Due to these constraints, evaluation results on this dataset tend to be much higher, with the state-of-the-art being 98.7% [91] compared to that of LRS2-BBC being 66.5% [95]. The results on this dataset also indicate that with proper context and understanding of language, ALR systems can obtain very encouraging results. The GRID dataset is often used as a metric for sentence ALR systems; it is, however, important to keep in mind that the results are not as real-world indicative as other datasets' results.

Alghamdi et al. collected an expanded version of the GRID dataset named GRID-Lombard [99]. They added faces from varying poses, similar to [92]. This dataset has not been evaluated on ALR methods as other datasets have but has been found useful for audio-to-video generation evaluation [25].

3.2.2. OuluVS2

Anina et al. collected a dataset titled OuluVS2 [92]. This dataset was recorded with multiple cameras mounted at five different angles, including frontal, profile, 30°, 45°, and 60° views. Naturally, the dataset was recorded in a controlled environment to be able to obtain such viewpoints of individuals. The number of speakers and the number of classes are also very low in this dataset, containing 52 individuals with each individual repeating 10 phrases (repeated 3 times), which are the same phrases as in the less popular OuluVS [104] dataset: a 10-digit sequence (repeated three times) and 10 sentences (recorded only once per subject). The phrases and digit sequences were the same across individuals, but the sentences differed per individual. Thus, this dataset is much less varied compared to similar datasets that were collected in the wild [45,84,95,97] with varied head pose positions and no limitations on classes.

The results of research on this dataset can, however, be informative for real-world applications of lip-reading when determining the position of the camera or cameras in such systems. See Table 4 for individual angle results from different research. Based on those results, it seems as though which angle is bests is dependent on the network that is being used to achieve the lip-reading. On average, however, the frontal view obtains the highest accuracy. We can also see the bidirectional LSTMs and GRUs result in higher accuracies on this dataset.

Maeda et al. trained their network using both frontal and 90° videos [105]. They then evaluated each camera position individually and found that 30° was the most accurate view with this method. Zimmermann et al. analyzed the results with an LSTM network when using the five different angles individually as well as in pairs [106]. They found that individually, the 30° viewpoint resulted in the highest accuracy and that the frontal view paired with the 30° viewpoint resulted in the best results compared to individual angles as well as paired angles. Petridis et al. further explored these angles by evaluating all possible combinations of the angles provided in the OuluVS2 dataset [107]. They found that the best results are achieved when the frontal, profile, and 45° views are used in parallel to determine what was spoken. They were able to achieve a then-state-of-the-art result on the OuluVS2 dataset of 96.9%.

The current state of the art is 98.31% [93]. The results on this dataset tend to be much higher than those collected in the wild due to the limitation of vocabulary (digit sequences and assigned phrases) as well as the constrained lab scenario the dataset was collected in. This dataset remains a benchmark dataset for those wanting to test networks on specific angles. The results on the OuluVS datasets pair well with results from the LRS line-up of datasets [45,84,95,97] to indicate good performance on specific angles as well as in-the-wild varying angles.

3.2.3. MODALITY

The MODALITY dataset [94] is relatively small compared to other sentence-level ALR datasets. It was collected in a lab under controlled conditions. The videos were recorded on a stereo pair of time-of-flight cameras and thus include depth data for each individual, which is useful for specific research. The size of this dataset, its limited variation, and the focus on depth information, however, led to it not being a commonly used benchmark dataset. Thus, the state-of-the-art results for this dataset were achieved by its creators with an accuracy of 54%. Many succeeding datasets compare themselves to MODALITY to show the progress in amounts of videos, vocabulary, and variety.

Table 4. Comparison of accuracy from different camera angles in the OuluVS2 dataset across different network architectures. This informs the collection and selection of data for future datasets and methods to ensure real-world applicability. Neural network architectures are listed for convenience and are not discussed at length in this work. Data retrieved from [107] and individual papers. * These models are first trained with data from more than one view and then fine-tuned with data from the corresponding view. ** This model was pre-trained on the MV-LRS dataset [84] and then fine-tuned on the OuluVS2 dataset. Bolded results indicate the camera angle on which the given method performs best. DA: data augmentation, LVM: latent variable model; ATT: attention variable models.

Network Architecture + Additions	0°	30°	45°	60°	90°
CNN + DA [108]	**85.6%**	82.5%	82.5%	83.3%	80.3%
End-to-end CNN + LSTM [109]	81.1%	80.0%	76.9%	69.2%	**82.2%**
CNN + LSTM * [109]	82.8%	81.1%	85.0%	83.6%	**86.4%**
PCA + LSTM + GMM-HMM [106]	74.1%	**76.8%**	68.7%	63.7%	63.1%
Raw Pixels + LVM [107]	73.0%	75.0%	**76.0%**	75.0%	70.0%
VGG-M+LSTM+ATT ** [84]	**91.1%**	90.8%	90.0%	90.0%	88.9%
Multi-view 3DCNN * [93]	88.6%	**89.4%**	88.1%	85.6%	83.9%
CNN + Bi-LSTM [110]	90.3%	84.7%	**90.6%**	88.6%	88.6%
CNN + Bi-LSTM * [110]	**95.0%**	93.1%	91.7%	90.6%	90.0%
End-to-end Encoder + Bi-LSTM [107]	**94.7%**	89.7%	90.6%	87.5%	93.1%
3D CNN + SAM + Bi-GRU + Local self-attention + CTC + DA [93]	**98.31%**	97.89%	97.21%	96.78%	97.55%
Average	**86.78%**	85.54%	85.21%	83.08%	84.00

3.2.4. LRS

The Lip-Reading Sentences dataset (LRS) [45] was one of the first in-the-wild sentence-level ALR datasets. It was created by the same team of researchers as the LRW dataset [43], and thus, the methods used to extract the data are practically the same, with slight differences to account for sentences instead of only words. They extracted their videos from the BBC television network's collection of interviews, just like the LRW dataset. These videos offer a large range of individuals with high variation in age, gender, ethnicity, etc. This utilization of a large database of videos introduced much-needed variation and noise that sentence-level ALR systems needed to challenge them and test how robust they can be. During the collection of this dataset, the research team determined that a head pose of any angle between 0° and 30° was acceptable, and thus, the dataset is also more varied in head position compared to other datasets with head pose limitations. The LRS dataset set a new standard for sentence-level datasets in size and variation.

The LRS team had a professional lip-reader attempt to read the lips of the dataset as a comparison for the results. The professional lip-reader achieved a word error rate (WER) of 73.8%, meaning that he correctly annotated 26.2% of words. The initial evaluation carried out by the LRS team utilizing a CNN + LSTM front-end and an LSTM + attention back-end achieved 49.8% accuracy. While these results are not incredibly impressive, they do show that machines can outperform even a professional lip-reader. Due to licensing restrictions, the LRS dataset is not public; thus, the same team quickly followed up by creating the MV-LRS [84], LRS2-BBC [95], and LRS3-TED [97] datasets, as described below.

3.2.5. MV-LRS

The Mulit-View Lip-Reading Sentence (MV-LRS) dataset [84] expands upon the LRS [45] dataset. Instead of only using videos from BBC TV interviews, they also used videos from dramas and factual programs to increase the variety and, likely, the emotional variance in the speakers' faces. They also increased the allowed range of angles for head position to up to 90°. This introduced a lot more variety and difficulty in the dataset. Due to licensing restrictions, the MV-LRS dataset is not public, thus increasing the need for LRS2-BBC and LRS3-TED, as described below. This restriction results in the state-of-the-art results being achieved by the creators of the dataset. They achieved 47.2% accuracy using a VGG-M [111] and LSTM front-end and an LSTM and attention back-end. They also

evaluated this method on the OuluVS2 dataset and found then state-of-the-art results, as seen in Table 4.

3.2.6. LRS2-BBC

Due to the license limitations on the LRS [45] and MV-LRS [84] datasets, the LRS team created the LRS2-BBC [95] and LRS3-TED [97] datasets to supersede the LRS and MV-LRS datasets. The LRS2-BBC dataset was collected in the same manner as the LRS dataset with an expanded video set for the BBC TV programs. As seen in Table 3 the LRS-BBC dataset has more videos and higher vocabulary coverage than the LRS dataset. It has fewer videos than the MV-LRS dataset but higher vocabulary coverage. This dataset has become a benchmark dataset for many new ALR systems thanks to its variation, its large number of videos, and its vocabulary.

