

Audiovisual speech recognition: A review and forecast

Linlin Xia¹ , Gang Chen¹, Xun Xu², Jiashuo Cui¹ and Yiping Gao¹

Abstract

Audiovisual speech recognition is a favorable solution to multimodality human–computer interaction. For a long time, it has been very difficult to develop machines capable of generating or understanding even fragments of natural languages; the fused sight, smelling, touching, and so on provide machines with possible mediums to perceive and understand. This article presents a detailed review of recent advances in audiovisual speech recognition area. After explicitly representing audiovisual speech recognition development phase divided by timeline, we focus on typical audiovisual speech database descriptions in terms of single view and multi-view, since the public databases for general purpose should be the first concern for audiovisual speech recognition tasks. For the following challenges that are inseparably related to the feature extraction and dynamic audiovisual fusion, the principal usefulness of deep learning-based tools, such as deep fully convolutional neural network, bidirectional long short-term memory network, 3D convolutional neural network, and so on, lies in the fact that they are relatively simple solutions of such problems. As the principle analyses and comparisons related to computational load, accuracy, and applicability of well-developed audiovisual speech recognition frameworks have been conducted, we further illuminate our insights into the future audiovisual speech recognition architecture design. We argue that end-to-end audiovisual speech recognition model and deep learning-based feature extractors will guide multimodality human–computer interaction directly to a solution.

Keywords

AVSR, multimodality HCI, public databases, deep learning, neural networks

Date received: 30 July 2019; accepted: 26 September 2020

Topic Area: AI in Robotics; Human Robot/Machine Interaction

Associate Editor: Oscar Mozos

Topic Editor: Fernandez-Caballero

Introduction

When it comes to the issues of human–computer interaction (HCI), human–computer harmony is as eternal a pursuit of HCI as human–computer collaboration. As the booming of artificial intelligence dramatically promotes the intellectualization of machines, the HCI technology, therefore, is currently facing increasing challenges and complications than ever before. Under this background, a more powerful and humanized means in terms of audio perception seems to be realistic when dealing with high-accuracy rate HCI problems, regardless of a tranquil work envelope or a noisy work volume within which the machines operate. Audio speech recognition (ASR)

essentially functions as a favorable bridge joining these two parts: humans and machines.

At present, machines share the same ASR cognition level with humans under quiet environments, but in cases

¹ School of Automation Engineering, Northeast Electric Power University, Jilin, China

² Institute for Superconducting and Electronic Materials, University of Wollongong, Wollongong, Australia

Corresponding author:

Linlin Xia, School of Automation Engineering, Northeast Electric Power University, Jilin City, Jilin 132012, China.

Email: xiall521@neepu.edu.cn



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without

further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

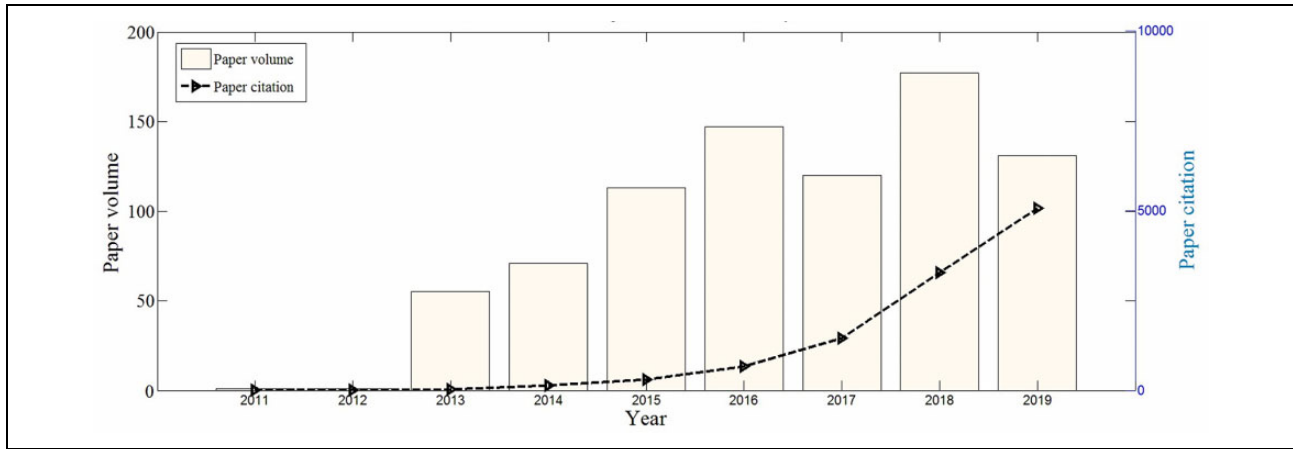


Figure 1. The retrieval statistics of subject word “AVSR” by Web of Science. AVSR: audiovisual speech recognition.

where the work spaces are noisy, generally, the recognition result by machines lags far behind in accuracy when compared to humans.¹ The reason that humans are more competent to identify noisy speech consists of the fact that speech perception by them appears to be a multimodal process. As the superior intelligent creatures that exist in nature, humans, respectively, acquire acoustic information and visual information by cochlear and eyes; with the derived multimodal information has been comprehensively analyzed and processed by brains, they consequently present the desired perceptual speech results.² By simulating human brain mechanism on multimodality-based speech recognition, the audiovisual speech recognition (AVSR) has been put forward. The perceptual vision results favorably remain reliable irrespective of the noise fed to the vision channel; the extracted visual modal information in AVSR frame is therefore applied to compensate or correct the missing or insufficient audio recognition results. Specifically, the adoption of video information source with dynamically adjusting the weight between audio output (or we say stream) and visual output (stream) between audio and visual channel results provides solutions for robust AVSR under severe noise conditions. It has been proved that robust AVSR significantly improves the recognition accuracy of ASR systems under a variety of unfavorable acoustic conditions.³

Lately, there has been more research in the area of AVSR-based HCI. According to the retrieval statistics of Web of Science, the subject word “AVSR” is highly cited. The overall increase in paper publication amount and paper citation amount is diagrammatically represented in Figure 1 (by mid-2019). As shown, AVSR is now attracting more and more research attention.

AVSR system can be applied to a vastly wider range of applying scenarios, such as command recognition in land vehicles,⁴ text translating of mobile phones,^{5,6} lipreading for people who are hearing impaired,^{7,8} speech recognition of individual speaker from lots of people who are talking at once,⁹ and so on. In practice, the difficulties in building an

efficient AVSR system mainly concentrate on the following three aspects:

- How to build public databases for general purpose.
- How to extract the representative visual features for a specific use.
- How to dynamically fuse audio features and visual features for a better speech recognition.

As illustrated, the knowledge associated with image or vision is closely related to AVSR research. In fact, the visualized information was the original inspiration for these fundamental ideas of AVSR. Research shows that the extraction of visual features largely relies on robust face recognition,^{10–12} mouth region tracking,^{13,14} and visual feature representation.^{15,16} The eminent computing capacity of deep learning-based algorithms enables AVSR to precisely locate the region of interest (ROI), to efficiently extract the visual features, and to accurately determinate the weights (which balance audio stream and visual stream), which therefore facilitates the algorithmic framework enhancements of AVSR.

The main contributions to this article are shown in the following aspects:

First, we analyzed the historical phase of AVSR. With mathematically expressing the AVSR process, we summarized the traditional and deep learning-based tools that are employed in AVSR systems. Meanwhile, we conducted a comparative analysis of existing AVSR databases in terms of task, number of testers, corpus, recording angle (or view), and so on, and we explicitly expound our views of open-source audiovisual databases qualified by actual scenario applications.

Second, we emphasized two main concerns of structuring an AVSR anatomy. In principle, the feature extraction (in terms of audio/visual modality) and dynamic audiovisual fusion are fully analyzed. The convolutional neural network (CNN)-based extractors that enable the efficient feature extractions and the advanced neural networks or

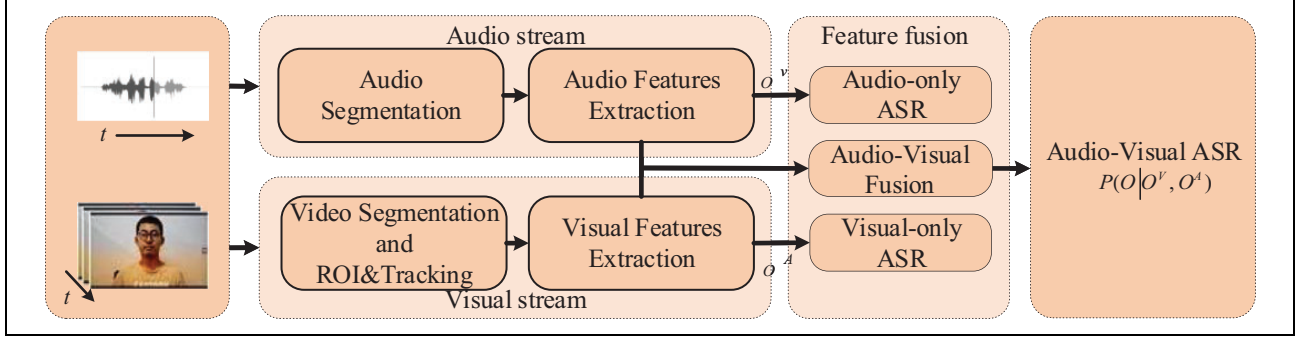


Figure 2. Block diagram of a typical AVSR. AVSR: audiovisual speech recognition.

encoders-based weight training models are summarized. From robustness point of view, we explicitly illuminate that we remain bullish on the prospects of end-to-end AVSR models under environments with multisource noises.

Third, in terms of case studies, a possible AVSR architecture design that fuses multisource information derived from audio, vision, and depth channels is proposed. The formalization of the speech recognition process using deep learning ideas has been directed at making HCI practicable in many situations. We believe, with fully developed frameworks, the deep learning-based AVSR that fuses reliable depth for image description will provide an important model reference for a class of understanding-driven multi-modality HCI.

