Image and Vision Computing 32 (2014) 590–605

Contents lists available at ScienceDirect

Image and Vision Computing

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

Editor's Choice Article

A review of recent advances in visual speech decoding[☆]

Ziheng Zhou [*], Guoying Zhao, Xiaopeng Hong, Matti Pietikäinen

Center for Machine Vision Research, Computer Science and Engineering, University of Oulu, Oulu, Finland

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form 16 May 2014
Accepted 26 June 2014
Available online 3 July 2014

Keywords:
Visual speech decoding
Automatic speech recognition
Lip-reading
Audio-visual speech recognition
Review

ABSTRACT

Visual speech information plays an important role in automatic speech recognition (ASR) especially when audio is corrupted or even inaccessible. Despite the success of audio-based ASR, the problem of visual speech decoding remains widely open. This paper provides a detailed review of recent advances in this research area. In comparison with the previous survey [97] which covers the whole ASR system that uses visual speech information, we focus on the important questions asked by researchers and summarize the recent studies that attempt to answer them. In particular, there are three questions related to the extraction of visual features, concerning speaker dependency, pose variation and temporal information, respectively. Another question is about audio-visual speech fusion, considering the dynamic changes of modality reliabilities encountered in practice. In addition, the state-of-the-art on facial landmark localization is briefly introduced in this paper. Those advanced techniques can be used to improve the region-of-interest detection, but have been largely ignored when building a visual-based ASR system. We also provide details of audio-visual speech databases. Finally, we discuss the remaining challenges and offer our insights into the future research on visual speech decoding.

Contents

1. Introduction . . . 591
2. Region of interest . . . 591
3. Visual feature extraction . . . 593
   3.1. Speaker dependency . . . 593
   3.2. Pose variation . . . 594
   3.3. Temporal information . . . 596
4. Dynamic audio-visual fusion . . . 597
5. Audio-visual speech databases . . . 599
   5.1. AVICAR . . . 599
   5.2. AVLetters . . . 599
   5.3. AVLetter2 . . . 599
   5.4. AV-TIMIT . . . 599
   5.5. CUAVE . . . 600
   5.6. Grid . . . 600
   5.7. IBMIH . . . 600
   5.8. IBMSR . . . 600
   5.9. LILiR . . . 600
   5.10. MOBIO . . . 600
   5.11. OuluVS . . . 600
   5.12. XM2VTSDB . . . 600
6. Discussions . . . 600
   6.1. Visual feature extraction . . . 600
   6.2. Dynamic audio-visual speech fusion . . . 602
   6.3. Audio-visual speech databases . . . 603

[☆] Editor's Choice Articles are invited and handled by a select rotating 12 member Editorial Board committee. This paper has been recommended for acceptance by Vladimir Pavlovic.

[*] Corresponding author.
E-mail addresses: ziheng.zhou@ee.oulu.fi (Z. Zhou), guoying.zhao@ee.oulu.fi (G. Zhao), xiaopeng.hong@ee.oulu.fi (X. Hong), mkp@ee.oulu.fi (M. Pietikäinen).

http://dx.doi.org/10.1016/j.imavis.2014.06.004
0262-8856/© 2014 Elsevier B.V. All rights reserved.

7. Conclusion . . . 603
Acknowledgment . . . 603
References . . . 603

1. Introduction

It is well known that speech perception is a bi-modal process that takes into account both the acoustic and visual speech information [70]. There has been clear evidence that visual information plays a key role in automatic speech recognition when audio is corrupted by, for example, background noise, or even inaccessible [97]. During the last two decades, there have been significant advances in the research of audio-based ASR [119], resulting in various commercial systems. Many believed that *visual speech decoding* would be relatively easily done following the success of audio-based ASR. However, early attempts did not achieve the anticipated results. As reported by Potamianos et al. [95], visual features provided very weak speech information in a large vocabulary continuous speech recognition (LVCSR) task. Despite the poor performance, the visual features still helped boost the ASR performance on some low-quality audio data through audio-visual (AV) speech information fusion.

Since then, we have seen research focused on two major tasks: extracting better visual features and developing a better AV fusion scheme. For feature extraction, three particular questions have been asked and tackled recently:

> *How to deal with speaker dependency in visual data*? Visual data are typically stored as video sequences which contain a lot of information that is irrelevant to the uttered speech. Such information includes the variability of the visual appearances among different speakers. The first question addresses this important issue and is mainly concerned with the methods that attempt to suppress the speaker-dependent information in visual features.
>
> *How to cope with head-pose variation*? This question arises naturally when building a practical ASR system that uses visual information. In a real-world situation, it is unreasonable to assume that users would face the video camera all the time during speaking. Therefore, there could be various head poses in the captured visual data. Since the visual appearance of a talking mouth could vary significantly in images due to the view change, pose variation poses a serious challenge to any practical ASR system.
>
> *How to encode temporal information in visual features*? At first glance, this question seems less important. The temporal information of a talking mouth might well be characterized by statistical models (e.g., HMMs) built upon visual features. However, a number of studies [95,120,123] show that the use of statistical models alone may not be sufficient to capture the video dynamics. A solution is to encode temporal information to improve informativeness and stability of the extracted visual features.

There have been various models proposed for fusing AV speech information in the past [95,97]. The recent research has been focused on answering the practical question: *How to automatically adapt the fusion rule when the quality of the two individual (audio and visual) modalities varies*? It is a critical question since there are many factors that could affect the quality. For instance, some abrupt background noise could significantly worsen the quality of the recorded audio, making the acoustic information unreliable temporally. In addition, a sudden loss of tracking of the speaker's facial landmarks could result in the failure of localizing the talking mouth, making the extracted visual features less informative. Therefore, it is important to develop a dynamic AV fusion scheme that is capable of handling the varying quality.

In this paper, we review the recent studies concerning visual speech decoding. They are mostly organized and described with respect to the particular questions raised above. In addition, attentions are given to the state-of-the-art approaches to facial landmark localization. They can be used to accurately extract the region of the talking mouth, which is utterly important for any ASR system that uses visual speech information. However, we feel that they have been largely ignored in the recent work and would like to promote the use of these advanced methods in the future. Note that it is NOT our intention to provide a comprehensive review that covers every aspect of visual-based ASR and replaces previous surveys [95,97].

Instead, this paper is aimed to serve as a supplement to them and focus on the recent advances in the area of visual speech decoding.

This paper is organized as follows: Section 2 introduces some recently developed methods for facial landmark localization. Section 3 provides a detailed review of the recent studies that attempt to design better visual features. Section 4 describes the development of dynamic AV speech fusion schemes. AV speech databases are described in Section 5. In Section 6, we discuss the challenges and provide our vision for future research. Finally, Section 7 concludes this paper.