Ref. [96] has achieved the state-of-the-art results at 64.8% on the LRS2-BBC dataset. To achieve this level of accuracy, they created a neural network pipeline that first classifies visemes, then detects what words are likely said based on those visemes. The visual front-end is made up of 3D convolution and 2D ResNet; this is followed by the transformer-based viseme classifier. The visemes are then fed into the transformer-based word classifier, which predicts what words are said. This method is very intriguing because visemes are what can be discerned from visual-only systems, so separating it out at that point allows for the network to become very good at viseme classification. Then, the word classifier has to use the context from all the visemes to determine the actual words.

3.2.7. LRS3-TED

The LRS3-TED [97] was collected in the same way as the LRS2-BBC dataset. It was extracted from TED and TEDx talks from YouTube. As seen in Table 3, it is a very large dataset. It does not have as many videos as the MV-LRS [84] dataset, but unlike the MV-LRS dataset, it is publicly available and, therefore, has become a benchmark dataset for many new ALR systems. The current state-of-the-art accuracy is 63.7%, achieved by [98], who utilized a 3DCNN and ResNet-18 front-end and a conformer encoder [112] as the back-end. They found that supplementing their training dataset with other publicly available datasets further improved their system to achieve 80.9%. This is not listed as the state-of-the-art because there is possible overlap between individuals, but it does show that even better results are possible. The conformer encoder is an example of another NLP network that has been found to yield impressive results in ALR systems.

3.2.8. CMLR

The CMLR dataset [102] is a large, highly varied Mandarin Chinese sentence-level ALR dataset collected from Chinese TV programs, much like LRS [45]. It is small when compared to the LRS family of datasets [45,84,95,97] but is useful for evaluating models on other possibly more difficult languages. The state-of-the-art results on this dataset were achieved by the creators, with an accuracy of 67.52%.

3.2.9. LSVSR

Shillingford et al. trained on a large custom dataset named Large-Scale Visual Speech Recognition (LSVSR; see Table 3), which contains 3886 h of content [100]. For reference, the LRS3-TED dataset contains under 500 h. They evaluated their model on LRS3-TED and achieved a then-state-of-the-art result of 53%. These results, however, are very insightful, as they indicate that higher-variation datasets will bring incredible results to the ALR research community.

3.2.10. YTDEV18

Makino et al. created and trained on a YouTube dataset named YTDEV18, which is not publicly available [102]. They did, however, say that it contains 31,000 h of lip-reading footage. They evaluated their network on LRS3-TED [97], which is publicly available, and

achieved a higher-than-state-of-the-art accuracy of 66.4%. This is not listed as the state-of-the-art accuracy in Table 3 due to the fact that this model was not exclusively trained on LRS3-TED. They ensured that the training data in YTDEV18 and the test data in LRS3-TED had no specific video overlap; they could, however have overlapped individuals. They implemented an RNN-T or transducer [113]. They exhibited that using a high-variation dataset (YTDEV18) can result in very good results, as indicated by their LRS3-TED results.

Serdyuk et al. improved upon this work. Using YTDEV18 for training and evaluating the LRS3-TED dataset, they achieved a staggering 74.1% accuracy [114]. This was achieved by implementing the first (to their knowledge) purely transformer-based model for this task.

This team at Google further improved their results by using a conformer encoder [115]. Again, they used the large YTDEV18 dataset for training, as well as an even larger dataset scraped from YouTube for pretraining. They achieved an immense increase in performance with an accuracy of 87.2% on the LRS3-TED dataset.

These examples demonstrate one of the huge issues in visual lip-reading right now. These researchers' models are trained on a dataset that had much more variation, which led to state-of-the-art results, but that dataset, along with its labels, is not accessible to the public and thus is unable to be authenticated as a reliable dataset. Furthermore, it cannot be used in future works to improve results [65].

3.2.11. Other Supplemental Datasets

The landscape of ALR system training has recently been transformed, largely due to the significant accuracy gains achieved through the use of the massive YTDEV18 dataset [115]. However, the exclusivity of this dataset necessitates that other researchers seek alternative supplementary datasets. In this context, the work of Ma et al. [98] is particularly noteworthy. They demonstrated that augmenting the training dataset with automatically transcribed data from the VoxCeleb2 [69] and AVSpeech [116] datasets using open-source speech-to-text software could substantially enhance network performance. Additionally, their use of the LRW [82] and LRS2-BBC [95] datasets further enriched the training process. This comprehensive approach resulted in a leap from 63.7% accuracy without supplemental data to 80.9% accuracy with the addition of approximately 3000 h of extra training data. While these results do not quite match the performance achieved with the 31,000-h YTDEV18 dataset [115], the public availability of these alternative datasets offers a valuable resource for future research, holding the promise of even further advancements in ALR system training.

3.2.12. SynthVSR

Building on the need for supplemental datasets to improve performance, Liu et al. demonstrated that incorporating synthetic data significantly enhances performance in visual lip-reading (VLR) systems [103]. Their study revealed that training solely on the 438 h LRS3-TED training set yielded an accuracy of 63.3%. However, integrating an additional 3652 h of synthetic data boosted this figure to 71.6%. The inclusion of previously mentioned auto-labeled data from the VoxCeleb2 [69] and AVSpeech [116] datasets, amounting to 2630 h, further augmented the accuracy to an impressive 83.1%. This synthetic dataset was meticulously crafted by animating CelebA [117] images with lip movements synchronized to audio samples from the LibriSpeech [118] and TED-LIUM 3 [119] databases using a specially developed speech-driven lip animation model. Regrettably, this extensive 3652 h synthetic dataset is not yet available for public use. The authors point out that the data used to train and create their synthetic data are, however, publicly accessible. The findings of Liu et al. underline the significant role synthetic data can play in elevating the accuracy of VLR systems, showcasing its potential as a rapid and efficient means to enrich training datasets.

3.2.13. Methods

The methods of data collection for sentence-level ALR datasets are very similar to those used for word-level ALR datasets. Some initial datasets were similarly collected in controlled, consistent environments, such as GRID [90], OuluVS2 [92], and MODALITY [94]. These datasets were integral for preliminary sentence-level lip-reading and are often used as a way for researchers to validate new datasets that are released; i.e., when Chung et al. released the LRS dataset [84], they trained a model on the LRS dataset and validated it on the GRID dataset [90] to authenticate their proposed dataset. The other method of collection is extracting videos of individuals speaking from large video libraries. These datasets are termed "in-the-wild" datasets due to their large variety of individuals, vocabulary, lighting, resolution, etc. The common in-the-wild datasets used are LRS [45], MV-LRS [84], LRS2-BBC [95], LRS3-TED [97], LSVSR [100], CMLR [101], and YTDEV18 [102]. This method of dataset extraction has proven very useful to increase accuracy across all datasets. It is also crucial to highlight the variance in speech characteristics across these datasets. Some offer more structured and formal speech contexts, while others encompass a broader and more varied range of speech situations. For a comprehensive comparison of the speech contexts in each dataset, refer to Table 5.

The advancements in synthetic data utilization, as illustrated by Liu et al. [103], signal an imperative for continued research in this area. The YTDEV18 dataset exemplifies the principle that larger datasets typically yield superior results. However, compiling extensive datasets such as YTDEV18 poses significant challenges, often being both resource-intensive and costly. Synthetic data emerges as a promising solution to this dilemma, offering a cost-effective and rapid means of acquiring substantial data volumes. This approach could potentially mitigate the challenges associated with the collection of large-scale datasets, underscoring the necessity for further exploration and development in the field of synthetic data generation and application.

The neural network architecture and training methods vary greatly in the literature. This will not be covered in this work as we focus on the data-driven advancements. It is important to note, however, that the method of input and output to these sentence-level ALR systems is quite consistent. The goal of such a system is to take an image sequence and dictate what the subject is saying based purely on visual information. Because the goal of these systems is aligned, they can be compared by evaluating them on the datasets highlighted.