The outline of the remainder of the article is as follows. In the following section, the principles of AVSR and related theories are firstly concerned and stated. The third section stresses a detailed description of recently developed audiovisual speech databases, which covers the applicability analyses under different types of tasks for solving “public databases for general purpose” problem. The independent audio/visual feature extraction is theoretically illustrated in the fourth section, together with the graphic interpretations of dynamic audiovisual fusion in terms of feature level and decision level in the fifth section, which underpins our vision for future AVSR architecture design. The sixth section also emphasizes the discussions of deep leaning-based frameworks for efficient feature extraction and dynamic feature fusion. The seventh section presents the main conclusion of this review study.

Principles of AVSR and related theories

Mathematical model of AVSR system

AVSR (also being referred to as dual-modality-based speech recognition) is to combine visual information (reflecting mouth movements) with original acoustic information to improve the accuracy rate of HCI. As in Figure 2, a typical AVSR system is constructed of a visual stream channel (VSC), an audio stream channel (ASC), and a dynamic audiovisual fusion part. Together with the acoustic

features derived from the paralleled ASC, the VSC results via ROI location of video sources, vision-to-image transformation, digital signal processing, and visual feature extraction constitute the inputs of dynamic audiovisual fusion part. Let $O^V = (O_1^V, O_2^V, \dots, O_n^V)$ and $O^A = (O_1^A, O_2^A, \dots, O_n^A)$, respectively, denote the VSC feature vector and ASC feature vector, and let $O = (O_1, O_2, \dots, O_l)$ denote the AVSR output vector, then the AVSR process can be mathematically represented by the following expression

$$P(O|O^V, O^A) = \prod_i P(O_i|O^V, O^A) \sum_{i=1}^n X_i \quad (1)$$

Relevant models and theories of AVSR during different developing stages

Stimulated by practical demands of HCI, the models and theories of AVSR evolve during the past decades of years. As illustrated in Figure 3, the development of AVSR, from the macroscopical point of view, experiences three distinguishable stages, in which theories research, traditional tools-based solutions, and deep learning-based solutions are concerned.

- Theories research:

This stage is characterized by the facts that related AVSR research has not concluded with well-established theories. This stage lasts for about 40 years long (from 1950 to 1990). As early as 1954, Pollack and Sumbly were the advocate of AVSR, who explicitly stated that visual information contributes to the understandings of speech in noisy environments.¹⁷ Representative McGurk effect (described in article titled “Hearing Lips and Seeing Voices” by Harry McGurk and John MacDonald) further presented the positive roles of visual modality information on speech perception. The theories research remained a limited prosperity until 1984; the first AVSR system was simply structured, which was mainly concerned with notably different mouth features of width, length, perimeter, and area as it aims to extract lip images in terms of image thresholds. With dynamic time warping algorithm, it

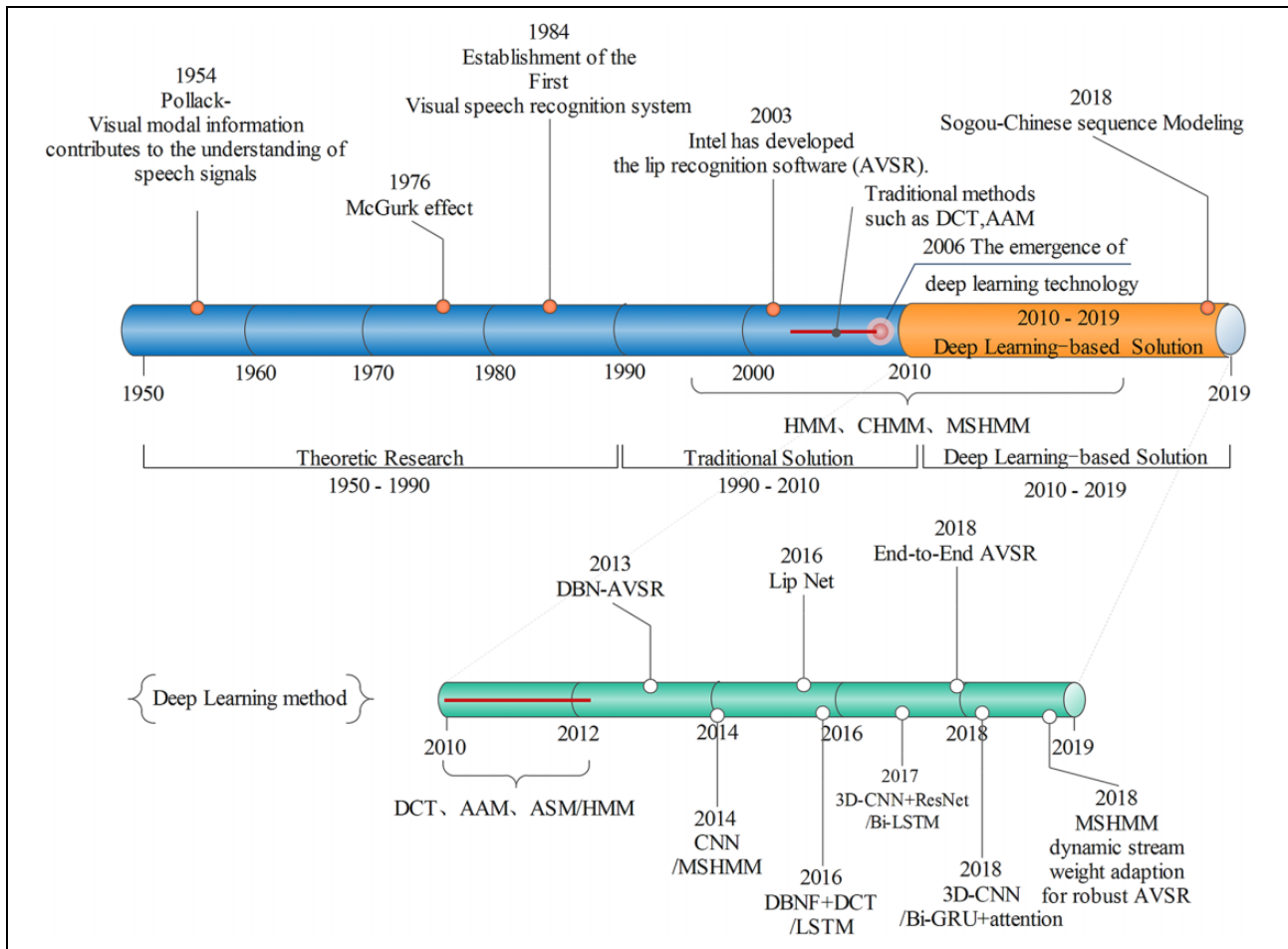


Figure 3. AVSR development timeline. AVSR: audiovisual speech recognition.

significantly accomplished the speech recognition of isolated words.¹⁸

- Traditional tools-based solutions:

The above model forms the basis of lip feature-based methods for efficient speech recognition under nonideal circumstances likely to be encountered. Generally, considering the manners of simultaneous speech-related feature maximization and extraneous feature restrain, the solutions with respect to visual feature extractions can be classified into two categories, one is to match a statistical model of lip shape and appearance, for instance, active appearance modal (AAM) is used for visual feature acquisition via image sequences of speakers.^{19,20} Another one is to directly extract the lip features in terms of pixels without regard to the priori lip models. In fact, the former is termed model-based feature extraction.²¹ In AAM, high-grade features of lip shape and appearance constitute the source from which a training supervisor derives accurate coordinates of landmarks via quantitatively training phase. The latter is termed image-based feature extraction, which distinguishes itself by no need for specific lip

models or coordinates of landmarks for training; a solution by directly extracting inferior pixel-based features via discrete cosine transform (DCT) or principal components analysis (PCA) methods may be possible.^{22,23} However, this pattern is essentially restricted to lighting conditions, translational motions, or rotational motions of images. The AVSR system may evolve with the desirable feature extraction results. It is important to appreciate that hidden Markov model (HMM) and multi-stream hidden Markov model (MSHMM)²⁴ are preferred methods for time sequence modeling to underpin the following dynamic audiovisual fusion.

- Deep learning-based solutions:

The deep learning methods are devoted to an enhancement of multimodality information fusion in terms of vision, auditory, tactility, and so on.²⁵ An illustrative system is “Lip Net,” which was developed at Oxford University in 2017. With multiple sets of testing on GRID data set, the superiority of Lip Net over lip-reading experts on total accurate rate of voice recognition has been numerically validated by a 93.4–52.3% lead.²⁶ Last year, by a method

Table 1. Descriptions of single view audiovisual databases.^a

Database	Birth year	Language	Task	Corpus	Testers	Resolution	Detailed description
AV Letters ³³	1998	English	A	26 individual English letters/repeating 3 times for each letter	10 (5 males, 5 females)	V:376 × 288 25 fps A:16bit 22.5 kHz	Top-most used database for alphabetic AVSR
GRID ³⁴	2006	English	P	Combination of commands, colors, numbers, etc.	34 (18 males, 16 females)	V-H:720 × 576 25 fps V-L:360 × 288 25 fps A: down-sampled to 25 kHz	Recorded in quiet labs with blue background
AV Letters-2 ³⁵	2008	English	D	26 individual English letters/repeating 7 times for each letter	5	V:1920 × 1080 50 fps A: 16 bit 48 kHz	Smaller database with improved recording conditions (compared with AV Letters)
OuluVS1 ³⁶	2009	English	P	10 phrases	20 (17 males, 3 females)	V:720 × 576 25 fps	10 daily-used English phrases, like “Excuse me,” “see you”
LRS-TED ³⁷	2018	English	S	More than 400 h of lecture videos from TED and TEDx	1000+	V: 224 × 224 25 fps A: 16 bit 16 kHz	Representative audiovisual database with large scale for general purpose
LRW-1000 ³⁸	2019	Mandarin	W	More than 500 h of broadcast news and conversational programs	2000+	V-H:1920 × 1080 25 fps V-L:1024 × 576 25 fps A: 128–160 Kbps audio bitrates	Largest word-level lip-reading database, only public large-scale Mandarin lip-reading database

AVSR: audiovisual speech recognition.