2. Region of interest

Given a video of a talking face, to extract useful visual speech information, the first step is to locate the mouth region that contains the motion relevant to speech, or in other words, the region-of-interest (ROI). It is important since the quality of ROIs could significantly affect the speech recognition performance [97]. In general, we need to localize certain facial landmarks, such as eye corners, nostrils and lip corners, to correctly box the talking mouth in images. The ROI can then be cropped off and its size normalized for further visual feature extraction. Tracking facial landmarks has been an active research topic in recent years due to its wide applications to face-related vision problems. The state-of-the-art systems have been shown to be capable of accurately tracking facial points under various image qualities, head poses, facial expressions and partial occlusions.

Unfortunately, the recent advances in facial landmark localization have largely been ignored by researchers in the speech recognition community especially when building up ASR systems using visual speech information. To extract ROIs, most of the systems relied on the active appearance model (AAM)[1] [98,15,86,50,49,78], skin color thresholding [61,24], Haar-like feature based boosted classification framework[2] [59,60,62,26,120,124,125,123] or other lip-region classifiers based on linear discriminant analysis (LDA) [31] or support vector machines (SVMs) [103]. These methods are often heuristic and lack either tracking accuracy or capabilities of generalizing to new faces and handling large pose and illumination changes which are often encountered in a real-world environment. In this section, we briefly introduce some of the techniques recently developed for robust facial landmark localization. Note that it is NOT our intention to provide a comprehensive survey on this important and active topic that has a large literature. We expect that future research concerning visual speech decoding would benefit from the described techniques in terms of better ROI detection and extraction.

The methods included in this section can be broadly grouped into two categories: the point-distribution-model (PDM) based and non-

[1] See [12,13,67] for details.

[2] First introduced by Viola and Jones [114] and extended by Leinhart and Maydt [53].

PDM based. Here we first describe the PDM-based methods. Let x_i denote the location of the ith facial landmark and they are very often modeled by the *point distribution model* introduced by Coote and Taylor [14]:

$$\mathbf{x}_i = s\mathbf{R}(\overline{\mathbf{x}}_i + \Phi_i\mathbf{q}) + \mathbf{t} \tag{1}$$

where the shape formed by the facial landmarks is modeled by a rigid transformation defined by scale s, rotation $\mathbf{R}$ and translation $\mathbf{t}$, and a non-rigid transformation by $\mathbf{q}$. Here $\overline{\mathbf{x}}_i$ denotes the mean location of the ith landmark in the training set and Φi the corresponding submatrix of the basis. Given a test image $\mathcal{I}$, from a probabilistic point of view, the objective is to maximize the posterior of the PDM parameter $\mathbf{p} = \{\mathrm{s},\mathbf{R},\mathbf{t},\mathbf{q}\}$. Saragih et al. [106] defined the posterior as

$$p\left(\mathbf{p}|\{l_i = 1\}_{i=1}^{n}, \mathcal{I}\right) \propto p(\mathbf{p}) \prod_{i=1}^{n} p(l_i = 1|\mathbf{x}_i, \mathcal{I}) \tag{2}$$

where l_i is a binary variable denoting whether the ith landmark is correctly located. The problem can then be formulated as

$$\begin{aligned} \mathbf{p}^* &= \arg\min_{\mathbf{p}} \mathcal{Q}(\mathbf{p}) \\ &= \arg\min_{\mathbf{p}} \left\{ -\sum_{i=1}^{n} p(l_i = 1|\mathbf{x}_i, \mathcal{I}) - \log p(\mathbf{p}) \right\}. \end{aligned} \tag{3}$$

The shape prior $p(\mathbf{p})$ is often defined as a Gaussian with a diagonal covariance whose non-zero entries are set as the eigenvalues of the modes of the non-rigid deformation. Parameters $\mathbf{p}$ are often optimized in an iterative manner. In each iteration, parameter update $\Delta\mathbf{p}$ that leads to a local minimum of the lost function $\mathcal{Q}(\mathbf{p})$ starting from the current values of $\mathbf{p}$ is calculated. After that, $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$. This process ends when $\mathbf{p}$ converges or the maximum number of iterations is reached. We will describe some recent systems built within such a framework below.

In [17], Cristinacce and Cootes built a linear appearance model for the patches sampled at landmarks, similar to AAM. During each iteration of parameter updating, the appearance model generated a set of templates that best described the patches sampled at x_i. The likelihood took the form $p(l_i = 1|\mathbf{x}_i, \mathcal{I}) \propto \exp\{-\alpha R_i\}$ where R_i was the normalized correlation response of the ith template.

Gu and Kanade [33] modeled $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$ using the Gaussian mixture model (GMM) to account for the possible multiple modes. The modes were chosen as the K-largest responses in the response map calculated around $\mathbf{x}_i$. Saragih et al. [106] represented $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$ as a kernel density estimate upon a set of candidate locations around x_i. Instead of modeling the posterior, Asthana et al. [4] took a discriminative regression based approach. They proposed to learn a set of functions to predict $\Delta\mathbf{p}$ from the response maps calculated around $\{x_i\}$. These functions were trained one by one such that the current function targeted the more difficult samples that previous ones failed to predict the correct $\Delta\mathbf{p}$. During each iteration, $\Delta\mathbf{p}$ was estimated by the function that maximized the sum of the responses measured at the updated $\{x_i\}$.

The non-PDM based methods tend to estimate $X = \{x_i\}$ directly from $\mathcal{I}$. Belhumeur et al. [5] adopted a Bayesian approach to tackle the problem. An SVM was trained as a local detector for each landmark and the problem was formulated as

$$\mathbf{X}^* = \arg\max_{\mathbf{x}} p(\mathbf{X}|\mathbf{D}) \tag{4}$$

where $\mathbf{D}$ denoted the detector responses. They assumed that $\mathbf{X}$ could be generated from the global model $\mathbf{X}^{k,t}$ which was the kth labeled sample transformed by similarity transformation t. The posterior was then computed as

$$p(\mathbf{X}|\mathbf{D}) = \sum_k \int_{t \in T} p\left(\mathbf{X}|\mathbf{X}^{k,t}, \mathbf{D}\right) p\left(\mathbf{X}^{k,t}|\mathbf{D}\right) \mathrm{d}t. \tag{5}$$

During optimization, a RANSAC-like procedure was introduced to select a suitable subset of $\{\mathbf{X}^{k,t}\}$ to lessen the computational burden.

Ong and Bowden [83] proposed a person-specific tracking system based on linear predictors (LPs). An LP is defined by a reference point c surrounded by a set of support positions, a linear mapping $\mathbf{H}$, the base support pixel values $\mathbf{v}$ and bias $\mathbf{b}$. Given image $\mathscr{I}$, the displacement of c is predicted by $\mathbf{H}\delta\mathbf{v} + \mathbf{b}$ where $\delta\mathbf{v}$ is the difference between $\mathbf{v}$ and the pixel values at the support positions in $\mathscr{I}$. A number of LPs (a flock of LPs) were used to predict one point simultaneously and the output of the flock defined as the displacement averaged over all the LPs. For each facial point, two flocks were used to predict the horizontal and vertical displacements, respectively. Instead of randomly placing LPs, an iterative training step was introduced for selecting LPs within a flock based on their displacement prediction mean errors from training ground-truth data.