Table 5. This table delineates key sentence-level audio–visual lip-reading datasets, outlining their release year, language, peak accuracy achieved, and the specific context of data collection. It underscores the variety in dataset environments ranging from controlled, structured settings to diverse, real-world scenarios, including TV programs, formal lectures, and YouTube content. Additionally, it reflects on the innovative use of synthetic data, offering a holistic view for researchers to assess the datasets' relevance and potential applicability in the evolving domain of audio–visual speech recognition.

Dataset	Year	Language	SOTA Accuracy	Speech Scenario
GRID [90]	2006	English	98.7% [91]	Structured sentences
GRID-Lombard [99]	2018	English	N/A	Structured sentences
OuluVS2 [92]	2015	English	98.31% [93]	Controlled sentences
MODALITY [94]	2017	English	54.00% [94]	Controlled sentences
LRS [45]	2017	English	49.8% [45]	TV interviews
MV-LRS [84]	2017	English	47.2% [84]	TV programs
LRS2-BBC [95]	2018	English	64.8% [96]	TV programs
LRS3-TED [97]	2018	English	63.7% [98]	Formal lectures
CMLR [101]	2019	Mandarin	67.52% [101]	TV programs
LSVSR [100]	2018	English	59.1% [100]	YouTube videos
YTDEV18 [102]	2019	English	N/A	YouTube videos
SynthVSR [103]	2023	English	N/A	Synthetic data

3.2.14. Findings

Comparing the datasets by looking at Table 3, we can see that the in-the-wild datasets firstly contain much more data (as indicated by the utterances and hours columns) and are much more difficult for ALR systems to achieve high accuracy on. Due to this difficulty and variation, many previous works choose these datasets as the benchmark to validate their methods. In particular, the LRS3-TED dataset has been used most frequently due to its large hour and utterance counts, its difficulty, and its availability. The LRS family of datasets brought about new standards and challenges to the ALR research community. Due to the medium in which the videos were collected, they contain highly variant videos when it comes to head position, lighting, ethnicity, and vocabulary.

The supplemental training datasets, including the LSVSR, YTDEV18 and SynthVSR datasets, illustrated that there is even more room to grow when it comes to large-scale datasets. These datasets, unfortunately, are not made publicly available and thus cannot become benchmark datasets for future research. This illustrates a need for even larger datasets that are publicly accessible for the research community to make even more progress. The enhancement in performance achieved by incorporating substantial supplemental datasets is a recurring theme in recent research, as evidenced in a range of studies [98,114,115,120,121], including those utilizing synthetic datasets [103]. These findings reflect a broader trend within the field of deep learning, emphasizing the pivotal role of large datasets in augmenting training processes. The significant improvements garnered through the integration of extensive data volumes not only validate the current methodologies but also lay a foundation for future breakthroughs. This trend underscores the ever-increasing importance of data quantity in driving advancements in deep learning, suggesting that further exploration and utilization of large and diverse datasets will continue to be a key factor in propelling the field forward.

With the transition from word- to sentence-level ALR datasets and systems, new solutions were found to address issues such as visemes and provide much-needed context. The results on the sentence-level datasets are still not very high when compared to audio-based dictation or when compared to word-level ALR systems. Therefore, the next logical step to enable even more context and challenges is to expand the context past single sentences. In the case of the LRS3-TED dataset, this could mean using the context from the previous sentence to increase results on the current one. New datasets must also be created to allow for dialogue level ALR systems that take the context from the previous sentence independent of who said it to inform the network on the current sentence.

4. Lip Motion as Authentication

Lip motion authentication is a less-researched yet interesting biometric authentication method that allows an individual to authenticate by repeating a facial password. Due to the lower development of these systems, there are many different ways to achieve authentication with lip motion. This makes it difficult to determine which system is more accurate and to determine what dataset is the best benchmark for these systems. A common metric for these authentication systems is the equal error rate (EER). This is the percentage at which the false-positive and false-negative rates are equal. Thus, if the EER is 5%, then that system will result in 5% false positives as well as 5% false negatives.

4.1. Datasets

Because this area of research is less explored compared to other biometric methods, there are fewer datasets to utilize when attempting a new method for lip motion authentication. There are a few, however, that can prove useful.

4.1.1. XM2VTS

Messer et al. created the XM2VTS dataset, which many lip authentication methods use to validate their methods (see Table 6) [122]. This dataset consists of 295 speakers who repeated the same 3 sentences. The sentences are as follows: "0 1 2 3 4 5 6 7 8 9", "5 0 6 9 2 8 1 3 7 4", and "Joe took fathers green shoe bench out". Each subject repeated the sentences during four separate sessions, which were uniformly spaced out over 5 months. These individuals speaking the same and different sentences are tested against each other to discover how well a system works. This dataset has advantages and disadvantages compared to other datasets. For example, there is a high number of individuals, but the amount of "passwords" is relatively low.

4.1.2. VidTIMIT

Sanderson et al. created a dataset named VidTIMIT [123]. It contains recordings of 43 individuals across 3 separate sessions. The sessions were, on average, 6.5 days apart. Each person was assigned 10 different sentences chosen from the TIMIT speech dataset [124], which were recorded only once, resulting in 430 recordings. While this dataset's variation in sentences is better than that of VM2VTS, it does not contain multiple recordings of each sentence and thus is less useful in authentication settings.

4.1.3. qFace and FAVLIPS

Wright et al. created two datasets that are available upon request [125]. The first dataset is qFace. It was collected to determine if networks trained on XM2VTS would be applicable to the real world. It contains 10 individuals saying 10 different digit sequences 8 times each. The videos were collected on a mobile device. This dataset is limited but fulfills its intended purpose of testing the real-world applicability of the XM2VTS dataset. Using the same network, the authors achieved an equal error rate (EER) of 1.65% on the XM2VTS dataset and an EER of 6.25% on the qFace dataset. This discrepancy is somewhat expected when the XM2VTS dataset's lack of variation is considered.

The second dataset is the FAVLIPS dataset, which contains data collected on a mobile device from 42 individuals over 4 sessions with one month of time between each session. During each session, the individuals would repeat the following: utter ten digits in series, utter a randomized 10-digit sequence (which was the same for all users), subvocalize the same two digit sequences, utter the randomized 10 digit sequence in 3 different lighting conditions, and utter a randomly selected sentence from the TIMIT speech dataset [124].

The FAVLIPS dataset is a useful dataset to evaluate change over time. Table 7 demonstrates the utility of the FAVLIPS dataset. We can see that only training on the XM2VTS dataset does not result in very high accuracies on the FAVLIPS dataset due to the real-world variation. Updating a model with new weights based on the FAVLIPS training data improves the results immensely. Further improvement is seen in some lighting conditions when a network is trained concurrently on the XM2VTS and FAVLIPS datasets. While the FAVLIPS dataset lacks in the subject count compared to XM2VTS, it has more diverse sentences, as well as more diverse lighting and background conditions. The qFace results, as well as the FAVLIPS results, illustrate a need for a large-scale, highly varied dataset for this authentication method. The FAVLIPS dataset is a dataset that could be used by other researchers to determine how robust their lip motion authentication systems are on a more varied dataset.

4.1.4. Existing Datasets

Another method for testing lip motion authentication is by evaluating on an already existing lip-reading dataset. Ref. [126] trained and tested on the OuluVS [104] dataset mentioned previously and found that adding lip motion tracking to a conventional facial authentication system increased accuracy from 83.75% to 93.25%. Ref. [127] similarly evaluated their lip motion authentication method on a portion of the AV Digits lip-reading dataset [128]. These methods of using more varied lip-reading datasets appear to be very viable due to their variation and large size.

4.1.5. Private Datasets

Most of the other lip motion authentication methods utilize datasets collected for their specific task. These are most often not publicly available and therefore will not be covered in detail. This illustrates one of the large issues with lip motion authentication research at present. Each system comes at the problem in a different way.

4.2. Methods

Faraj et al. utilizes lip motion as a form of liveness detection [129]. Their system ensures that the lips have temporal changes in shape, thus indicating if the person is, in fact, there compared to a static image. Similarly, Ref. [130] designed a system that authenticates based on if the person's lips move, as well as their lip structure.