^aFor “resolution” item, A stands for audio, V stands for video, V-H stands for visual higher quality version, V-L stands for visual lower quality version.

of modeling Chinese lip-reading sequences, a lip-reading program was launched by Sogou Inc. It reveals that it got 60% correct in a speaker-independent open oral test and up to 90% correct in certain (smart home and in-vehicle) scenarios.²⁷ Throughout the timeline, the beginning of deep learning era in AVSR is 2010, breakthroughs have occurred ever since. Algorithmically, CNN is now taking the places of traditional DCT and AAM in aspects of feature extractions. Also, HMM and MSHMM are currently being substituted by long short-time memory network (LSTM)²⁸ or bidirectional long-term memory networks (Bi-LSTM)²⁹ in aspects of time sequence modeling. Due to space limitations, please refer to the enlarged region of Figure 3 for more detailed illustrations of deep learning-based solutions.

The audiovisual speech databases

Special attention should be paid to audiovisual speech databases because suitable data sets for certain purpose are so essential. As a consequence of multiplicity and complexity of languages, together with diverse research aims and requirements, it may be not realistic to create one audiovisual speech database that can meet all kinds of research needs, and also, it is impossible to assign one universally accepted standard which is sufficiently good

to evaluate such databases or compare them in a wide range of potential applications. Differing from simple speech databases, the establishment of corpus linguistics with certain scale is dependent on many other factors. Interests are focused not only upon the subjects of recorded or the number of times that testers repeat the corpus, but upon the lighting conditions of testing environments, the head posture of testers, the amount of recorded vocabularies, the resolution of recorded video, and so on.

The entire section is devoted to summarize the typical audiovisual speech databases and further analyze their applicabilities in different types of tasks. Five forms, more precisely Alphabet (A), Digits (D), Words (W), Phrases (P), and Sentences (S), are concerned, to one of which all speech sources approximate. Considering the possible views of recording videos, we have single view databases and multiview databases, whose principal attributes (like task, language, corpus, etc.) are, respectively, described in Tables 1 and 2.

Single view database

The typical single view databases and relative descriptions are listed in Table 1. As indicated, from the alphabetic AV Letters to the phrases-type OuluVS1 or from the sentence-

Table 2. Descriptions of multi-view audiovisual databases.

Database	Birth year	Language	Task	Corpus	Testers	Recording angle or view	Detailed description
CUAVE ⁴¹	2004	English	A/D	Arabic numerals 0–9	36 (individual or pair work)	–90°, 0°, 90°	Recorded in quiet indoor volume with blue background
AVICAR ⁴²	2006	English	A/D/S	Single digits, alphabet, phone number, and sentences	86 drivers	Four near front views	In-vehicle scenario under situations of engine idling, 35 mph running and 55 mph running with window opening and closing
IBMSR ⁴³	2008	English	D	Total of 38 subjects uttering connected-digit strings,	38	–90°, 0°, 90°	Two microphones and three cameras used for AV data collection. Two synchronous audio streams at 22 kHz and three visual streams at 30 Hz and 368 × 240-pixel frames are available
HIT-AVDB-II ⁴⁴	2009	English	S	Consists of Chinese and English poems, tongue twister, digits, Greek alphabet, and music	30	0°, 30°, 60°, 90°	Facilitating the investigation of multi-view biometrics technology and visual speech reading
Qulips ⁴⁵	2018	English	D	Arabic numerals 0–9	2	0°, 10°, 20°, ..., 90° (interval of 10°)	Only two cameras used; 180 digits are available for each of the 10 angles and each speaker, giving a total of 3600 digits
LilIR ⁴⁶	2010	English	S	Consists of 20 speakers uttering 200 sentences from the Resource Management Corpus	12	0°, 30°, 45°, 60°, 90°	Has a vocabulary size of approximately 1000 words
OuluVS2 ⁴⁷	2015	English	S	10 phrases	53 (41 males, 12 females)	0°, 30°, 45°, 60°, 90°	Thousands of videos simultaneously recorded by six cameras from five different views spanned between the frontal and profile words
TCD-TIMIT ⁴⁸	2015	English	S	Total of 6913 phonetically rich sentences	62	0°, 30°	Three of the speakers are professionally trained lipspeakers, recorded to test the hypothesis that lipspeakers may have an advantage over regular speakers in automatic visual speech recognition systems
AV Digits ⁴⁹	2018	English	D/P	Arabic numerals 0–9 and 10 phrases	53 (41 males, 12 females) for first part; 39 (32 males, 7 females) for second part	0°, 45°, 90°	Two testing parts in labs: first, reading numerals 0–9 in random order with normal, whispered soundless voice; second, reading 10 phrases (same with those in OuluVS2)
RGB-D ⁵⁰	2019	English	W/P	20 daily-used words/phrases	53	3D Head pose angles: yaw, pitch, and roll	Using Kinect facial tracking record the RGB data, depth data mapping between RGB and depth data 3D head orientation

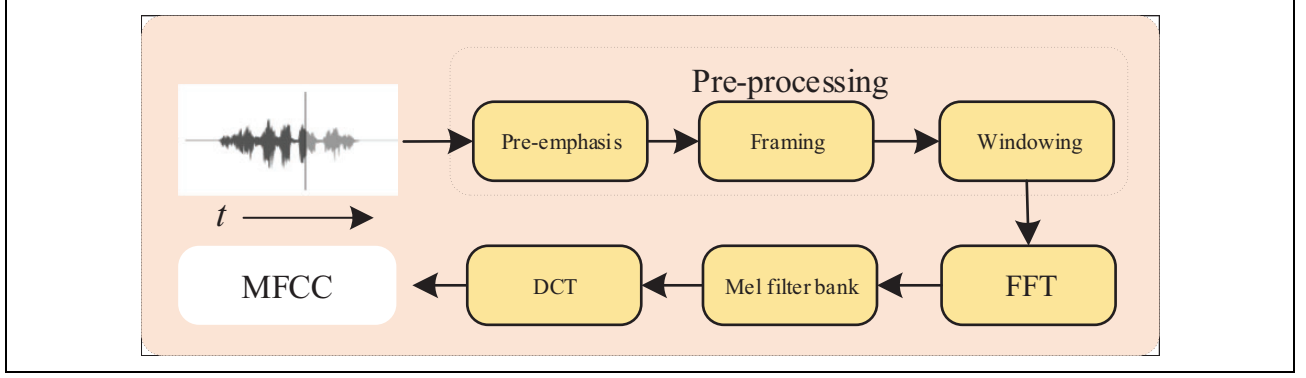


Figure 4. Flow diagram of MFCC feature extraction process. MFCC: Mel-frequency cepstral coefficients.

level LRS-TED for reference to the complex Mandarin-based LRW-1000, we may conclude the newly developed single view databases appear to be sentence-level, enriched language-type, large scale, and more testers involved. The above research indispensably benefit from enhanced data processing performances of machine learning (which has been largely applied to the industrial areas such as fault recovery,³⁰ parameter prediction,³¹ or classification³²). Meanwhile, as the optimized databases stand up to the requirements of high-grade HCI, it allows the AVSR to be conducted with more favorable and natural human-machine interaction means.

Multi-view database

Table 2 summarizes the multi-view audiovisual databases used in recent work in AVSR area. It lists the birth year, language, task type, corpus, recording angle or view, and so on. As indicated, the database transition from single view to multi-view describes the requirement of scenarios likely to be encountered in practical speech perception. During the last few years, multi-view database-based AVSR technology has advanced rapidly and now is laying the foundation of harmonious human-machine interactions. Even though audiovisual databases are being recognized as powerful and eminently practical benchmarks for the solution to comparative analyses problems, these tools, however, still face some challenges like the language of open-source database is basically English, the contents of corpus are short of pertinence with being not truly encompassed by the complex realities, the recorded audio or video is generally of inferior quality, and so on. It's worth mentioning that, as in Table 2, the newly developed database “RGB-D” employs an RGB-D (Kinect) camera to record visual resources, and the multisource inputs could be audio, 2D image, and 3D visual information. The significance of “depth” can also be found in the literature.^{39,40}

We further summarize the following qualifications of a much-needed audiovisual database based on the analysis hereinabove:

- Publicly accessible, rich in subject matter, widely covering visemes and phonemes
- Being recorded under high-quality visual and audio conditions
- Multimodality information (audio, vision, depth, etc.) fusion acceptable
- Multi-view video acquisition acceptable
- Large amount of testers with balanced sex ratio.

We are convinced of their progress in AVSR with the advance of deep learning technology, and we look forward to their improvements in the near future.

Feature extraction

This section mainly discusses the principles of audio feature extraction and visual feature extraction in algorithmic domain. Together with the following dynamic audiovisual fusion section, it theoretically leads to the final discussion with our vision for future development.

Audio feature extraction

Mel-frequency cepstral coefficients (MFCCs) are basically recognized as desirable audio features because of their superior robustness when dealing with channel noise and spectral distortion.⁵¹ The detailed feature extraction process for MFCC is given in Figure 4.