Zhu and Ramanan [126] proposed to use a mixture of trees to capture the appearance and shape variations of facial landmarks. In particular, landmarks were connected such that they formed a tree. Each tree in the model encoded the landmark topology from a particular view. For mixture m, they defined the cost function

$$\mathcal{Q}^m(\mathbf{X}, \mathcal{I}) = \mathcal{A}^m(\mathbf{X}, \mathcal{I}) + \mathcal{S}^m(\mathbf{X}) + \alpha^m. \tag{6}$$

where $\mathcal{A}^m(\mathbf{X}, \mathcal{I})$ was the appearance evidence, $\mathcal{S}^m(\mathbf{X})$ the spatial configuration score and α^m a bias term.

Function $\mathcal{Q}^m$ was minimized over m and $\mathbf{X}$ to locate landmarks. Zhao et al. [122] also used a tree structure for landmark representation and a similar cost function. A cascaded strategy was proposed to prune the shape space efficiently. They first trained some AdBoost classifiers to reject the false positions for individual landmarks. They then further constrained the search space such that the true shape was assumed to be near one of the shapes in training samples.

Martinez et al. [65] used a regression-based approach to localize facial landmarks. Two support vector regressors were learned to predict the horizontal and vertical displacements from a test location to the true one for every landmark. Landmarks were updated iteratively. Within the kth iteration, one test location was sampled from a sampling region for each $\mathbf{x}_i$ and its prediction $\hat{\mathbf{t}}_i^k$ together with previous predictions $\left\{\hat{\mathbf{t}}_i^j\right\}_{j=1}^{k-1}$ were used to approximate the true distribution of $\mathbf{x}_i$ as $\sum_j \mathcal{N}\left(\mathbf{x}_i|\hat{\mathbf{t}}_i^j, \Sigma\right)$. The mode of the distribution was used as the estimate of the true location so far.

A Markov random field was trained to impose global shape constraints on all the centers of the sampling regions.

Xiong and De la Torre [116] formulated facial landmark localization as a non-linear least square problem. The positions of facial landmarks $\mathbf{X}$ are updated iteratively. In the kth iteration, $\mathbf{X}_k$ was computed as:

$$\mathbf{X}_k = \mathbf{X}_{k-1} + \mathbf{R}_k\phi_{k-1} + \mathbf{b}_{k-1} \tag{7}$$

where ϕ_{k-1} was the SIFT features extracted at $\mathbf{X}_{k-1}$ and $\mathbf{R}_k$ and $\mathbf{b}_{k-1}$ parameters learned from training data.

The above described methods share the same principle that the non-rigid shape of facial landmarks is recovered based on evidence gathered from different local image patches. Such an evidence-gathering mechanism allows us to incorporate robust patch experts (e.g., the HOG [18] and SIFT [57] descriptor) and to train discriminative models that generalize well to unseen data. On the contrary, the widely used AAM and its variants [67,87,55,105] rely on a linear statistical model to generate the whole face texture. Such a model is often insufficient to represent the variations due to changes in identity, facial expression, pose and illumination [4].

Finally, we list in Table 1 the datasets used to evaluate the above-mentioned methods. It indicates, to some extent, their capabilities of

Table 1
Summary of the datasets used to evaluate the described methods for facial landmark localization. Symbol * marks those collected for facial expression recognition. A dataset listed without mentioning the view range means no restrictions applied on the set during testing.

Method	Evaluation datasets (view range)
Cristinacce and Cootes [17]	BioID [43] XM2VTS [73] (frontal)
Gu and Kanade [33]	Multi-PIE (frontal) AR [64]
Saragih et al. [106]	Multi-PIE [32] (frontal) XM2VTS (frontal) LFW [40]
Asthana et al. [4]	Multi-PIE (−30°, 30°) XM2VTS (frontal) LFPW [5]
Belhumeur et al. [5]	LFPW BioID
Ong and Bowden et al. [83]	ASL [20]
Zhu and Ramanan [126]	Multi-PIE AFW [126]
Zhao et al. [122]	BioID LFW
Martinez et al. [65]	*MMI [113] FERET [92] XM2VTS (frontal) BioID *SEMAINE [71] Multi-PIE (−30°, 30°)
Xiong and De la Torre [116]	LFPW LFW

handling factors such as image qualities, illumination changes, pose variations and facial expressions, as presented in the datasets.

3. Visual feature extraction

Despite of the many years of research, we have not seen any visual feature set universally accepted for representing visual speech, in contrast to the well-established features (e.g., MFCC [117]) for acoustic speech. Ideally, the extracted visual features should be relatively compact and sufficiently informative regarding the uttered speech, meanwhile showing a certain level of invariance against irrelevant information or noise in videos. It is a challenging problem largely due to the facts that there are uncertainties (e.g., speaker identity and head pose) that could significantly affect the visual appearance of a talking mouth in images and that visual features are extracted to describe a dynamic process (uttering) rather than static images.

Traditionally, most of the feature extraction methods fall under one or some of the following categories as summarized by Dupont and Luettin [22]:

1. *Image-based* [8,93,31,103,39] — raw pixel values are either used directly or undergone some image transformation as visual features.
2. *Motion-based* [66,118] — features are designed to describe the motion observed during uttering.
3. *Geometric-feature-based* [9,3,77,10] — geometric information of the talking mouth (e.g., the width and height of the mouth opening) is extracted as features.
4. *Model-based* [68,25,48,50] — a model of the visible articulators is built and the compact model parameters are used as visual features.

In this work, however, we categorize the recent development of visual feature extraction from a problem-oriented perspective which may provide more insight into the current progress in visual speech decoding. As mentioned in Section 1, there are three particular problems related to visual features. This section is therefore comprised of three subsections below. We describe those efforts that attempt to tackle speaker dependency and pose variation in Sections 3.1 and 3.2, respectively. Section 3.3 reviews the methods aimed to extract useful temporal information from video sequences.

3.1. Speaker dependency

As illustrated in Fig. 1, speakers' mouths may look very different and so do their appearances in images. Such speaker dependency causes the major variation that troubles any attempt to extract useful speech related information from the cropped mouth images as pointed out by Cox et al. [15]. In acoustic speech recognition, techniques such as the vocal-tract normalization [52] and the maximum likelihood linear transformations [28] have been developed to effectively counter the variability in the acoustic signal among different speakers. In the visual domain, however, there has not been any universally accepted approach to tackling such speaker dependency.

In the rest of this section, we review those methods that take into account the visual variability among speakers when extracting visual features. We first describe the efforts that attempt to search for a linear transformation that results in a low-dimensional subspace where speaker dependency is suppressed. We then introduce the articulatory feature based methods. Finally, we describe the latest work that uses the generative latent variable model to explicitly model the inter-speaker variations.