The systems [125,131–134] that utilize the XM2VTS dataset compare individuals speaking the same phrases and distinguish between them by determining how each individual utters the same phrases. Similar approaches are often used by other systems as well. Ref. [135] designed a system that distinguishes between individuals based on how they smile. Ref. [136] designed a system that authenticates based on the physiological characteristics of lip contour movements while a person speaks. Ref. [137] utilizes audio biometrics in addition to lip motion tracking carried out by ultrasonic signals to determine an individual's unique way of speaking audibly as well as physically.

Many systems allow the user to choose their own unique lip motion passphrase or password. One system was designed to track the motion of the entire face for authentication [138]. It allows each individual to select a face or lip motion as their password to authenticate. This is combined with conventional facial authentication to increase security. Ref. [139] introduced a visual passphrase system that separates out each word in the user's phrase or sentence passphrase and creates a unique feature vector for each word, which is compared to the stored feature vectors to authenticate. Refs. [140,141] both designed a system that authenticates a person based on the structure of their lips as well as a lip motion passphrase. Refs. [67,142] utilized acoustic signals to track the lip motion patterns of individuals as they speak their passphrase to authenticate in combination with conventional facial authentication. Refs. [67,143] both similarly used ultrasonic lip motion tracking to authenticate a user when they speak their chosen passphrase. Ref. [144] designed a four-factor authentication system. They used conventional facial authentication, lip-reading to ensure the person said the same phrase, speech authentication to ensure the person sounds the same, and lip motion authentication to ensure that the user moved their lips in the same way as they did for their stored password.

Table 6. Comparison of lip motion authentication systems' architecture, the datasets used, and the results achieved. XM2VTS is the only generally used benchmark dataset, but the number of speakers is still immensely small compared to datasets used for ALR systems. It can be very difficult to compare these systems due to dataset differences, method differences, and metric differences. Neural network architectures are listed for convenience and are not discussed at length in this work. It is important to acknowledge that the comparison of datasets can be challenging when looking at the various metrics. These metrics presented are intended to illustrate potential areas for improvement in the field rather than be used for direct comparison between datasets. This perspective recognizes the inherent variability and specific conditions of each test environment, emphasizing that these metrics are indicative benchmarks rather than definitive comparisons. * Number of videos (number of speakers not provided); ** Motion only results; HMM: hidden Markov model [145]; GMM: Gaussian mixture model; GMVE: generalized minimum volume ellipsoid; DP: dynamic programming matching; LMK: CNN-based landmark detector [146]; EV: eigenvectors; ELM: extreme learning machine; UT: ultrasonic; AL: audio–lip motion tracking; SNN: Siamese neural network.

Paper (Year)	Method(s)	Dataset	Speakers	Network Architecture	Metric	Results
[131] (2000)	Motion	XM2VTS	295	GMVE	EER	14%
[147] (2003)	Motion	M2VTS	36	HMM	EER	19.7%
[148] (2004)	Voice and motion	VidTIMIT [123]	43	GMM	EER	1.0%
[141] (2004)	Motion	Private	40	HMM	EER	5.1%
[149] (2006)	Face and motion	BioID [150]	25	2D-DCT + KNN	Accuracy	86%
[129] (2006)	Voice and motion	XM2VTS	295	GMM	EER	22% **, 2%
[151] (2007)	Voice and motion	XM2VTS	295	GMM	Accuracy	78% **, 98%
[140] (2011)	Structrue and motion	Private	21	DTW	Accuracy	99.5%
[132] (2012)	Voice and motion	XM2VTS	295	GMM	Accuracy	94.7%
[152] (2013)	Motion	Private	43	DP	FAR@ 3% FRR	14.5%
[139] (2014)	Motion	Private [153]	20	KNN + DTW [154]	Accuracy	92.4%
[155] (2015)	Motion	XM2VTS	295	GMM	EER	2.2%
[136] (2017)	Motion	Private	20	GMM	Accuracy	96.2%
[143] (2018)	UT Motion	Private	50	SVM	TNR & TPR	86.7% & 76.7%
[126] (2018)	Face and motion	OuluVS [104]	20	EV + ELM	Accuracy	71% **, 93.25%
[135] (2018)	Face and motion	Other [135,156]	400 & 104	CNN + LSTM	EER	0.37%
[133] (2019)	Motion	XM2VTS	295	STCNN + Bi-GRU	EER	1.03%
[125] (2020)	Motion	XM2VTS	295	SNN	EER	1.65%
[127] (2020)	Motion	AV Digits [128]	39	3DCNN + Bi-LSTM	EER	9%
[67] (2021)	AL, voice, and face	Private	44	CNN + LSTM	EER	5%
[144] (2021)	Voice and motion	Private	240 *	LMK + 3D Resnet	FAR & FRR	0.25% & 18.25%
[137] (2021)	Voice and motion	Private	50	N/A	Accuracy	95.89%
[138] (2021)	Face and motion	Private	10	LMK + RNN + FC	AP	98.8% **
[157] (2022)	Face and motion	Private	10 [138] + 38	LMK + Transformer	AP	94.9% **
[158] (2022)	Face and motion	Private	48 [157] + 11	CNN + Transformer	AP	98.8% **
[134] (2022)	Motion	XM2VTS	295	DWLSTM + GRU	Accuracy	96.78%

4.3. Findings

As illustrated, lip motion authentication research is very fragmented. The datasets that are publicly available are small and very limited when it comes to variation. Chowdhury et al. concluded that the work that has been conducted largely avoids unconstrained scenarios and larger population evaluations [70]. This coincides with what we have determined here. The datasets built for lip motion authentication are small and very con-

strained. While implementing our facial and lip motion authentication system [138,157,158], we found that the real-world unconstrained scenario resulted in a decrease in accuracy and revealed many oversights in our dataset [159]. For these types of authentication methods to succeed, immensely large datasets are required to verify their validity. As shown, there is also a very wide range of implementations for lip motion authentication systems. This has made it difficult to compare the various implementations and thus difficult to gauge the contribution of the research that has been conducted. A new benchmark dataset would unite the lip motion biometric research field to allow for comparisons and reviews that would increase the validity of the research conducted.

Table 7. Write et. al found that the small lip motion authentication benchmark dataset (XM2VTS) was not sufficient for training [125]. They collected the FAVLIPS dataset to validate their lip motion authentication system, which was trained on the XM2VTS dataset. The FAVLIPS dataset contains more variation when it comes to lighting and spaces in time that are very beneficial when analyzing the real-world applicability of such a system. As seen, the XM2VTS dataset is not sufficient for real-world applicability. They then used a section of the FAVLIPS dataset to train the network and saw improvements, as shown. It is apparent that further improvements are needed as the lowest error achieved is only 13.79% even in neutral lighting. All numbers reported are the equal error rate (EER).

	Training Data		
Evaluation Data	**XM2VTS Only**	**Pretrained on XM2VTS; Updated on FAVLIPS**	**XM2VTS + FAVLIPS**
XM2VTS: evaluation set	1.21%	1.95%	5.60%
FAVLIPS: neutral nums	22.43%	13.79%	10.83%
FAVLIPS: light front	28.44%	17.50%	20.54%
FAVLIPS: light side	42.29%	36.67%	30.00%
FAVLIPS: light behind	44.91%	24.17%	29.12%

5. Future Directions

5.1. Automated Lip-Reading

The ALR datasets have driven immense improvements in the literature. There are signs that training on more and more data will continue to improve accuracy [98,103,114,115,120]. The highest accuracies achieved are often from models trained on as much data as possible. Unfortunately, the data they train on are either not verified to be distinct from the validation/test set, or the data they train on are not publicly accessible. This makes it difficult to validate the results and makes it more difficult to progress past the work conducted previously. We propose that there is a need for even larger benchmark datasets that are publicly accessible to drive the research solutions to new capabilities.