As shown, the process of extracting MFCC features consists in first performing pre-emphasis of speech signals that are separated from videos. A filter $[H(z) = 1 - \mu z^{-1}]$ ($0.9 < \mu < 1$ denoting preemphasis filter coefficients) is primarily implemented to minimize losses of high-frequency components, and as the following operations of framing and frequency-windowing are sequentially conducted, the speech signals are further optimized. To derive frequency components, each frame of signal in time domain is transformed with fast Fourier transformation (FFT). For the FFT result, calculate the amount of energy in a wide range of frequency regions via Mel filter bank and take the DCT of log energy derived from each bank for

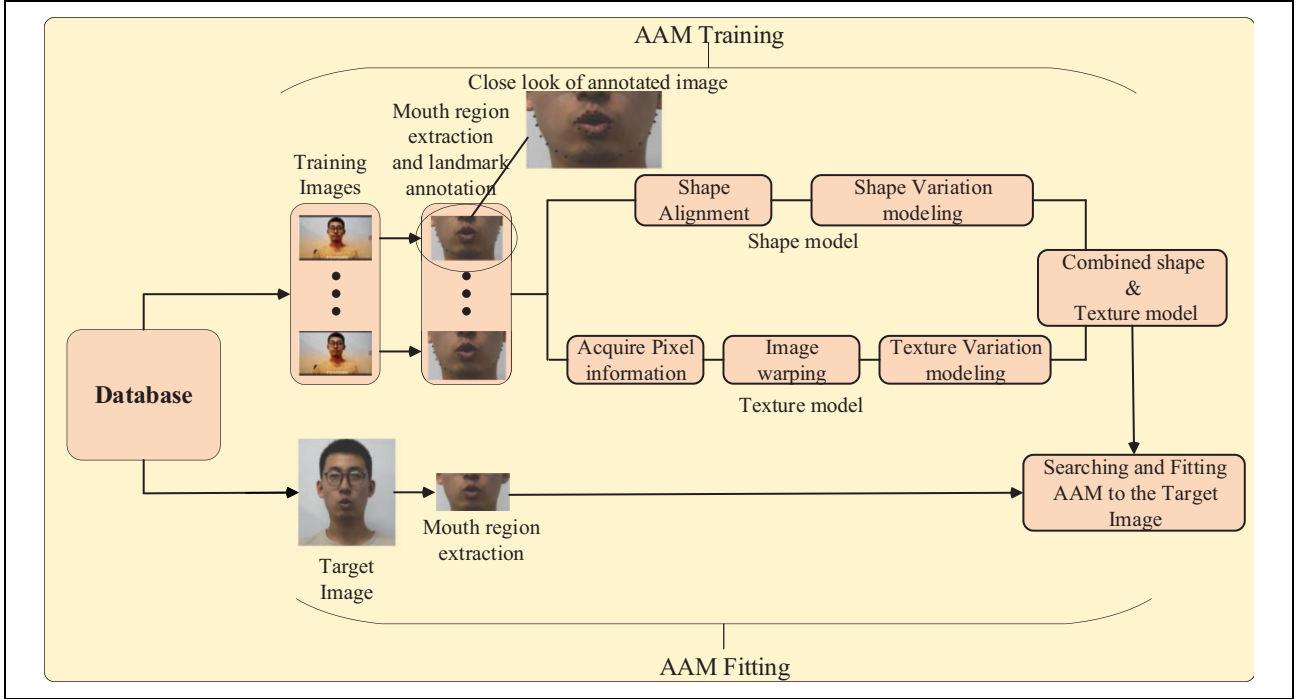


Figure 5. Framework of AAM training and fitting. AAM: active appearance modal.

purpose of de-correlation, we subsequently obtain expected MFCC features on Mel scale spectrum.

Visual feature extraction

As stated in “Relevant models and theories of AVSR during different developing stages” of the second section, early studies on visual feature extraction focus upon model-based AAM method or image-based DCT/PCA methods. Differing from the lower pixel-level DCT, AAM is now being recognized as a powerful statistical tool which is characterized by constructing its model in terms of shape and texture, more precisely, which is followed by the gradient-descent fitting algorithm to fit the trained model into the target image.⁵² The paralleled training and fitting parts form the basic AAM anatomy. Take RIO of lip as an example of such AAM frameworks, for the images to be trained, the mouth region extraction and landmark annotation tasks should be firstly taken, as illustrated by the amplified “close look of annotated image” part in Figure 5.

Three models contribute to the paralleled AAM training and AAM fitting, they are shape model, texture model, and appearance model. Based on the annotated images, the shape feature vector of all samples (images) is defined and represented as $s = (s_1, s_2, \dots, s_N)$ (N denotes number of images); once the shape coordinates are aligned to the unified coordinates, the dimensionality reduction will be conducted by means of classic PCA. The shape model of each image can be approximated by

$$s = \bar{s} + p_s b_s \quad (2)$$

where \bar{s} denotes mean shape, p_s denotes a set of orthogonal modes of variation, and b_s denotes a set of shape parameters. The texture model of AAM refers to those pixel points that are mapped in the matched regions of newly established shape model. The combined shapes and textures of faces, however, are the source from which the texture model derives its pixel points. Analogously, we have

$$a = a_0 + P_a b_a \quad (3)$$

where a_0 denotes mean texture, p_a denotes a set of orthogonal modes of variation, and b_a denotes a set of texture parameters.

Combine b_s and b_a into one vector and define

$$c = [b_s^T, b_a^T]^T \quad (4)$$

In consequence, formula (4) mathematically represents the appearance model of AAM. Note that, vector c is directly used to search and fit AAM to the target image. By a method similar to that adopted for de-correlation and dimensionality reduction (more precisely PCA method), the AAM may be mathematically idealized by

$$s = s_0 + Q_s c \quad (5)$$

$$a = a_0 + Q_a c \quad (6)$$

where Q_a and Q_s , respectively, represent shape feature matrix and texture feature matrix.

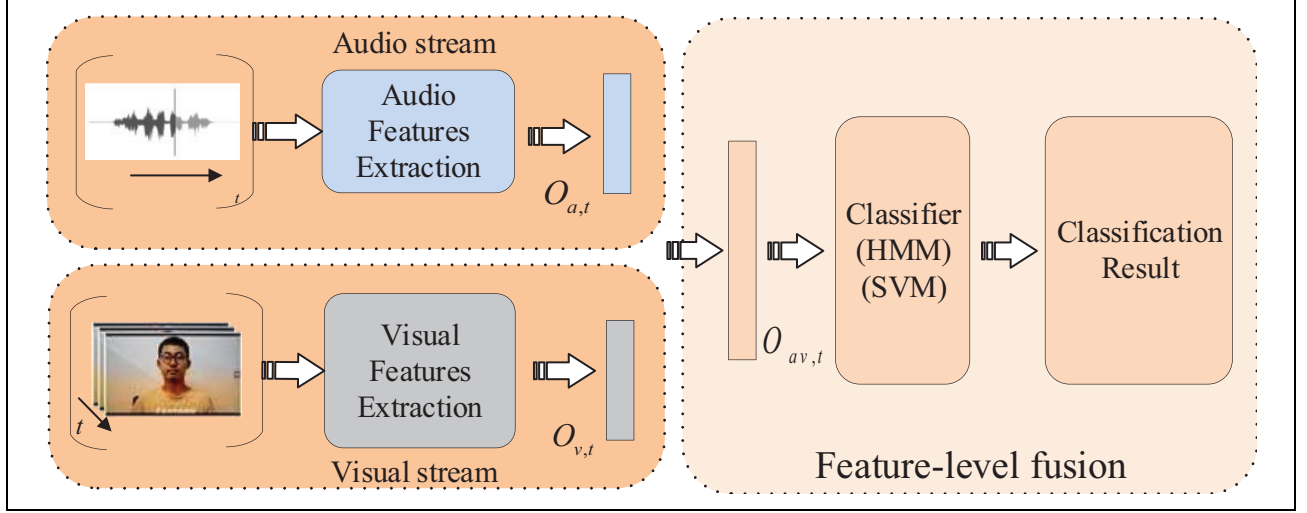


Figure 6. Flow diagram of feature-level fusion.

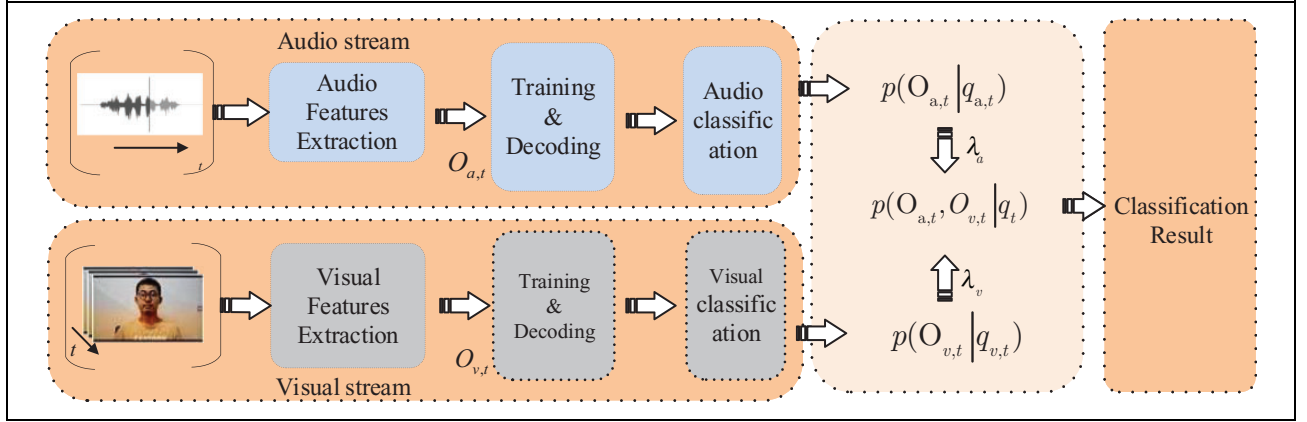


Figure 7. Flow diagram of decision-level fusion.

Dynamic audiovisual fusion

The principal importance of audiovisual fusion lies in its adaption to quality change of audio/visual signals. Let's consider a destroyed audio file or a lost face, the unavailability of which does not allow AVSR system to directly combine the corresponding features. Dynamic audiovisual fusion is the answer to this challenge issue. Fusion can be performed at either feature level or decision level.