The linear discriminant analysis has been widely used to deal with speaker dependency since it tries to pull the class means away from each other and push data points of the same class together at the same time. Here a class often corresponds to a speech unit (e.g., a viseme). Potamianos et al. [96] applied some image transformation (e.g., the DCT, DWT, or PCA) on ROI images and removed the mean from feature vectors' output by the transformation over each utterance. They then used LDA to further reduce the dimensionality. Later, they extended their method through applying the 'inter-frame' LDA on the concatenation of consecutive feature vectors output by the previous LDA which they referred to as the 'intra-frame' [95]. The extracted feature was named 'HiLDA'. In this way, temporal information was encoded (we will discuss this issue in detail in Section 3.3). Lan et al. [50] adopted a similar strategy. Instead of the intra-frame LDA, they used AAMs to calculate features from images and applied the *z*-score normalization on a per-speaker basis. Note that their method required to acquire data of every test speaker to calculate the means and STDs for the *z*-score normalization before testing. Finally, they performed the inter-frame LDA on the normalized features.

As argued by Yan et al. [117], the LDA projection can be considered as the linearized solution to a graph embedding (GE) problem. Within a graph, data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ are represented by vertices and connected by edges with assigned weights that quantify their similarities. A graph can be described by a matrix $\mathbf{W}$ whose element $W_{i,j}$ records the similarity between the *i*th and *j*th data points. The GE defines $\mathbf{W}$ and $\mathbf{W}^{\mathrm{p}}$ to encode the desired and undesired geometrical relationships of data points, respectively. The linearized GE (LGE) searches for linear projections $\mathbf{w}^*$ that preserve the desired and penalize the undesired geometrical information. Mathematically,

$$\mathbf{w}^* = \underset{\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{L}^{\mathrm{p}}\mathbf{X}^{\mathrm{T}}\mathbf{w}=c}{\arg\min}\ \mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{w} \tag{8}$$

where $\mathbf{L}$ and $\mathbf{L}^{\mathrm{p}}$ are the Laplacian matrices of $\mathbf{W}$ and $\mathbf{W}^{\mathrm{p}}$ and c a constant. The effectiveness of LGE has been demonstrated by its use in face recognition [37,117].

Fu et al. [27,26] adopted the LGE framework for extracting visual speech features. For every data points $\mathbf{x}_i$, they searched for its K neighbors $\left\{\mathbf{x}_{j_k^i}\right\}_{k=1}^{K}$ and weights $\left\{w_{j_k^i}\right\}_{k=1}^{K}$ such that $\mathbf{x}_i$ was best reconstructed by $\sum_k w_{j_k^i}\mathbf{x}_{j_k^i}$ under the constraints that $\mathbf{x}_i$ and $\mathbf{x}_{j_k^i}$ belonged to the same class, $w_{j_k^i} \geq 0$ and $\sum_k w_{j_k^i} = 1$. Let $\mathbf{M}$ be the matrix such that $\mathrm{M}_{i,j_k^i} = w_{j_k^i}$. They defined $\mathbf{L} = (\mathbf{I} - \mathbf{M})^{\mathrm{T}}(\mathbf{I} - \mathbf{M})$. Matrix $\mathbf{L}^{\mathrm{p}}$ was defined in the same

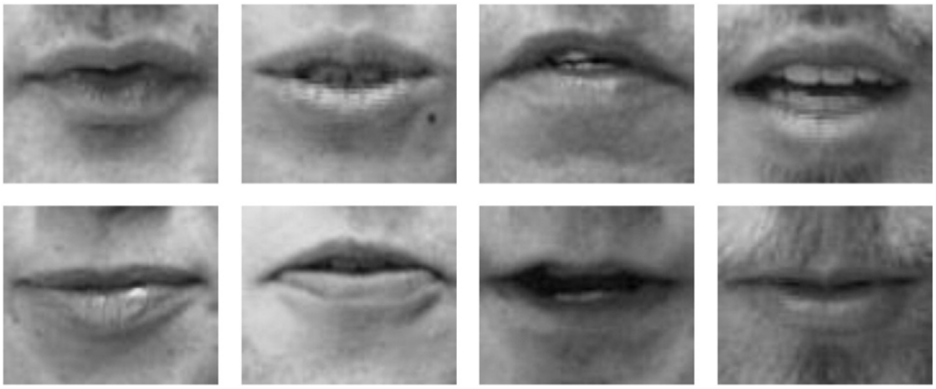


Fig. 1. Sample mouth images from the OuluVS database [120].

way except that the number of neighbors was set as K^{p} and $\mathbf{x}_i$ and each of its neighbors $\mathbf{x}_{j_k^i}$ were from different classes. In this way, the learned subspace preserved the desired and suppressed the undesired local neighboring relationships. Note that $\mathbf{W}$ and $\mathbf{W}^{\mathrm{p}}$ were not explicitly defined. However, Yan et al. [117] proved that $\mathbf{L}$ and $\mathbf{L}^{\mathrm{p}}$ were indeed Laplacian matrices and $\mathbf{W}$ and $\mathbf{W}^{\mathrm{p}}$ existed. For feature extraction, they concatenated DCT coefficients computed from every four consecutive frames and used PCA to reduce the dimensionality. The final visual features were projected from the DCT-PCA features using the learned linear mapping.

The articulatory features (AFs) have been studied as an alternative to the traditional phonemic subword units for modeling speech. Kirchhoff [45] described them as the *abstract classes which characterize the most essential aspects of articulation in a highly quantized, canonical form, leading to a representational level intermediate between the signal and the level of lexical units*. Regarding visual speech, AFs could, for example, be the lip opening, lip rounding or labio-dental articulation as defined by Livescu et al. [56]. Instead of being described by a particular phone, a speech observation is characterized by multiple AFs simultaneously, resulting in a multivariate representation. Such a mechanism allows us to use just a few AFs to model speech that may otherwise requires dozens of phonemic units if they are context independent or hundreds (even thousands) if context dependent.

Since we only need to train a small number of classifiers (or observation probability distributions) for AFs, we may gather a large number of training samples for each AF from training data, which potentially makes the trained classifiers more robust against the variability in the signal among speakers. Moreover, as shown by Papcun et al. [88], AFs themselves are to a certain extent speaker independent. Although it has been an active topic in acoustic speech recognition eLivEtAl07, there are few studies of AFs in the visual domain. Saenko et al. [101,103,102] used AFs for visual-only ASR, sometimes also referred to as automatic lip-reading. In their work, SVMs were constructed to classify AFs from ROI images. The classification scores were converted to probabilities by a fitted sigmoid function. They then fed the probabilities into a multi-stream dynamic Bayesian network (which allowed asynchrony between AFs) for speech recognition.