As described, we propose the next modality shift in ALR systems is to move from sentence-level lip-reading to dialogue-level lip-reading. A dialogue-level ALR system has the potential to use dialogue-level context to improve results. Such a system would be able to have a broader understanding of the context of given speech and thus work in more cases that are not present in datasets as they stand thus far. This would help systems become more general ALR systems rather than be specific to a particular type of speech, e.g., a model trained on the LRS3-TED dataset [97] is trained only on technical presentation speech and thus would likely perform more poorly in a conversation context or in the context of general speeches. This is an advancement that cannot occur without a large dataset first being collected for future works to train and test on. A dataset that is collected in the wild and contains even more data than the datasets thus far is required to advance past the current systems.

5.2. Lip Motion Authentication

As discussed, the previous systems in lip motion authentication struggle to validate their results due to the lack of an unrestricted, large dataset. We propose that the lip motion authentication research requires such a dataset for advancements to occur and real-world applications to be viable. This dataset should learn from the dataset extraction pipelines used to collect lip-reading datasets, such as LRS3-TED [97]. Such a dataset collected from a large video library would have the much-needed real-world variation in lighting, timing, background, etc.

The other large issue in the current literature of lip motion authentication systems is due to the variations in methods of authentication. Some use an enrolled passphrase to compare against, thus ensuring the subject uttered the same word, phrase, or sentence. Others attempt to train a neural network to recognize a person based on how their lips move in general, thus ensuring the subject is the same because their lips move similarly enough to the enrolled subject. Future work needs to compare these two methods and consider the viability of both. Benchmark datasets for both of these cases would be paramount to determine which of these methods is more viable and accurate.

6. Ethics

As with many applications that neural networks are being used in, there are ethical considerations that need to be addressed. Many of the previous works in datasets and methods do not address the ethical concerns surrounding their work.

6.1. Dataset Collection

A growing industry in the era of deep learning is data collection, labeling, acquisition, and magnetization. The ethics behind the collection and use of this data has been brought into question recently, particularly by the European Union [160]. Many of the datasets collected thus far in the areas of automated lip-reading and lip motion authentication fail to discuss how they are addressing the ethical concerns around the data collected. This is something that is becoming ever more important to the general public. Thus, future works that aim to collect the large-scale datasets required for these tasks should consider and mention the ethical collection of said datasets.

6.2. Dataset Usage

The use of the datasets collected can also bring up many ethical considerations that have not been discussed sufficiently. The neural networks trained with these large, diverse datasets are very intriguing for improving human–computer interfaces and aiding disabled individuals, but they can also be used to invade the privacy of individual's conversations and privacy. Prajwal et al. points out that the limitations of the current datasets, such as resolution and head poses, make it more difficult to use in these situations. It is, however, important to keep this in mind with future datasets and development [120]. As the datasets described and proposed in this work will enable real-world applicability, use, and commercialization, the ethical considerations need to be addressed by those collecting and releasing the datasets, as well as those that use them.

6.3. Ethnic and Dialect Issues

A significant yet under-explored factor impacting the performance of automatic lip-reading (ALR) and lip motion authentication systems is the influence of dialects, accents, and the effects of subjects not natively speaking the language under evaluation. This element becomes crucial in the collection and utilization of datasets for training, as well as in the real-world application of these systems in unconstrained environments. Ethnicity and dialect variations can introduce complexities in accurately interpreting lip movements, which, in turn, can affect the efficacy of these technologies. Furthermore, these considerations raise ethical questions regarding inclusivity and bias, aspects that have received limited attention in prior research on ALR and lip motion authentication. Addressing

these issues is essential to ensure that these technologies are equitable and effective across diverse populations.

7. Conclusions

Lip-reading has progressed immensely in the last twenty years. There have been many advancements in the realms of datasets that have ushered in new advancements in the algorithms. As discussed, there has been a shift in the focus and goal of lip-reading models as advancements have been made, moving from single sounds, to digits, to words, to phrases, and now to sentences. While sentences have much room for improvement, Ref. [36] points out that there is a limit to how many words can be recognized by vision alone. This limit is reduced and partially overcome by sentence-level lip-reading due to the temporal context introduced by the sentence structure [161]. This temporal context can be even further improved by broadening the scope even further from sentences to larger sections of sentences and even to conversations between people.

The natural language processing (NLP) field of research has shown that more context can improve accuracy. Applying current ALR strategies to data throughout a dialogue could allow for ALR systems to gain even more context and result in even higher accuracy levels to further this exciting area of research. To move forward to this next step in increased accuracy and real-world unconstrained scenarios, new datasets are required. We propose such datasets will usher in a new round of advancements in the ALR research community and help these systems come much closer to applicability and generalizability in the real world.

There is also strong evidence, as shown in this work, that the larger the training dataset a system is trained on, the higher the accuracy. We propose the need for even larger, more diverse, publicly accessible benchmark datasets to further the research in the ALR field.

The lip motion as authentication world also needs an overhaul when it comes to datasets. There has not been a dataset similar to the LRS family of datasets that has high amounts of variation and in-the-wild data that also has large amounts of individuals. Due to the limited datasets available, the previous lip motion authentication research works struggle to compare their contributions to previous works and struggle to prove their applicability in unconstrained scenarios. To reach the desired levels of accuracy, applicability, and comparability, new datasets are required for training and evaluating such systems.

This work offers an in-depth analysis of the data-driven advancements that have emerged in the fields of visual lip-reading and visual lip motion authentication. We highlight the significance of large datasets and the potential growth in these fields as new datasets become available. By conducting a comprehensive review of previous advancements facilitated by new benchmark datasets, our objective is to assist future research endeavors in locating the datasets required for their tasks and to foster further dataset creation, thereby facilitating even more data-driven advancements in these exciting fields.

Author Contributions: Conceptualization, S.T. and D.-J.L.; Methodology, S.T.; Validation, S.T.; Formal analysis, S.T., A.S. and Z.S.; Investigation, S.T.; Resources, D.-J.L.; Writing—original draft preparation, S.T.; Writing—review and editing, A.S., D.-J.L. and Z.S.; Visualization, S.T.; Supervision, D.-J.L.; Project Administration, D.-J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
4. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
6. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
7. Oneață, D.; Lőrincz, B.; Stan, A.; Cucu, H. FlexLip: A Controllable Text-to-Lip System. *Sensors* **2022**, *22*, 4104. [CrossRef]
8. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 251–263.
9. Li, L.; Wang, S.; Zhang, Z.; Ding, Y.; Zheng, Y.; Yu, X.; Fan, C. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1911–1920.
10. Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D.B.; Genova, K.; Jin, Z.; Theobalt, C.; Agrawala, M. Text-based editing of talking-head video. *ACM Trans. Graph. (TOG)* **2019**, *38*, 68. [CrossRef]
11. Taylor, S.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A.G.; Hodgins, J.; Matthews, I. A deep learning approach for generalized speech animation. *ACM Trans. Graph. (TOG)* **2017**, *36*, 93. [CrossRef]
12. Sha, T.; Zhang, W.; Shen, T.; Li, Z.; Mei, T. Deep Person Generation: A Survey from the Perspective of Face, Pose and Cloth Synthesis. *arXiv* **2021**, arXiv:2109.02081.
13. Liu, J.; Zhu, Z.; Ren, Y.; Huang, W.; Huai, B.; Yuan, N.; Zhao, Z. Parallel and High-Fidelity Text-to-Lip Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2022; Volume 36, pp. 1738–1746.
14. Yang, Y.; Shillingford, B.; Assael, Y.; Wang, M.; Liu, W.; Chen, Y.; Zhang, Y.; Sezener, E.; Cobo, L.C.; Denil, M.; et al. Large-scale multilingual audio visual dubbing. *arXiv* **2020**, arXiv:2011.03530.
15. Zhou, H.; Sun, Y.; Wu, W.; Loy, C.C.; Wang, X.; Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4176–4186.
16. Kumar, R.; Sotelo, J.; Kumar, K.; de Brébisson, A.; Bengio, Y. Obamanet: Photo-realistic lip-sync from text. *arXiv* **2017**, arXiv:1801.01442.
17. Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 484–492.
18. Yamamoto, E.; Nakamura, S.; Shikano, K. Lip movement synthesis from speech based on Hidden Markov Models. *Speech Commun.* **1998**, *26*, 105–115. [CrossRef]
19. Ling, J.; Tan, X.; Chen, L.; Li, R.; Zhang, Y.; Zhao, S.; Song, L. StableFace: Analyzing and Improving Motion Stability for Talking Face Generation. *arXiv* **2022**, arXiv:2208.13717.
20. Almajai, I.; Milner, B. Visually derived wiener filters for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.* **2010**, *19*, 1642–1651. [CrossRef]
21. Adeel, A.; Gogate, M.; Hussain, A.; Whitmer, W.M. Lip-reading driven deep learning approach for speech enhancement. *IEEE Trans. Emerg. Top. Comput. Intell.* **2019**, *5*, 481–490. [CrossRef]
22. Kumar, Y.; Aggarwal, M.; Nawal, P.; Satoh, S.; Shah, R.R.; Zimmermann, R. Harnessing ai for speech reconstruction using multi-view silent video feed. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1976–1983.
23. Kumar, Y.; Jain, R.; Salik, M.; ratn Shah, R.; Zimmermann, R.; Yin, Y. Mylipper: A personalized system for speech reconstruction using multi-view visual feeds. In Proceedings of the 2018 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 10–12 December 2018; pp. 159–166.
24. Kumar, Y.; Jain, R.; Salik, K.M.; Shah, R.R.; Yin, Y.; Zimmermann, R. Lipper: Synthesizing thy speech using multi-view lipreading. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 2588–2595.
25. Kumar, N.; Goel, S.; Narang, A.; Lall, B. Multi Modal Adaptive Normalization for Audio to Video Generation. *arXiv* **2020**, arXiv:2012.07304.
26. Salik, K.M.; Aggarwal, S.; Kumar, Y.; Shah, R.R.; Jain, R.; Zimmermann, R. Lipper: Speaker independent speech synthesis using multi-view lipreading. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 10023–10024.
27. Hassid, M.; Ramanovich, M.T.; Shillingford, B.; Wang, M.; Jia, Y.; Remez, T. More than Words: In-the-Wild Visually-Driven Prosody for Text-to-Speech. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10587–10597.

28. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef]
29. Ma, P.; Petridis, S.; Pantic, M. End-to-end audio-visual speech recognition with conformers. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7613–7617.
30. Ivanko, D.; Ryumin, D.; Karpov, A. Automatic Lip-Reading of Hearing Impaired People. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 97–101. [CrossRef]
31. Ebert, D.A.; Heckerling, P.S. Communication with deaf patients: Knowledge, beliefs, and practices of physicians. *JAMA* **1995**, *273*, 227–229. [CrossRef]
32. Barnett, S. Clinical and cultural issues in caring for deaf people. *Fam. Med.* **1999**, *31*, 17–22.
33. Davenport, S. Improving communication with the deaf patient. *J. Fam. Pract.* **1977**, *4*, 1065–1068.
34. Steinberg, A. Issues in providing mental health services to hearing-impaired persons. *Psychiatr. Serv.* **1991**, *42*, 380–389. [CrossRef]
35. Fernandez-Lopez, A.; Martinez, O.; Sukno, F.M. Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017; pp. 208–215.
36. Altieri, N.A.; Pisoni, D.B.; Townsend, J.T. Some normative data on lip-reading skills (L). *J. Acoust. Soc. Am.* **2011**, *130*, 1–4. [CrossRef]
37. Hilder, S.; Harvey, R.W.; Theobald, B.J. Comparison of human and machine-based lip-reading. In *AVSP*; University of East Anglia: Norwich, UK, 2009; pp. 86–89.
38. Sooraj, V.; Hardhik, M.; Murthy, N.S.; Sandesh, C.; Shashidhar, R. Lip-reading techniques: A review. *Int. J. Sci. Technol. Res.* **2020**, *9*, 4378–4383.
39. Oghbaie, M.; Sabaghi, A.; Hashemifard, K.; Akbari, M. Advances and Challenges in Deep Lip Reading. *arXiv* **2021**, arXiv:2110.07879.
40. Agrawal, S.; Omprakash, V.R.; Ranvijay. Lip reading techniques: A survey. In Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, 21–23 July 2016; pp. 753–757. [CrossRef]
41. Hao, M.; Mamut, M.; Yadikar, N.; Aysa, A.; Ubul, K. A survey of research on lipreading technology. *IEEE Access* **2020**, *8*, 204518–204544. [CrossRef]
42. Fernandez-Lopez, A.; Sukno, F.M. Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput.* **2018**, *78*, 53–72. [CrossRef]
43. Chung, J.S.; Zisserman, A. Lip reading in the wild. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 87–103.
44. Tsourounis, D.; Kastaniotis, D.; Fotopoulos, S. Lip reading by alternating between spatiotemporal and spatial convolutions. *J. Imaging* **2021**, *7*, 91. [CrossRef] [PubMed]
45. Son Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6447–6456.
46. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552.
47. Stafylakis, T.; Tzimiropoulos, G. Combining residual networks with LSTMs for lipreading. *arXiv* **2017**, arXiv:1703.04105.
48. Cheng, S.; Ma, P.; Tzimiropoulos, G.; Petridis, S.; Bulat, A.; Shen, J.; Pantic, M. Towards pose-invariant lip-reading. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4357–4361.
49. Wang, C. Multi-grained spatio-temporal modeling for lip-reading. *arXiv* **2019**, arXiv:1908.11618.
50. Courtney, L.; Sreenivas, R. Using deep convolutional LSTM networks for learning spatiotemporal features. In Proceedings of the Asian Conference on Pattern Recognition, Jeju Island, Republic of Korea, 9–12 November 2019; pp. 307–320.
51. Luo, M.; Yang, S.; Shan, S.; Chen, X. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 273–280.
52. Weng, X.; Kitani, K. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. *arXiv* **2019**, arXiv:1905.02540.
53. Xiao, J.; Yang, S.; Zhang, Y.; Shan, S.; Chen, X. Deformation flow based two-stream network for lip reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 364–370.
54. Zhao, X.; Yang, S.; Shan, S.; Chen, X. Mutual information maximization for effective lip reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 420–427.
55. 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. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 356–363.
56. Feng, D.; Yang, S.; Shan, S.; Chen, X. Learn an effective lip reading model without pains. *arXiv* **2020**, arXiv:2011.07557.