Feature-level fusion

Assume vector $O_{a,t}$ (dimension $l_a = N$) and vector $O_{v,t}$ (dimension $l_v = M$) are, respectively, extracted from audio stream and visual stream at time (please see Figure 6). The combined vector

$$O_{av,t} = [O_{a,t}, O_{v,t}] \in R^{l_{av}} \quad (7)$$

represents the newly formed audiovisual feature with dimension $l_{av} = N + M$. For the vector $O_{av,t}$ represented by formula (7), whose high dimension l_{av} may suggest that

linear discriminant analysis or PCA⁵³ should be adopted for dimension reduction. The modified form of $O_{av,t}$ (with lower dimension), accordingly, will be further modeled by HMM or other classifiers.

As indicated, with simple cases, it is relatively easy to construct the fused audiovisual feature vector that yields to independent streams. But for complicated cases where the weight between these two forms of features should dynamically changes with the variation of the reliable channels, the decision-level fusion must be used.

Decision-level fusion

Decision-level fusion is more adaptable to changes than feature-level fusion. A basic decision-level fusion can be constructed from two paralleled channels as connected in Figure 7. Similarly, assume vector $O_{a,t}$ and vector $O_{v,t}$ are, respectively, extracted from audio stream and visual stream at time t . The emission probability represents a joint form of two independent probabilities $p(O_{a,t} | q_{a,t})$ and

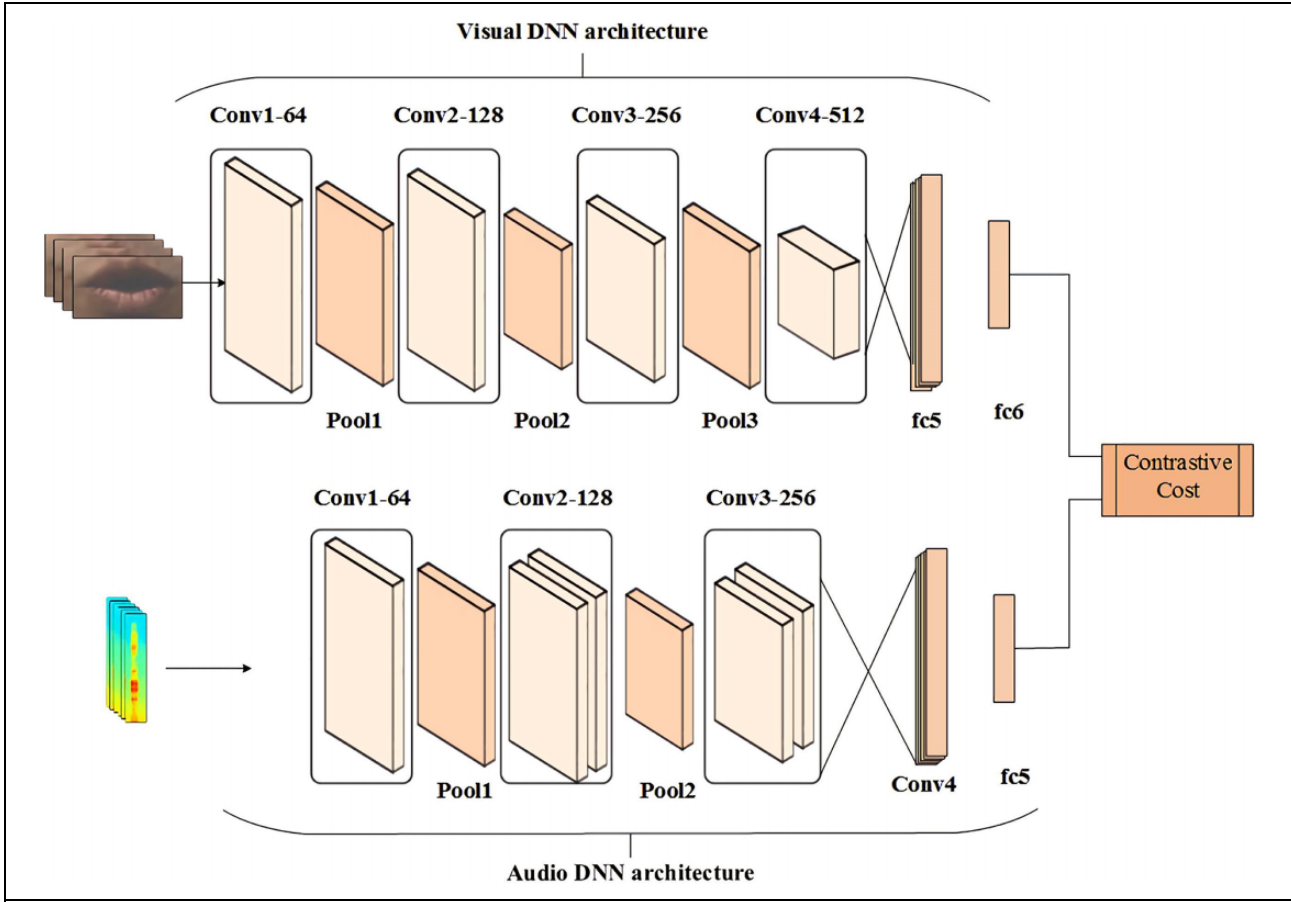


Figure 8. Architecture of 3D CNN. CNN: convolutional neural network.

$p(O_{v,t}|q_{v,t})$, it follows that

$$p(O_{v,t}, O_{a,t}|q_t) = p(O_{v,t}|q_{v,t})^{\lambda_v} p(O_{a,t}|q_{a,t})^{\lambda_a} \quad (8)$$

where $q = [q_{v,t}, q_{a,t}]^T$ represents state numbers. λ_v and λ_a , respectively, represent the audio weight and visual weight that directly determine the separate rate of contribution to the joint probability.

The key to dynamically adjust weights consists in determining the most relevant features used for reliability scaling of matched modalities. In this sense, the results of adjustments are typically measured in quality of audio signals. In research of AS Saudi et al.,⁵⁴ six-dimensional vector (three in time domain and three in frequency domain) extracted from each audio frame is used to describe this so-called quality. The stream weight adjusting tests are firstly conducted under different signal-to-noise ratios (SNRs), and the optimized weights are subsequently assigned to the matched SNR. In fact, the principal attribute of dynamic weight adjusting is that the optimal weights under different SNR are eventually determined by multilayer perceptron (MLP) model. By comparison, decision-level fusion deals with two channel streams by dynamic weight adjusting strategy rather than directly combine the corresponding

features.⁵⁵ This solution makes it more applicable to high robustness-required circumstances.

Discussions

Based on the foregoing research work on AVSR, we would like to discuss the core techniques from certain aspects: feature extraction and dynamic audiovisual fusion. Meanwhile, we would like to further state our vision for future AVSR architecture design.

Feature extraction

Audio feature extraction. A gloomy fact in AVSR field has always been that researchers would like to pay their attention to well-established visual feature extraction techniques rather than audio feature extraction means. Lately, an increasing number of papers in the fields of AVSR are devoted to MFCC-based acoustic feature acquisition. The first-order transformation (denoted by Δ) for frame correlation enhancements and second-order transformation (denoted by $\Delta\Delta$) for dynamic feature extraction are both conducted to the formation of desired audio-stream features, videlicet.

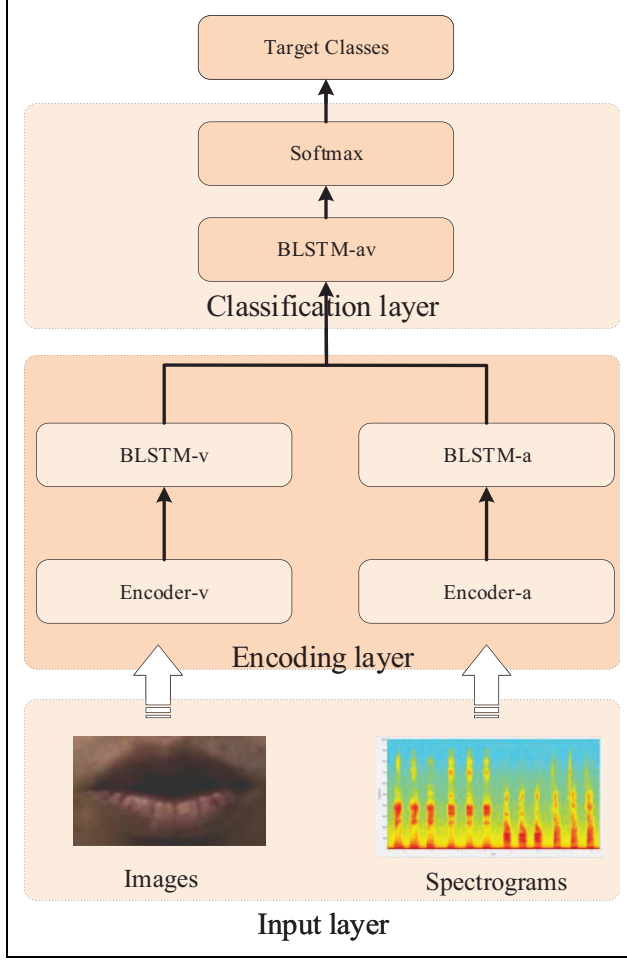


Figure 9. The graphic interpretation of end-to-end audiovisual model.

$$\text{Audio features} = \text{MFCC} + \Delta + \Delta\Delta \quad (9)$$

It is gratifying to see lots of novel frameworks on audio recognition have been put forward. An example of such a framework is deep fully convolutional neural network (DFCNN) fused model (given by iFLYTEK).⁵⁶ The significant advantage of which over the classical frameworks is that it directly implements “speech to spectrogram transformation” without concerning the matched MFCC features. Here, the combined convolutional layer and pooling layer of DFCNN enable the models of speeches to be formed, so that the output explicitly corresponds to the final recognition results, like syllables or Chinese characters, and so on. A notable comparison between DFCNN-based framework and BLSTM-CTC (known as the best in audio recognition area) shows, via tens of thousands of hours testing with Chinese data sets, the former presents an additional 15% recognition rate enhancement.