The visual appearance of a talking mouth can be considered as the output of the combination of multiple sources of information. The desirable visual features need to preserve the relevant speech-related variation while suppressing the others. One way to do that is to use low-dimensional latent variables [75] to *explicitly* (in contrast to the above-mentioned methods) represent these sources of information and model the process that generates the observed images.

Zhou et al. [123] identified two sources of major variations, caused either by the appearance variability among speakers or by speaking utterances, in frontal view mouth images. They considered the former as irrelevant, modeled by the latent speaker variable $\mathbf{h}$ and the latter as relevant, modeled by the latent utterance variable $\mathbf{w}_t$. The process of generating a video sequence was described as:

$$\mathbf{x}_t = \mu + \mathbf{F}\mathbf{h} + \mathbf{G}\mathbf{w}_t + \epsilon_t. \qquad (9)$$

Here $\mathbf{x}_t$ stands for the observed image at time t, μ a global mean image, $\mathbf{F}$ and $\mathbf{W}$ the factor matrices and ϵ_t the normally distributed noise term. They placed $p(\mathrm{w}_t)$ along a low-dimensional curve to preserve the temporal relationships among video frames (see Section 3.3 for details). To extract visual features from a video sequence $\{\mathbf{x}_t\}$, they fitted model $\mathcal{M}$ that was trained for a particular utterance to the sequence through measuring the posterior $p(\mathbf{h}, \{\mathbf{w}_t\}|\{\mathbf{x}_t\}, \mathcal{M})$. The MAP estimates $\{\hat{w}_t\}$ were directly used as visual features.

Comparative studies among some features described above were conducted in their experiments. They were carried out through the task of recognizing ten short daily-use utterances among twenty speakers in a speaker-independent setting. Fig. 2 shows the recognition rates averaged over the speakers on the OuluVS database [120]. Here the dark gray bars display the recognition rates obtained on the manually localized ROI images, while the light gray ones on those normalized automatically. The error bars represent one standard deviation. It can be seen that the quality of ROI images could significantly affect the ASR performance.

Furthermore, they showed that the use of intermediate features such as the LBP [80] and LBP-TOP [121] (marked by 'LBP-TOP' in the figure) instead of raw pixel values (marked by 'Raw') could improve the recognition performance.

Note that the proposed latent variable models were tested on recognizing isolated words or phrases, rather than continuous visual speech. They showed that their methods outperformed the HiLDA features on classifying smaller speech units such as visemes, indicating the possibility of extending the system to do continuous speech.

3.2. Pose variation

It is impractical to assume that speakers would face the video camera all the time. Therefore, the talking face could be filmed not only from the frontal view (FV) but from other angles resulting in various head poses in images. Since the camera view can significantly affect the appearance of a talking mouth, pose variation challenges any system that tries to use visual speech information. Most of the methods concerning pose variation

- either extracted some pose-dependent features (PDFs) from non-frontal view (NFV) images and directly used them for speech recognition
- or transformed the PDFs into some pose-independent features (PIFs) before classification.

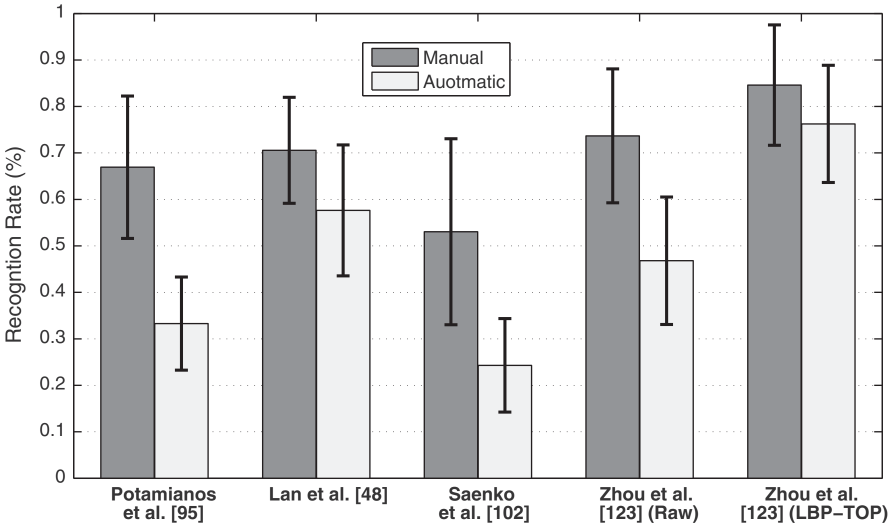


Fig. 2. Comparative results for various visual features concerning speaker dependency.

For those extracting PDFs, the advantage is that there would be no information loss or added noise that may occur during the feature transformation. However, the disadvantage is that for every camera view, we will have to train a particular system using the corresponding PDFs. These methods may suffer from the lack of representative training data for a particular view. Hence, we may have to collect extra NFV data for training. On the contrary, the latter methods attempt to transform PDFs into a common PIF space such that they are comparable. Therefore, only one system needs to be trained based on the available training data. However, they often suffer some substantial performance drop due to the information loss or added noise caused by the feature transformation.

Here we first describe those methods that extracted visual features directly from NFV images for speech recognition. Yoshinaga et al. [118] used the side-view videos recorded by a small camera installed in a headset, near the microphone. Optical flow was computed for each frame and the horizontal and vertical variances of the flow vector components as PDFs.

In [58], Lucey and Potamianos used the cascade method described in [95] to extract PDFs from profile-view (PV) speech videos.

In [60], Lucey et al. divided PV ROI images into some overlapping patches and from each path, calculated a set of PDFs using the same feature extraction method. They reported that features extracted from individual patches did not outperform those from the whole image. The fusion of them slightly improved the recognition performance.

Kumar et al. [47] defined some geometric PDFs for speech recognition. To extract features from PV images, they first segmented foreground pixels using color thresholding. They then calculated the facial contour and located the nose tip and centers of the lip and chin as feature points. They measured four PDFs based on the points. Same geometrical features were measured from FV images. Experiments were conducted to recognize isolated words in a speaker-dependent setting and the PV features were reported to outperform the FV features. Saitoh and Konishi [104] adopted a similar approach. They defined an extra feature point, the lip corner and extracted eight geometrical PDFs. The features were used to recognize vowels and words in a speaker-dependent setting. Their system was reported to outperform human viewers.

Having described the methods based on PDFs, we now focus on those aiming to design PIFs. Inspired by the pose-invariant face recognition work done by Blanz et al. [7], Lucey et al. [59,62] used linear regression to transform PDFs from an unwanted camera view to the wanted one. Let $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}N]$ and $\mathbf{T} = [[\mathbf{t}_1,1]^{\mathrm{T}},\ldots,[\mathbf{t}_N,\mathbf{1}]^{\mathrm{T}}]$ where $\mathbf{x}_n$ is the nth training sample of the unwanted view and $\mathbf{t}_n$ the synchronized counterpart of the wanted view. Linear transformation $\mathbf{W}$ was learned by

$$\mathbf{W} = \mathbf{T}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I}\right)^{-1}. \tag{10}$$

Given a sample $\mathbf{x}$ of the unwanted view, its transformed vector was computed as $\hat{\mathbf{t}} = \mathbf{W}\mathbf{x}$. In their work, linear transformations were learned to project features between PV and FV. From ROI images, PDFs were extracted using a cascade method similar to one in [95]. They showed that the regression-based strategy significantly improved the recognition performance when training and test datasets were of different poses. However, the performance was still substantially worse than the one obtained when training and test datasets were of the same pose.