57. Martinez, B.; Ma, P.; Petridis, S.; Pantic, M. Lipreading using temporal convolutional networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6319–6323.
58. Ren, S.; Du, Y.; Lv, J.; Han, G.; He, S. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13325–13333.
59. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Stat* **2015**, *1050*, 9.
60. Peng, C.; Li, J.; Chai, J.; Zhao, Z.; Zhang, H.; Tian, W. Lip Reading Using Deformable 3D Convolution and Channel-Temporal Attention. In Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022; pp. 707–718.
61. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards practical lipreading with distilled and efficient models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2021; pp. 7608–7612.
62. Koumparoulis, A.; Potamianos, G. Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 8467–8471.
63. Ma, P.; Wang, Y.; Petridis, S.; Shen, J.; Pantic, M. Training strategies for improved lip-reading. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 8472–8476.
64. Fenghour, S.; Chen, D.; Guo, K.; Li, B.; Xiao, P. Deep learning-based automated lip-reading: A survey. *IEEE Access* **2021**, *9*, 121184–121205. [CrossRef]
65. Pu, G.; Wang, H. Review on research progress of machine lip reading. *Vis. Comput.* **2022**, *39*, 3041–3057. [CrossRef]
66. Kaur, P.; Krishan, K.; Sharma, S.K.; Kanchan, T. Facial-recognition algorithms: A literature review. *Med. Sci. Law* **2020**, *60*, 131–139. [CrossRef]
67. Zhou, M.; Wang, Q.; Li, Q.; Jiang, P.; Yang, J.; Shen, C.; Wang, C.; Ding, S. Securing face liveness detection using unforgeable lip motion patterns. *arXiv* **2021**, arXiv:2106.08013.
68. Raji, I.D.; Fried, G. About face: A survey of facial recognition evaluation. *arXiv* **2021**, arXiv:2102.00813.
69. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv* **2018**, arXiv:1806.05622.
70. Chowdhury, D.P.; Kumari, R.; Bakshi, S.; Sahoo, M.N.; Das, A. Lip as biometric and beyond: A survey. *Multimed. Tools Appl.* **2022**, *81*, 3831–3865. [CrossRef]
71. Kumar, K.; Chen, T.; Stern, R.M. Profile view lip reading. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 16–20 April 2007; Volume 4, pp. IV–429–IV–432
72. Chitu, A.G.; Driel, K.; Rothkrantz, L.J. Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In Proceedings of the International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 6–10 September 2010; pp. 259–266.
73. Chiţu, A.; Rothkrantz, L.J. Automatic visual speech recognition. In *Speech Enhancement, Modeling and Recognition—Algorithms and Applications*; InTech Open: London, UK, 2012; p. 95.
74. Antar, S.; Sagheer, A.; Aly, S.; Tolba, M.F. Avas: Speech database for multimodal recognition applications. In Proceedings of the 13th International Conference on Hybrid Intelligent Systems (HIS 2013), Gammarth, Tunisia, 4–6 December 2013; pp. 123–128.
75. Fix, E.; Hodges, J.L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev. Int. Stat.* **1989**, *57*, 238–247. [CrossRef]
76. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. A new visual speech recognition approach for RGB-D cameras. In Proceedings of the International Conference Image Analysis and Recognition, Vilamoura, Portugal, 22–24 October 2014; pp. 21–28.
77. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. An adaptive approach for lip-reading using image and depth data. *Multimed. Tools Appl.* **2016**, *75*, 8609–8636. [CrossRef]
78. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
79. Parekh, D.; Gupta, A.; Chhatpar, S.; Yash, A.; Kulkarni, M. Lip reading using convolutional auto encoders as feature extractor. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019; pp. 1–6.
80. Estival, D.; Cassidy, S.; Cox, F.; Burnham, D. AusTalk: An audio-visual corpus of Australian English. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 3105–3109.
81. Sui, C.; Togneri, R.; Bennamoun, M. A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition. *Speech Commun.* **2017**, *90*, 26–38. [CrossRef]
82. Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; Chen, X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, France, 14–18 May 2019; pp. 1–8.
83. Wang, H.; Pu, G.; Chen, T. A Lip Reading Method Based on 3D Convolutional Vision Transformer. *IEEE Access* **2022**, *10*, 77205–77212. [CrossRef]
84. Chung, J.S.; Zisserman, A. Lip reading in profile. In Proceedings of the Ritish Machine Vision Conference, London, UK, 4–7 September 2017.

85. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
86. iFlyRec Team. iFlyRec: A Speech Recognition Tool. Available online: https://www.iflyrec.com/ (accessed on 11 July 2019).
87. SeetaFaceEngine2 Team. SeetaFaceEngine2. Available online: https://github.com/seetaface (accessed on 11 July 2019).
88. He, Z.; Kan, M.; Zhang, J.; Chen, X.; Shan, S. A Fully End-to-End Cascaded CNN for Facial Landmark Detection. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), Washington, DC, USA, 30 May–3 June 2017.
89. Xiao-Ding, C.; Chang-Chong, S.; Gang-Yao, K.; Li, L. The state of the art and prospects of lip reading. *Acta Autom. Sin.* **2020**, *46*, 2275–2301.
90. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef]
91. Margam, D.K.; Aralikatti, R.; Sharma, T.; Thanda, A.; Roy, S.; Venkatesan, S.M. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. *arXiv* **2019**, arXiv:1906.12170.
92. Anina, I.; Zhou, Z.; Zhao, G.; Pietikäinen, M. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 1, pp. 1–5.
93. Jeon, S.; Kim, M.S. End-to-End Sentence-Level Multi-View Lipreading Architecture with Spatial Attention Module Integrated Multiple CNNs and Cascaded Local Self-Attention-CTC. *Sensors* **2022**, *22*, 3597. [CrossRef]
94. Czyzewski, A.; Kostek, B.; Bratoszewski, P.; Kotus, J.; Szykulski, M. An audio-visual corpus for multimodal automatic speech recognition. *J. Intell. Inf. Syst.* **2017**, *49*, 167–192. [CrossRef]
95. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *44*, 8717–8727. [CrossRef] [PubMed]
96. Fenghour, S.; Chen, D.; Guo, K.; Xiao, P. Lip reading sentences using deep learning with only visual cues. *IEEE Access* **2020**, *8*, 215516–215530. [CrossRef]
97. Afouras, T.; Chung, J.S.; Zisserman, A. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv* **2018**, arXiv:1809.00496.
98. Ma, P.; Haliassos, A.; Fernandez-Lopez, A.; Chen, H.; Petridis, S.; Pantic, M. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
99. Alghamdi, N.; Maddock, S.; Marxer, R.; Barker, J.; Brown, G.J. A corpus of audio-visual Lombard speech with frontal and profile views. *J. Acoust. Soc. Am.* **2018**, *143*, EL523–EL529. [CrossRef]
100. Shillingford, B.; Assael, Y.; Hoffman, M.W.; Paine, T.; Hughes, C.; Prabhu, U.; Liao, H.; Sak, H.; Rao, K.; Bennett, L.; et al. Large-scale visual speech recognition. *arXiv* **2018**, arXiv:1807.05162.
101. Zhao, Y.; Xu, R.; Song, M. A cascade sequence-to-sequence model for chinese mandarin lip reading. In Proceedings of the ACM Multimedia Asia, Beijing, China, 16–18 December 2019; pp. 1–6.
102. Makino, T.; Liao, H.; Assael, Y.; Shillingford, B.; Garcia, B.; Braga, O.; Siohan, O. Recurrent neural network transducer for audio-visual speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 905–912.
103. Liu, X.; Lakomkin, E.; Vougioukas, K.; Ma, P.; Chen, H.; Xie, R.; Doulaty, M.; Moritz, N.; Kolar, J.; Petridis, S.; et al. SynthVSR: Scaling Up Visual Speech Recognition with Synthetic Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–19 June 2023; pp. 18806–18815.
104. Zhao, G.; Barnard, M.; Pietikainen, M. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimed.* **2009**, *11*, 1254–1265. [CrossRef]
105. Maeda, T.; Tamura, S. Multi-view Convolution for Lipreading. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 1092–1096.
106. Zimmermann, M.; Mehdipour Ghazi, M.; Ekenel, H.K.; Thiran, J.P. Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 264–276.
107. Petridis, S.; Wang, Y.; Li, Z.; Pantic, M. End-to-end multi-view lipreading. *arXiv* **2017**, arXiv:1709.00443.
108. Saitoh, T.; Zhou, Z.; Zhao, G.; Pietikäinen, M. Concatenated frame image based cnn for visual speech recognition. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 277–289.
109. Lee, D.; Lee, J.; Kim, K.E. Multi-view automatic lip-reading using neural network. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 290–302.
110. Han, H.; Kang, S.; Yoo, C.D. Multi-view visual speech recognition based on multi task learning. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3983–3987.
111. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
112. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
113. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.