Concerning the reasons that spectrograms are generally superior to MFCC features, the unexpected information losses both in frequency domain and in time domain would certainly be highly mentioned. The former corresponds to

the losses of high-frequency components as described in part A of Section 4, and the latter refers to over condensed computation caused by large frame shifts, especially in cases where the speakers speak fast.

As a summary, heavy emphasis should be placed upon developing various types of acoustic feature vectors. The latest delivered research confirmed our vision for audio recognition. According to the literature,⁵⁷ the authors directly extracted acoustic features from speech spectra via encoders and combined them with visual features, with modeling the fused features in terms of BLSTM, the accuracy and reliability of AVSR remain optimal with the variation of noise and visual angle interferences.

Visual feature extraction. As noted in Figure 3, the tools applied to AVSR are now experiencing the transition from surface feature-based AAM/DCT to deep learning-based CNN. Stimulated by deep neural networks (DNNs) and large-scale databases, K Noda et al.⁵⁸ found the deep-level features enhancements for visual extractions of images in translational sliding motion and rotary motion. In their work, the lip feature and audio feature are, respectively, extracted by CNN and deep denoise autoencoder, while no pretrained models by landmarks are required.

With the increasing enthusiasm for CNN, some research focuses on issues of temporal and spatial correlations that are frequently overlooked in deep-level feature extractions toward images in motion. A Torfi et al.⁵⁹ made their first successful attempts to fulfill AVSR using 3D CNN. As we shall see in Figure 8, two paralleled DNNs are initially trained by audio stream and visual stream; as the trainings end, they will be coupled together in the last layer which is fully connected by contrastive cost criterion. Note that, a cube that is composed of six successive lip gray-scale images (used as input of visual stream) represents the temporal information, and the lip’s motion represents the spatial information. In this sense, the temporal sequence (frames) and spatial sequence (motions) concepts are in separately related, and the fusion of these two via 3D CNN reflects an abstract temporal and spatial correlation idea.

In addition, lipreading is highly related to AVSR, and the developments of AVSR have been greatly promoted by advances in lipreading. Take GRID with task type P (phrases) as reference data set, the authors summarize the performances of typical methods for visual feature extraction in terms of word recognition rate (WRR); more detailed descriptions associated with methods, models, database, tasks, and WRR are presented in Table 3.

As in Table 3, the worthy-noted WRR by K Xu’s lip region extraction strategy with 3D CNN is up to 97.1%. It is fair to say that typical DCT and AAM for AVSR are now being replaced by deep learning-based CNN tools.

Table 3. Comparative WRR in GRID data set.

Year	Authors	Method for visual feature extraction	Model	Database	Task	WRR (%)
2009	YX Lan et al. ⁶⁰	DCT/AAM	HMM	GRID	P	40.0/65.0
2016	YM Assael et al. ²⁶	3D CNN	Bi-GRU	GRID	P	93.4
2016	M Wand and Schmidhuber ⁶¹	Feed-forward	LSTM	GRID	P	79.5
2017	JS Chung et al. ⁶²	CNN	LSTM attention	GRID	P	97.0
2018	K Xu et al. ⁶³	3D CNN	Bi-GRU attention	GRID	P	97.1

CNN: convolutional neural network; Bi-GRU: bidirectional gated recurrent unit; LSTM: long short-time memory network; AAM: active appearance model; DCT: discrete cosine transform; HMM: hidden Markov model; WRR: word recognition rate.

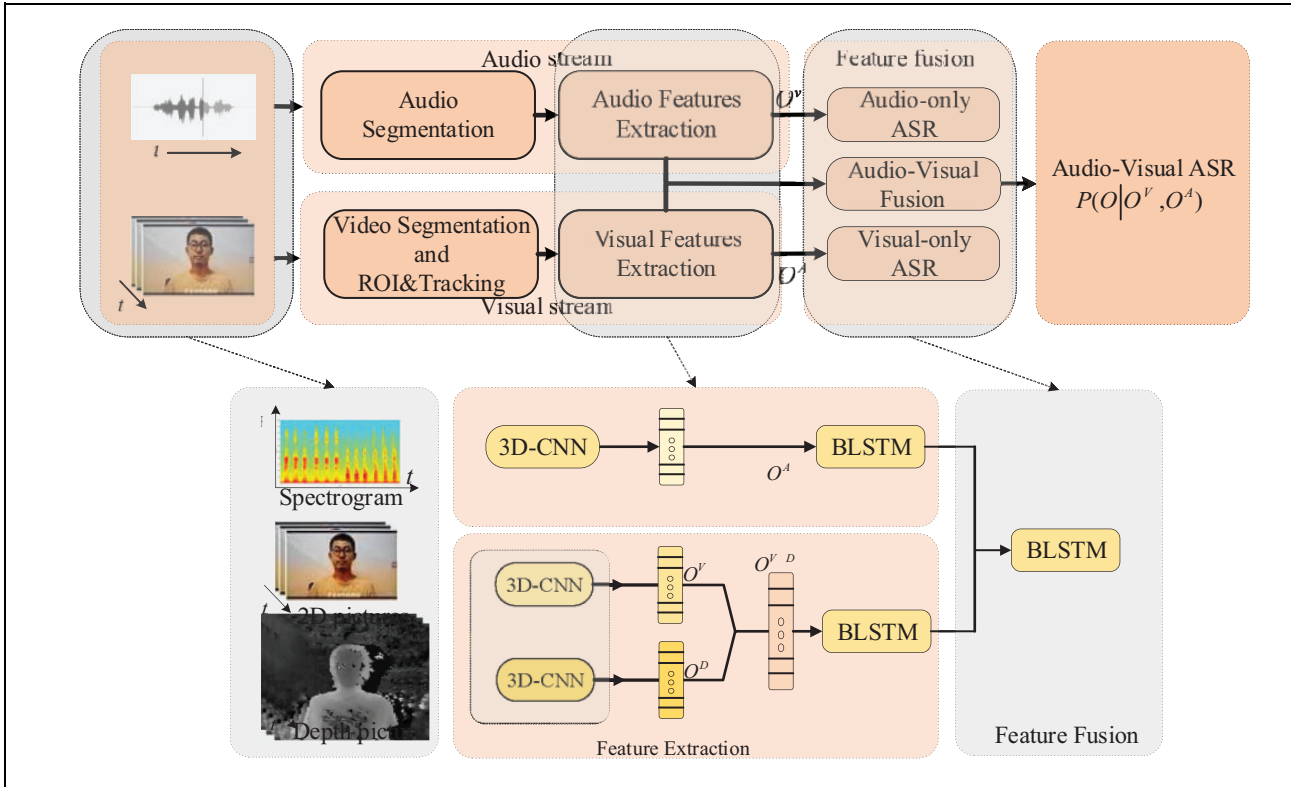


Figure 10. A possible AVSR architecture with BLSTM temporal sequence modeling. AVSR: audiovisual speech recognition; BLSTM: bidirectional long short-term memory.

Dynamic audiovisual fusion

The development and maturation of deep learning also opens the door to facilitate AVSR in terms of dynamic audiovisual fusion. S Petridis et al.⁵⁷ initially put forward an end-to-end audiovisual model that is based on bidirectional long short-term memory (BLSTM) networks, whose graphic interpretation is illustrated by an encoder layer and a classification layer in Figure 9. The functions of encoder layer are to extract features from raw images and spectrograms and to model their own temporal dynamics via a BLSTM network. The classification layer located above is mainly used to fuse the audio information that is derived from BLSTM networks in encoding layer. The final classification results (target classes) are given in terms of labels (one label per input frame, Softmax provides). By contrast,

the end-to-end audiovisual model presents a better classification performance over typical AVSR models in AVIC data set without any additional load in complexity.

In another similar research proposed by S Petridis et al.,⁶⁴ a new type of end-to-end model has been proposed. Whose model features are directly derived from waveforms and pixels, and the networks for modeling are bidirectional gated recurrent units (Bi-GRU) networks. Compared with the former models that mainly comprise of LSTM networks or encoders, this model may learn more knowledge on account of new form input “waveform.” However, more time is required for training. Overall, the quality of audio signals as the measurement for dynamic weight adjustments is no longer the only applicable solution to dynamic audiovisual fusion. Since the favorable input forms facilitate the learning abilities for parameter adjustments and

Table 4. Comparative analysis of typical case studies on AVSR.

Author	Database	Pose	Visual feature	Audio feature	Classifier	Fusion	Accuracy (%)			
							SNR	ASR	VSR	AVSR
JN Gowdy	CUAVE (training: 1200, test: 600)	Frontal	PCA (eye detection) 30-dimensional DCT2+ Δ	13 MFCC+ Δ + Δ	HMM (16 states, 4 Gaussian)	DF	-4 db	31	53.33	72.12
							4 db	65.57	53.33	82.67
							6 db	75.67	53.33	84.46
							10 db	84.82	53.33	90.00
							12 db	89.33	53.33	96.12
AS Saudi	CUAVE (training: 1200, test: 400)	Frontal	30 GVFs+ Δ	90 GAFS+ Δ	MSHMM (3-15 states, 1 Gaussian, dynamic stream weight adaption)	DF	Clean	98	53.33	98.62
							-10 db	75.56	69.23	80
							-5 db	81.11	69.23	85.71
							0 db	91.11	69.23	93.33
							5 db	92.22	69.23	94.44
S Petridis	LRW (training: 800-1000, test: 50, validation: 50)	Multi-view	Feature extraction directly from the raw images and audio waveforms Visual stream: 3D-Conv	Audio stream: ResNet,	Two-layer BGRU	—	10 db	94.17	69.23	95.83
							15 db	96.67	69.23	96.67
							20 db	97.78	69.23	97.78
							Clean	98.89	69.23	98.89
							End-to-end audio model	—	—	97.7
							End-to-end visual model	—	—	82.0
							End-to-end audiovisual model	—	—	98.0

GAF: Gabor audio feature; GVF: Gabor visual feature; DF: decision fusion; MSHMM: multi-stream hidden Markov model; MFCC: Mel-frequency cepstral coefficient; PCA: principal components analysis; AVSR: audiovisual speech recognition; HMM: hidden Markov model; DCT: discrete cosine transform; SNR: signal-to-noise ratio; BGRU: bidirectional gated recurrent unit.

further enhance the robustness of AVSR, the end-to-end models are supposed to be the mainstreams in the fields of predictable dynamic audiovisual fusion.