Esteller and Thiran [24] also used linear regression to normalize PDFs (extracted from views at 30°, 60° and 90°) to FV for pose-invariant speech recognition. In addition to the way used in Lucey et al. [59], linear regression was performed locally on individual patches of ROI images or partial PDFs. In their work, the former was named the global linear regression (GLR) and the latter the local linear regression (LLR). Features such as raw pixel values, DCT and LDA [95] were tested. They found that the GLR-projected LDA features achieved the best performance. LLR performed better than GLR on raw pixel values, but worse on DCT and LDA features.

Lan et al. [49] computed AAM-based PDFs from various views based on their previous work [50]. The extracted features were appended with their second order derivatives for enhancement. They first trained a lip-reading system for each view and tested it on data recorded from that view. It was found out that for the proposed PDFs, 30° was the optimal view instead of FV. After that, they learned linear transformations (Eq. (10)) to project PDFs to the optimal view. They then trained a system for 30° and tested it on data from other views. The results confirmed that those feature transformations helped to increase system performance when training and test datasets were of different poses.

Pass et al. [89] attempted to select DCT coefficients with the minimum cross-pose variances to form PIFs. Given two synchronized training datasets of different poses, they first applied 2D-DCT to obtain two feature streams F_1 and F_2. They then calculated $F_{1-2} = F_1 - F_2$ and matrix C_{1-2} that contained variances of elements of F_{1-2}.

Finally, they picked up the DCT coefficients corresponding to the smallest values in C_{1-2} as PIFs. Note that since the variances were calculated for a specific pair of wanted/unwanted views, the corresponding transformation was actually view-dependent. In case the selected coefficients were different, features transformed from different NFVs would not be able to share one ASR system built upon FV data.

3.3. Temporal information

The visual speech signal contains not only the spatial information of the visual appearance of a talking mouth, but the important temporal information that characterizes the dynamic process of uttering. Potamianos et al. [95] proposed to concatenate consecutive feature vectors and employ LDA to obtain the final compact features that encoded temporal information. Since then the method has been used in a number of systems [59,60,62,50,49,24]. Despite of its popularity, such a simple linear approach may not be sufficient to capture the dynamic speech information. Other intuitive methods include the use of optical flow to capture the motion information [66] and the use of B-splines to model the temporal trajectories of the extracted individual visual features [115]. The former relies on the accuracy of the computed flow information and the latter on the quality of the extracted features, and therefore, may both be sensitive to noise. Below we review the more sophisticated methods that have been recently developed to tackle the problem.

As illustrated in Fig. 3, a video can be viewed as a 3D volume from which we may obtain the temporal patterns (TPs) which are defined as the images formed by stacking a particular row/column of each frame along the temporal axis. Zhao and Pietikäinen [121] exploited the texture information within TPs to characterize video dynamics. They applied the LBP descriptors both on TPs to capture temporal and on video frames to extract spatial information. The concatenated spatio-temporal LBP histograms, named 'LBP-TOP', were successfully used for dynamic texture classification. Zhao et al. [120] used LBP-TOP features to handle the lip-reading task of classifying a limited number of short utterances. The features were computed from the whole utterance and classified by SVMs to recover utterance identities. Later, the above method was extended in [125] through adding a video normalization phase. Videos of the same utterance across speakers were linearly interpolated to have a pre-defined length before the calculation of LBP-TOP features. They found out that such a simple phase significantly improved the lip-reading performance (more than 20% better than the one reported in [120]).

Ong and Bowden [82] proposed to use the *temporal signatures* (TS) to capture the temporal information. In their work, a static image was represented as a binary feature vector. A TS was defined as an arbitrary set of '1' locations within a TP which is a binary image formed by stacking feature vectors extracted from a fixed number of consecutive video frames in this case. Weak classifiers were formed to detect TSs within an input TP. For utterances to be recognized, strong classifiers were constructed from the weak classifiers within the AdaBoost framework. Due to the huge number of TSs that could be defined within a TP, they proposed a gradient-descent based method to search for suitable TSs for learning strong classifiers.

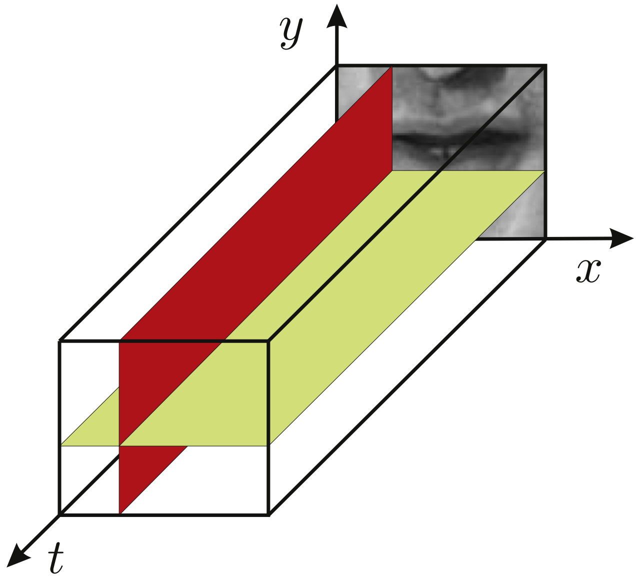


Fig. 3. Temporal patterns defined as the images formed by stacking a particular row (green) or column (red) of each video frame along the temporal axis.

Pachoud et al. [85] extracted visual features directly from 3D image volumes. To do that, they divided a video segment into several 'macro-cuboids'. Each macro-cuboid was then divided into cuboids and the SIFT descriptor adapted for cuboids [21] was used to calculate features from them. They collected a database of model templates (video segments of the utterances to be classified) and matched them with test video segments to recognize utterances.

Zhou et al. [124] adopted the LGE approach (see Eq. (8)). Instead of measuring the actual distances between visual observations which did not take their temporal arrangement into account, they defined a pseudo-distance measure based on the frame alignment results between sequences of the same utterance.

Interestingly, they found out that the training sequences were mapped onto temporal trajectories similar to sine waves. The waves were then Fourier transformed and the peak frequencies recorded for classifying utterances. Their method was used to tackle the same lip-reading problem as in [120] and found to perform well in a speaker-dependent setting.