114. Serdyuk, D.; Braga, O.; Siohan, O. Audio-Visual Speech Recognition is Worth 32 × 32 × 8 Voxels. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 15–17 December 2021; pp. 796–802.
115. Chang, O.; Liao, H.; Serdyuk, D.; Shah, A.; Siohan, O. Conformers are All You Need for Visual Speech Recogntion. *arXiv* **2023**, arXiv:2302.10915.
116. Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W.T.; Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv* **2018**, arXiv:1804.03619.
117. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
118. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QL, Australia, 19–24 April 2015; pp. 5206–5210.
119. Hernandez, F.; Nguyen, V.; Ghannay, S.; Tomashenko, N.; Esteve, Y. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Proceedings of the Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, 18–22 September 2018; pp. 198–208.
120. Prajwal, K.; Afouras, T.; Zisserman, A. Sub-word Level Lip Reading with Visual Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5162–5172.
121. Shi, B.; Hsu, W.N.; Lakhotia, K.; Mohamed, A. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv* **2022**, arXiv:2201.02184.
122. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication, Washington, DC, USA, 22–24 March 1999; Volume 964, pp. 965–966.
123. Sanderson, C.; Paliwal, K.K. Fast features for face authentication under illumination direction changes. *Pattern Recognit. Lett.* **2003**, *24*, 2409–2419. [CrossRef]
124. Lamel, L.F.; Kassel, R.H.; Seneff, S. Speech database development: Design and analysis of the acoustic-phonetic corpus. Speech Input/Output Assessment and Speech Databases. *Speech Commun.* **1989**, *9*, 161–170.
125. Wright, C.; Stewart, D.W. Understanding visual lip-based biometric authentication for mobile devices. *EURASIP J. Inf. Secur.* **2020**, *2020*, 1–16. [CrossRef]
126. Shang, D.; Zhang, X.; Xu, X. Face and lip-reading authentication system based on android smart phones. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 4178–4182.
127. Ruengprateepsang, K.; Wangsiripitak, S.; Pasupa, K. Hybrid Training of Speaker and Sentence Models for One-Shot Lip Password. In Proceedings of the International Conference on Neural Information Processing. Springer, Bangkok, Thailand, 23–27 November 2020; pp. 363–374.
128. Petridis, S.; Shen, J.; Cetin, D.; Pantic, M. Visual-only recognition of normal, whispered and silent speech. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6219–6223.
129. Faraj, M.I.; Bigun, J. Motion features from lip movement for person authentication. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 1059–1062.
130. Lu, Z.; Wu, X.; He, R. Person identification from lip texture analysis. In Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, 16–18 October 2016; pp. 472–476.
131. Sanchez, M.U.R. *Aspects of Facial Biometrics for Verification of Personal Identity*; University of Surrey: Surrey, UK, 2000.
132. Ichino, M.; Yamazaki, Y.; Jian-Gang, W.; Yun, Y.W. Text independent speaker gender recognition using lip movement. In Proceedings of the 2012 12th International Conference on Control Automation Robotics & Vision (ICARCV), Guangzhou, China, 5–7 December 2012; pp. 176–181.
133. Wright, C.; Stewart, D. One-shot-learning for visual lip-based biometric authentication. In Proceedings of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 7–9 October 2019; pp. 405–417.
134. Dar, S.A.; Palanivel, S.; Geetha, M.K.; Balasubramanian, M. Mouth Image Based Person Authentication Using DWLSTM and GRU. *Inf. Sci. Lett* **2022**, *11*, 853–862.
135. Kim, S.T.; Ro, Y.M. Attended relation feature representation of facial dynamics for facial authentication. *IEEE Trans. Inf. For. Secur.* **2018**, *14*, 1768–1778. [CrossRef]
136. Yuan, Y.; Zhao, J.; Xi, W.; Qian, C.; Zhang, X.; Wang, Z. SALM: Smartphone-based identity authentication using lip motion characteristics. In Proceedings of the 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 29–31 May 2017; pp. 1–8.
137. Wong, A.B. Authentication through Sensing of Tongue and Lip Motion via Smartphone. In Proceedings of the 2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Virtual Conference, 6–9 July 2021; pp. 1–2.
138. Sun, Z.; Lee, D.J.; Zhang, D.; Li, X. Concurrent Two-Factor Identify Verification Using Facial Identify and Facial Actions. *Electron. Imaging* **2021**, *2021*, 318-1–318-7. [CrossRef]
139. Hassanat, A.B. Visual passwords using automatic lip reading. *arXiv* **2014**, arXiv:1409.0924.

140. Sayo, A.; Kajikawa, Y.; Muneyasu, M. Biometrics authentication method using lip motion in utterance. In Proceedings of the 2011 8th International Conference on Information, Communications & Signal Processing, Singapore, 13–16 December 2011; pp. 1–5.
141. Mok, L.; Lau, W.; Leung, S.; Wang, S.; Yan, H. Lip features selection with application to person authentication. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings. Institute of Electrical and Electronics Engineers Inc., Montreal, QC, Canada, 17–21 May 2004; Volume 3.
142. Lu, L.; Yu, J.; Chen, Y.; Liu, H.; Zhu, Y.; Liu, Y.; Li, M. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications, Honolulu, HI, USA, 16–19 April 2018; pp. 1466–1474.
143. Tan, J.; Wang, X.; Nguyen, C.T.; Shi, Y. SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proc. Acm Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–18. [CrossRef]
144. Chen, J.; Cai, L.; Tu, Y.; Dong, R.; An, D.; Zhang, B. An Identity Authentication Method Based on Multi-modal Feature Fusion. *J. Phys. Conf. Ser.* **2021**, *1883*, 012060. [CrossRef]
145. Rabiner, L.; Juang, B. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16. [CrossRef]
146. Chen, C. PyTorch Face Landmark: A Fast and Accurate Facial Landmark Detector. 2021. Open-Source Software. Available online: https://github.com/cunjian/pytorch_face_landmark (accessed on 10 August 2023).
147. Lucey, S. An evaluation of visual speech features for the tasks of speech and speaker recognition. In Proceedings of the International Conference on Audio-and Video-Based Biometric Person Authentication, Guildford, UK, 9–11 June 2003; pp. 260–267.
148. Chetty, G.; Wagner, M. Automated lip feature extraction for liveness verification in audio-video authentication. In Proceedings of the Image and Vision Computing, Akaroa, New Zealand, 21–23 November 2004; pp. 17–22.
149. Shafait, F.; Kricke, R.; Shdaifat, I.; Grigat, R.R. Real time lip motion analysis for a person authentication system using near infrared illumination. In Proceedings of the 2006 International Conference on Image Processing, Las Vegas, NV, USA, 26–29 June 2006; pp. 1957–1960.
150. Jesorsky, O.; Kirchberg, K.J.; Frischholz, R.W. Robust face detection using the hausdorff distance. In Proceedings of the International Conference on Audio-and Video-based Biometric Person Authentication, Halmstad, Sweden, 6–8 June 2001; pp. 90–95.
151. Faraj, M.I.; Bigun, J. Audio–visual person authentication using lip-motion from orientation maps. *Pattern Recognit. Lett.* **2007**, *28*, 1368–1382. [CrossRef]
152. Nakata, T.; Kashima, M.; Sato, K.; Watanabe, M. Lip-sync personal authentication system using movement feature of lip. In Proceedings of the 2013 International Conference on Biometrics and Kansei Engineering, Tokyo, Japan, 5–7 July 2013; pp. 273–276.
153. Basheer Hassanat, A. Visual Words for Automatic Lip-Reading. *arXiv* **2014**, arXiv:1409.6689.
154. Hassanat, A.B.; Jassim, S. Visual words for lip-reading. In Proceedings of the Mobile Multimedia/Image Processing, Security, and Applications, Orlando, FL, USA, 5–9 April 2010; Volume 7708, pp. 86–97.
155. Wright, C.; Stewart, D.; Miller, P.; Campbell-West, F. Investigation into DCT feature selection for visual lip-based biometric authentication. In Proceedings of the Irish Machine Vision & Image Processing Conference Proceedings, Dublin, Irland, 28–30 August 2015; pp. 11–18.
156. Lander, K.; Chuang, L. Why are moving faces easier to recognize? *Vis. Cogn.* **2005**, *12*, 429–442. [CrossRef]
157. Sun, Z.; Sumsion, A.; Torrie, S.; Lee, D.J. Learn Dynamic Facial Motion Representations Using Transformer Encoder. In Proceedings of the Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 14–15 May 2022; pp. 1–5.
158. Sun, Z.; Sumsion, A.W.; Torrie, S.A.; Lee, D.J. Learning Facial Motion Representation with a Lightweight Encoder for Identity Verification. *Electronics* **2022**, *11*, 1946. [CrossRef]
159. Torrie, S.; Sumsion, A.; Sun, Z.; Lee, D.J. Facial Password Data Augmentation. In Proceedings of the Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 14–15 May 2022; pp. 1–5.
160. Perc, M.; Ozer, M.; Hojnik, J. Social and juristic challenges of artificial intelligence. *Palgrave Commun.* **2019**, *5*, 61. [CrossRef]
161. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. Lipnet: End-to-end sentence-level lipreading. *arXiv* **2016** arXiv:1611.01599.