Future AVSR architecture design

In this part, we would like to conduct a comparative analysis of three representative case studies on AVSR and further illuminate our insights into the future AVSR architecture design. As in Table 4, the studies mainly concern typical MFCC/DCT-based audio/visual feature acquisition, adaptive weight adjustments for dynamic audiovisual fusion, and end-to-end AVSR design. The research of JN Gowdy et al. represents a class of traditional tools-based AVSR solutions. In their studies, the eyes are chosen as the references to locate the lip, and the 30-dimensional matrices via 2D DCT (and its first order differential transform) are adopted as the visual features, together with 13-dimensional MFCC features, they constitute the VSC and ASC of an AVSR, whose audiovisual fusion is performed at a decision level. It has been learned from the experiments that the lower the SNR, the higher the recognition rate of such an AVSR exhibits.¹² When MSHMM-based time sequence modeling method is invited as a tool for dynamic audiovisual fusion, this type of AVSR also follows the traditional designing lines of thought.

However, AS Saudi et al.⁵⁴ focus more upon the studies of efficient audio/visual feature extraction and adaptive weight adjustment of audio/visual features for dynamic fusion. In this AVSR, a Gabor filter is used for the audio/visual feature extractions, and a MLP-based adaptive weighting framework that jointly connects six-dimensional feature vector and an optimal weight is proposed. Differing from the typical audiovisual fusion on feature level or decision level, this dynamic weight adjusting scheme enables an AVSR to present optimal weights in terms of different SNR, and the average accurate rate of speech recognition has been tested to be increased by 5.5%.

Even though the overall performance of AVSR designed by AS Saudi has enhanced a lot, the problem remains of how to efficiently extract the audio/visual features from an engineering perspective. When compared with DNNs, the Gabor filter can hardly fulfill the deep-level audio/visual feature acquisition; meanwhile, the MSHMM-based time sequence modeling also faces challenges such as nondeep-level modeling, repeated trainings, poor practicality, and so on. It is fair to say that the end-to-end model is favorable for the overall progress of AVSR. The experimental results verify its effectiveness even though the audio signal is under noising attacks by babble noise whose SNR ranges from -5 db to 20 db, and the accuracy of speech recognition has been numerically validated by a 14.1% (-5 db SNR) increase with comparison to that of ASR under the same conditions (more precisely, a 1.3% and 3.9% increases, respectively, correspond to 5 db SNR and 0 db SNR).⁶⁴

At present, the end-to-end AVSR model generally takes two of many forms of modality streams, for example, pixels, spectrogram, waveforms, and so on, as illustrated by the two-channel model in Figure 9. Taking such an AVSR (proposed by S Petridis) as an example, the training module of which directly extracts desired features from pixels and speech waveforms, and a two-layer classifier “bidirectional gated recurrent unit” is invited to classify the training results. When compared with the previous research, quite clear, the end-to-end-based AVSR adopts much larger data sets; also, it adapts to a much wider range of lip views (indicating that more angles of view are available).

Shrivastava et al.’s study⁶⁵ also discusses the end-to-end networks (MobiVSR) in lip-reading domain when the efficient lipreading itself is restricted to some cases with limited computational resources, like mobile devices, embedded systems, and so on. It has been proved that MobiVSR efficiently balances the speech accuracy and parameters involved, and this end-to-end design takes up less memory resources. These clues make us believe, the end-to-end AVSR has broad prospects. Also, we believe that the future AVSR would not be limited to two-channel information (such as visual and auditory modalities) fusion, videlicet, multimodality mechanism would be possible.

Refer to the physical structures of end-to-end networks, let’s consider an optimized framework that integrates superior audio feature estimation and the matched visual feature results. We present a possible AVSR architecture (see Figure 10). For the ASC, 3D CNNs may be chosen so that, as far as possible, the spectrogram exhibits as more comprehensive feature extraction source so replacing the features that are denoted by MFCC. For the multi-view lip-reading-based VSC, the pixel and depth used for image description could be extended to deal with the visual feature representation, so that the 3D structure understanding of lip spatial motion could be enhanced accordingly. BLSTM network is a solution to the audiovisual modality-based temporal sequence modeling. The major concerns of computational load, accuracy, and applicability should be given in future studies because these aspects are so essential. We may predicate, on basis of the maturing audiovisual recognition and representation, the further “sight-listening-touching-tasting-smelling” multimodality interaction tools will push the typical HCI to the direction of naturally complementary supervision pattern. Also, the fully developed AVSR that is applied to a range as small as voice inputting of mobile phones, lip-reading password, and biometric authentication or as large as robot, smart home, and in-vehicle scenarios will dramatically promote the development and maturation of HCI.

Conclusion

AVSR that used to guide the dialogue between humans and machines is now attracting more and more scientific

attention. In this review study, several problems that confront the designers of such complex speech recognition systems have been stated and elaborately discussed. From our point of view, finding solutions to large-scaled public databases for general purpose, efficient audio/visual feature representation, eminent feature extraction, and intelligent dynamic audiovisual fusion allows AVSR to move forward. Also, the evolution of such techniques has apparently exploited the opportunity for researchers to invite the deep learning-based tools in developing AVSR's algorithmic frameworks. One of the important contributions of this review is the attempt that has been made to describe a possible AVSR architecture in foreseeable future. As the comparative analysis of typical case studies presents, we hold that deep learning and AVSR are now inseparably related, and the favorable anti-noise performances of end-to-end AVSR model and deep-level feature extraction capacities of deep learning-based feature extractors will guide a class of multimodality HCI directly to a solution.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by Science and Technology Program of Department of Education, Jilin Province under Grant JJKH20200117KJ and Research Fund for Distinguished Young Scholars of Jilin City under Grant 20190104128.

ORCID iD

Linlin Xia  <https://orcid.org/0000-0002-5079-3788>

References

1. Panda SP. Automated speech recognition system in advancement of human-computer interaction. In: *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Erode, India, 18–19 July 2017, pp. 302–306. New York, NY, United States: IEEE.
2. McGurk H and MacDonald J. Hearing lips and seeing voices. *Nature* 1976; 264(5588): 746–748.
3. Dupont S and Luetttin J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans Multimed* 2000; 2(3): 141–151.
4. Biswas A, Sahu PK, and Chandra M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *Int J Speech Technol* 2016; 19(1): 159–171.
5. Koguchi Y, Oharada K, Takagi Y, et al. A mobile command input through vowel lip shape recognition. In: *International conference on human-computer interaction*, Las Vegas, NEVADA, US, 15–20 July, 2018. 15 July 2018, pp. 297–305. Cham, Switzerland: Springer.
6. Sun K, Yu C, Shi WN, et al. Lip-interact: Improving mobile device interaction with silent speech commands. In: *The 31st annual ACM symposium on user interface software and technology*. ACM, Berlin, Germany, 14–17 October 2018, pp. 581–593. New York, NY, United States: ACM.
7. Jang SB, Kim YG, and Ko Y. Mobile video communication based on augmented reality. *Multimed Tools Appl* 2017; 76(16): 893–909.
8. Mohammed A, Mansour A, Ghulam M, et al. Automatic speech recognition of pathological voice. *Indian J Sci Technol* 2015; 8(32): 1–6.
9. Afouras T, Chung JS, and Zisserman A. The conversation: deep audio-visual speech enhancement. *arXiv preprint arXiv: 1804.04121* 2018.
10. Ding CX and Tao DC. Robust face recognition via multi-modal deep face representation. *IEEE Trans Multimed* 2015; 17(11): 2049–2058.
11. Zhang WM, Zhao X, Morvan JM, et al. Improving shadow suppression for illumination robust face recognition. *IEEE Trans Pattern Anal Mach Intell* 2018; 41(3): 611–624.
12. Gowdy JN, Subramanya A, Bartels C, et al. DBN based multi-stream models for audio-visual speech recognition. In: *2004 IEEE international conference on acoustics, speech, and signal processing*. Montreal, Canada, 17–21 May 2004, pp. I–993. Washington, D.C. United States: IEEE.
13. Nainan S and Kulkarni V. Lip tracking using deformable models and geometric approaches. In: *Information and communication technology for intelligent systems*. Singapore: Springer, 2019, pp. 655–663.
14. Eveno N, Caplier A, and Coulon P. Automatic and accurate lip tracking. *IEEE Trans Circuit Syst Video Technol* 2004; 14(5): 706–715.
15. Wang XP, Hao YF, Fu DG, et al. Roi processing for visual features extraction in lip-reading. In: *2008 international conference on neural networks and signal processing*. Nanjing, China, 7–11 June 2008, pp. 178–181. Washington, D.C. United States: IEEE.
16. Ibrahim M, Mulvaney DJ, and Abas M. Feature-fusion based audio-visual speech recognition using lip geometry features in noisy environment. *ARPJ Eng Appl Sci* 2015; 10(23): 17521–17527.
17. Sumby WH and Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 1954; 26(2): 212–215.
18. Petajan Eeic D. Automatic lipreading to enhance speech recognition. In: *Proceeding CVPR'85*. San Francisco, United States, June, 1985, pp. 3582–3582. New York, NY, United States: IEEE.
19. Abboud B and Chollet G. Appearance based lip tracking and cloning on speaking faces. In: *ISPA 2005. Proceedings of the 4th international symposium on image and signal processing and analysis*. Zagreb, Croatia, 15–17 September 2005, pp. 301–305. Washington, D.C. United States: IEEE.