Inspired by the above finding, the same authors proposed to use a path graph to represent the temporal structure of a video sequence [125]. It was found out that the vertices of a path graph could be embedded onto a low-dimensional curve through graph embedding and more importantly, each dimension of the curve was exactly a sine wave, consistent with the results in [124]. Fig. 4 illustrates the path-graph representation and the embedded curve. For a reference video sequence, they learned a linear map that projected the features extracted from images (e.g., LBP) onto the embedded curve.

To match the reference, a test sequence was mapped using the learned projection and the resulting trajectory was compared with the embedded curve.

Once again, the method was used to solve the above speaker-dependent lip-reading problem.

An obvious drawback of the above model is that it is trained on a single video sequence and therefore, may generalize poorly to sequences of a different speaker. Zhou et al. [123] generalized the above representation. In particular, the representation was relaxed to allow graph vertices to represent data points that were either observed or unobserved. In such a way, the path graph could be used to model multiple video sequences of the same utterance. For feature extraction, they constructed a generative latent variable model (see Eq. (9)) that represented the speech related variations by latent variables $\mathbf{w}_t$. To preserve temporal information, they placed prior distributions $p(\mathbf{w}_t)$ along the embedded curve, as illustrated by Fig. 5, to penalize values of $\mathbf{w}_t$ that contradicted the temporal relationships among image frames.

Pei et al. [91] used AAM to track local feature points on the lip in video frames. Around each point, they extracted a small patch and calculated LBP and HOG features as the texture features. The point's in-image displacement from the previous frame was measured as the shape features. They then defined a *patch trajectory* which is the trajectory of the combined texture and shape features over time for the point. The random forest was constructed to measure similarities between patch trajectories and the multidimensional scaling algorithm performed to learn a low-dimensional manifold within which a patch trajectory was represented by a single data point. Video sequences were projected into sets of feature points in the manifold and their distances measured based on the points.

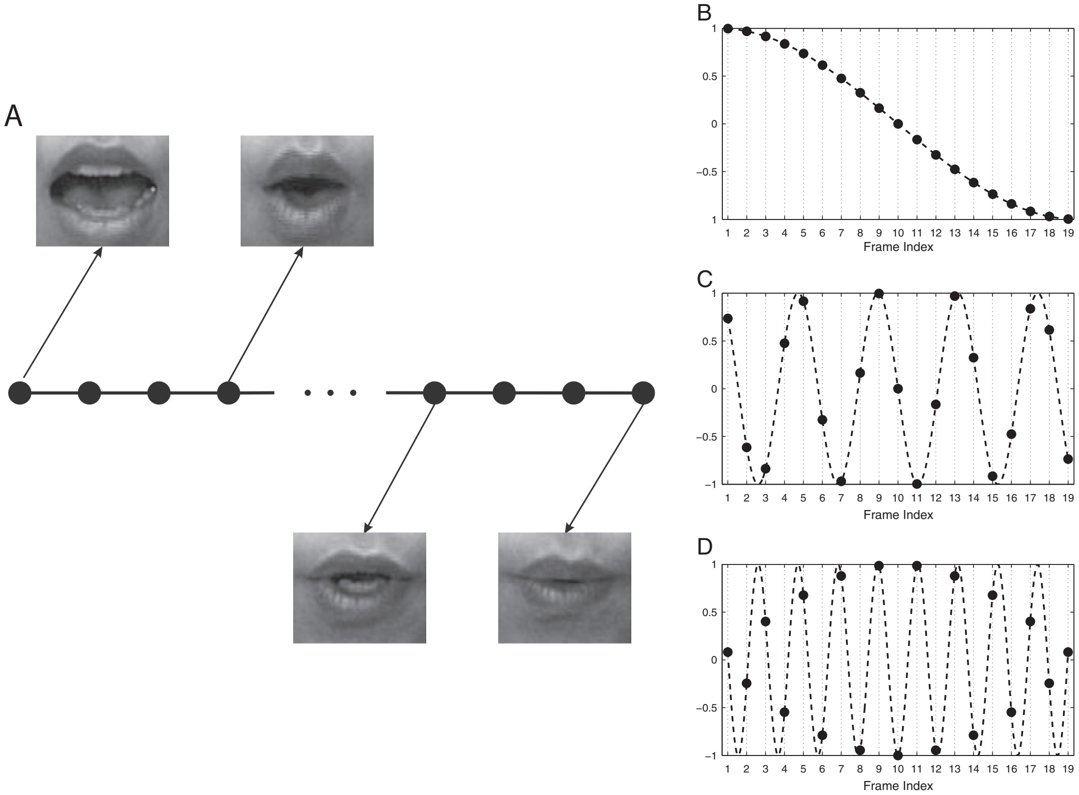


Fig. 4. (A) A path graph representation of a video sequence and (B)–(D) sine waves in the 1st, 9th and 18th dimensions of the curve embedded within a path graph with 19 vertices. All the figures were used in [125].

It can be seen that the problem of temporal information extraction has been tackled by the above methods at two different levels. In [85,120,82], they considered the local pixel-level spatio-temporal structures while in [124,125,91,123] the structure at the frame level was modeled and enforced in the extracted visual features. Note that none of the above sophisticated methods were tested on continuous visual speech. In [120,124,82,91,123], systems were trained and tested for classifying isolated words/phrases, while in [85], the problem of detecting isolated digits in continuous speech was considered. Only Zhou et al. [123] provided results on classifying smaller visual speech units. For other methods, it is unknown whether they are suitable for classifying short video segments, for example, at the viseme level.

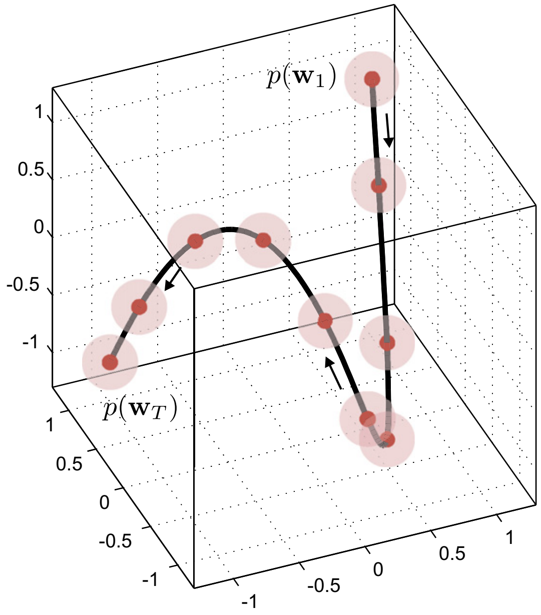


Fig. 5. Prior distributions $p(\mathbf{w}_t)$ constructed along the curve embedded within a path graph. The figure was used in [123].