20. Saenko K, Darrell T, and Glass JR. Articulatory features for robust visual speech recognition. In: *Proceedings of the 6th international conference on multimodal interfaces*. State College, PA, 14-15, October, 2004, pp. 152–158. New York, NY, United States: ACM.
21. Drugman T, Gurban M, and Thiran J. Relevant feature selection for audio-visual speech recognition. In: *2007 IEEE 9th workshop on multimedia signal processing*. Crete, Greece, 1–3 October 2007, pp. 179–182. Washington, D.C. United States: IEEE.
22. Hong XP, Yao HX, Wan YQ, et al. A PCA based visual DCT feature extraction method for lip-reading. In: *2006 international conference on intelligent information hiding and multimedia*. Pasadena, CA, USA, 18-20 December, 2006, pp. 321–326. New York, NY, United States: IEEE.
23. Biswas A, Sahu P, Bhowmick A, et al. Audio visual isolated Oriya digit recognition using HMM and DWT. In: *Proceedings of the conference on advances in communication and control system*. DIT University, 6-8 April 2013, pp. 234–238. Paris, France: Atlantis Press.
24. Huang J, Marcheret E, and Visweswariah K. Rapid feature space speaker adaptation for multi-stream HMM-based audio-visual speech recognition. In: *2005 IEEE international conference on multimedia and expo*. Amsterdam, the Netherlands, 6 July 2005, pp. 338–341. New York, NY, United States: IEEE.
25. Yang S and Guan YP. Fusion strategy-based multimodal human-computer interaction. *Int J Computat Vis Robot* 2018; 8(3): 300–317.
26. Assael YM, Shillingford B, Whiteson S, et al. LipNet: end-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* 2016.
27. Zhou P, Yang WW, Chen W, et al. Modality attention for end-to-end audio-visual speech recognition. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK, 12–17 May 2019, pp. 6565–6569. Washington, D.C. United States: IEEE.
28. Su RF, Wang L, and Liu XY. Multimodal learning using 3D audio-visual data for audio-visual speech recognition. In: *2017 International Conference on Asian Language Processing (IALP)*. Singapore, 5–7 December 2017, pp. 40–43. Washington, D.C. United States: IEEE.
29. Liu YH, Liu X, Fan WT, et al. Efficient audio-visual speaker recognition via deep heterogeneous feature fusion. In: *Chinese Conference on Biometric Recognition*. Shenzhen, China, 28–29 October 2017, pp. 575–583. Cham, Switzerland: Springer.
30. Xing XM, Sun Q, Zhang PY, et al. Research on distribution network fault recovery and reconstruction based on deep-first search and colony algorithms. *J Northeast Electr Power Univ* 2019; 39(03): 38–43.
31. Sun B, Qiao C, Yang D, et al. Prediction method of thermal conductivity of nanofluids based on deep belief network. *J Northeast Electr Power Univ* 2019; 39(01): 41–48.
32. Li PS, Liu X, and Li JD. Application study of self-organizing map network based on immune genetic algorithm. *J Northeast Electr Power Univ* 2018; 38(06): 82–85.
33. Matthews I, Cootes T, Cox S, et al. Lipreading using shape, shading and scale. In: *AVSP'98 International conference on auditory-visual speech processing*, Moreton Island, Australia, 4–6 December 1998. Terrigal, New South Wales, Australia: AVISA.
34. Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition. *J Acoust Soc Am* 2006; 120(5): 2421–2424.
35. Cox S, Harvey R, and Lan YX. The challenge of multi speaker lip-reading. In: *AVSP*. Brisbane, Australia, 26 September 2008, pp. 179–184. Terrigal, New South Wales, Australia: AVISA.
36. Zhao GY, Barnard M, and Pietikainen M. Lipreading with local spatiotemporal descriptors. *IEEE Trans Multimed* 2009; 11(7): 1254–1265.
37. Afouras T, Chung JS, and Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* 2018.
38. Yang S, Zhang YH, Feng DL, et al. LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild. In: *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. Lille, France, 14–18 May 2019, pp. 1–8. Washington, D.C. United States: IEEE.
39. Su RF, Liu XY, Wang L, et al. Cross-domain deep visual feature generation for mandarin audio-visual speech recognition. *IEEE/ACM Trans Audio Speech Language Process* 2019; 28: 185–197.
40. Hsieh HC, Zheng WZ, Chen KC, et al. Consonant classification in Mandarin based on the depth image feature: a pilot study. In: *The annual conference of the international speech communication association, INTERSPEECH*. Phoenix, Arizona, USA, 22-26 June 2019, pp. 2300–2304. New York, NY, USA: ACM.
41. Patterson EK, Gurbuz S, Tufekci Z, et al. CUAVE: a new audio-visual database for multimodal human-computer interface research. In: *2002 IEEE international conference on acoustics, speech, and signal processing*. Orlando, FL, 13–17 May 2002, Vol. 2, pp. II–2017. Washington, D.C. United States: IEEE.
42. Lee B, HasegawaJohnson M, Goudeseune C, et al. AVICAR: audio-visual speech corpus in a car environment. In: *Eighth international conference on spoken language processing*. Jeju Island, Korea, 4–8 October 2004. Terrigal, New South Wales, Australia: AVISA.
43. Lucey PJ, Potamianos G, and Sridharan S. Patch-based analysis of visual speech from multiple views., Moreton Island, Australia, 2008. pp. 69–74. Terrigal, New South Wales, Australia: AVISA.
44. Lin XX, Yao HX, Hong XP, et al. HIT-AVDB-II: a new multi-view and extreme feature cases contained audio-visual database for biometrics. In: *11th joint international conference on information sciences*. Shanghai, China, December 2008. Atlantis Press.

45. Pass A, Zhang JG, and Stewart D. An investigation into features for multi-view lipreading. In: *2010 IEEE international conference on image processing*. Hong Kong, China, 26–29 September 2010, pp. 2417–2420. Washington, D.C. United States: IEEE.
46. Lan YX, Theobald B, Harvey R, et al. Improving visual features for lip-reading. In: *Auditory-visual speech processing 2010*, Hakone, Kanagawa, Japan, 30 September–3 October 2010. Terrigal, New South Wales, Australia: AVISA.
47. Anina I, Zhou Z, Zhao G, et al. OuluVS2: a multi-view audio-visual database for non-rigid mouth motion analysis. In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Ljubljana, Slovenia, 4–8 May 2015, vol. 1, pp. 1–5. Washington, D.C. United States: IEEE.
48. Harte N and Gillen E. TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans Multimed* 2015; 17(5): 603–615.
49. Petridis S, Shen J, Cetin D, et al. Visual-only recognition of normal, whispered and silent speech. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Calgary, Canada, 15–20 April 2018, pp. 6219–6223. Washington, D.C. United States: IEEE.
50. Ahmed N. RGB-D dynamic facial dataset capture for visual speech recognition. In: Xu Huaiyu (ed). *2019 International conference on image and video processing, and artificial intelligence*. Shanghai, China, 27 November 2019, pp. 1132108(1)–1132108(4). Bellingham, WA, USA: SPIE.
51. Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition. *Int J Adv Res Eng Technol* 2013; 1(6): 1–4.
52. Biswas A, Sahu PK, and Chandra M. Multiple camera in car audio-visual speech recognition using phonetic and visemic information. *Comput Electr Eng* 2015; 47: 35–50.
53. Hu JP, Li L, Xie Q, et al. A novel segmentation approach for glass insulators in aerial images. *J Northeast Electr Power Univ* 2018; 38(2): 87–92.
54. Saudi AS, Khalil MI, and Abbas HM. Improved features and dynamic stream weight adaption for robust audio-visual speech recognition framework. *Digit Signal Process* 2019; 89: 17–29.
55. Petridis S and Pantic M. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. *IEEE Trans Affect Comput* 2015; 7(10): 45–58.
56. Wang HK, Pan J, and Liu C. Research development and forecast of automatic speech recognition technologies. *Telecommun Sci* 2018; 34(2): 1–11.
57. Petridis S, Wang YJ, Li ZW, et al. End-to-end audiovisual fusion with LSTMS. *arXiv preprint arXiv:1709.04343* 2017.
58. Noda K, Yamaguchi Y, Nakadai K, et al. Audio-visual speech recognition using deep learning. *Appl Intell* 2015; 42(4): 722–737.
59. Torfi A, Iranmanesh SM, Nasrabadi N, et al. 3D convolutional neural networks for cross audio-visual matching recognition. *IEEE Access* 2017; 5: 22081–22091.
60. Lan YX, Harvey R, Theobald B, et al. Comparing visual features for lipreading. In: *International conference on auditory-visual speech processing*. Norwich, England, 10–13 September 2009, pp. 102–106. Terrigal, New South Wales, Australia.
61. Wand M and Schmidhuber J. Improving speaker-independent lipreading with domain-adversarial training. *arXiv preprint arXiv:1708.01565* 2017.
62. Chung JS, Senior A, Vinyals O, et al. Lip reading sentences in the wild. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Honolulu, HI, 21–26 July 2017, pp. 3444–3453. Washington, D.C. United States: IEEE.
63. Xu K, Li DW, Cassimatis N, et al. LCANet: end-to-end lipreading with cascaded attention-CTC. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. Xi'an, China, 15–19 May 2018, pp. 548–555. Washington, D.C. United States: IEEE.
64. Petridis S, Stafylakis T, Ma P, et al. End-to-end audiovisual speech recognition. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Calgary, Canada, 15–20 April 2018, pp. 6548–6552. Washington, D.C. United States: IEEE.
65. Shrivastava N, Saxena A, Kumar Y, et al. MobiVSR: a visual speech recognition solution for mobile devices. *arXiv preprint arXiv: 1905.03968* 2019.