4. Dynamic audio-visual fusion

Audio-visual speech information fusion is a non-trivial task in speech recognition. As summarized in [95,97], fusion can be performed at either the *feature* or *decision* level. The former calculates a new set of features, out of the ones extracted from the audio and visual streams, which are more discriminative for speech recognition. For such a purpose, Ngiam et al. [79] employed multi-modal deep learning for feature-level fusion recently. However, its major disadvantage is that it cannot cope with speech data with varying qualities, which would be expected for a practical system. In contrast, decision-level fusion combines the decisions made, or in other words, the probability output by the HMMs that are trained for the individual modalities. It allows us to encode the reliability of each modality such that AV-ASR systems can adapt to a new environment efficiently.

Within the decision-level scheme, *dynamic* AV fusion (DAVF) has emerged as the focus of recent studies. One important reason is that it provides a system the capability of constantly adjusting itself to the volatility in the quality of the observed audio or visual streams. For instance, when there is a burst of noise in the background, the system should quickly adapt itself so as to temporarily reply more on the visual stream. Similarly, when there is a sudden illumination change, it should weigh more on the acoustic stream. In this section, we review the techniques recently developed for DAVF.

Before going into any detail, we first introduce the most widely used model for combining the likelihoods output by HMMs. In the model, the extracted audio and visual feature vectors at time t, $\mathbf{x}_t^A$ and $\mathbf{x}_t^V$, are assumed to be conditionally independent and the emission probabilities combined as:

$$p\left(\mathbf{x}_t^{\mathrm{A}}, \mathbf{x}_t^{\mathrm{V}} | \mathbf{q}_t\right) = p\left(\mathbf{x}_t^{\mathrm{A}} | q_t^{\mathrm{A}}\right)^{\lambda_t^{\mathrm{A}}} p\left(\mathbf{x}_t^{\mathrm{V}} | q_t^{\mathrm{V}}\right)^{\lambda_t^{\mathrm{V}}}. \tag{11}$$

Here $\mathbf{q} = [q_t^{\mathrm{A}}, q_t^{\mathrm{V}}]^{\mathrm{T}}$ stands for the state numbers and λ_t^{A} and λ_t^{V} the non-negative stream weights that control the contribution of each modality to the (unnormalized) joint likelihood. Note that when asynchrony is permitted between the audio and video HMM states, q_t^{A} and q_t^{V} may be different. Due to the dynamic nature of DAVF, the weights need to be set in an online manner and their values directly related to the modality reliabilities measured from the current observed data.

To simplify the problem, the sum of the weights is often fixed, e.g., $\lambda_t^{\mathrm{A}} + \lambda_t^{\mathrm{V}} = 1$ such that only one weight is required to be estimated. Next, we will first describe how the modality reliability is quantified and then the way it is converted into the stream weight.

The signal-to-noise ratio (SNR) is perhaps the most intuitive and straightforward reliability measure [2,72,16,112,109,23]. For the audio signal, SNR can be calculated as the ratio between the power of the speech signal and that of the noise. The latter is often measured from some silence interval. Estellers et al. [23] found out that SNR obtained during the silence intervals within utterances could significantly degrade the fusion performance if treated the same as those measured at the speech time. Therefore, they constructed a detector based on the trained HMM classifier [110] to detect the non-speech moment and learned different functions for converting the corresponding speech/non-speech SNRs into stream weights. Shao and Barker [109] considered the way of measuring SNR using the noise power calculated from silence intervals as unreliable since there could be voice from other speakers at the background, which should also be considered as noise. They argued that the true SNR be measured using both the acoustic and visual information and proposed to use the three-layer *multi-layer perceptrons* (MLPs) to measure it. They first trained two HMM classifiers for the audio and visual streams, respectively. At time t, to estimate SNR, the likelihoods (of individual Gaussian mixtures) output by all the audio and visual states were passed to the bottom layer of the MLPs. SNR was then output by the top layer.

Besides SNR, the modality reliability can also be quantified by the dispersion [2,100,94,38,95,63,23] and entropy measure [94,38,23]. The former is defined on the N-best state likelihoods output by the trained HMMs and mathematically, can be written as:

$$\mathcal{D}_t = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{m=n+1}^{N} \log \frac{p\left(\mathbf{x}_t | q_{n,t}\right)}{p\left(\mathbf{x}_t | q_{m,t}\right)} \tag{12}$$

where $\mathbf{x}_t$ is the observed feature vector and $\mathrm{q}_{n,t}$ the number of the state with the nth best likelihood. The latter is defined on the posteriors of all the HMM states q and can be expressed as:

$$\mathcal{H}_t = -\sum_q p(q|\mathbf{x}_t) \log p(q|\mathbf{x}_t). \tag{13}$$

Other reliability measures include the voicing index [6,30] and the N-best log-likelihood difference [94,95,63]. Recently, Estellers et al. [23] proposed to accumulate the transition probability between the previous and current most likely (ML) state to indicate the signal reliability. Mathematically, the measure $\mathcal{C}_t$ was defined as:

$$\mathcal{C}_t = \mathcal{C}_{t-1} + p\left(q_t^{\mathrm{ML}} | q_{t-1}^{\mathrm{ML}}\right). \tag{14}$$

Here q_t^{ML} was the number of the state with the best likelihood.

The motivation behind the measure was that if the observed signal was somehow corrupted, the transition between the ML states would be likely unnatural and therefore, $\mathcal{C}_t$ would be lowered by the corresponding small transition probability.

Given the reliability estimate(s), there have been various functions designed to map the estimate(s) to one stream weight. Meiel et al. [72] proposed a piece-wise linear function. However, the weight in their work was fixed globally. Glotin et al. [30] adopted the same function to convert the estimated voicing index into λ_t^{A}. Garg et al. [29] used a sigmoid function to evaluate λ_t^{A}, i.e., $\lambda_t^{\mathrm{A}} = 1/(1 + \exp(a + \mathbf{w}^{\mathrm{T}}\mathbf{y}))$ where $\mathbf{y}$ was the vector formed by multiple AV reliability estimates.

Marcheret et al. [63] quantized λ_t^{A} such that it had a fixed number of values that were evenly distributed between 0 and 1. They set λ_t^{A} either as the expectation or as the MAP estimate given the posterior distribution $p(\lambda^j|\mathbf{y})$ where λ^j was the jth quantized weight value. They modeled each distribution using a full-covariance GMM and learned the model parameters from the instances that maximized the word error rate (WER) on the evaluation dataset.

Gurbuz et al. [35] used a lookup table to convert the measured SNR. The table had two dimensions standing for the ratio and noise type. To learn its entry values, they first added various types and levels of noise to the audio evaluation data. They then filled in the table with the value that minimized WER on the modified data based on the HMMs learned from the original clean data. Shao and Barker [109] adopted the same approach. However, the table's dimensions represented SNR and a parameter used for normalizing the video stream likelihoods. Estellers et al. [23] defined the function as $\lambda_t^{\mathrm{A}} = a_1 e^{b_1 y} + a_2 e^{b_2 y}$ where y was the reliability measure (they tried various measures in their work) and α_1, α_2, b_1, and b_2 were the parameters. They learned the global λ_t^{A} values for various SNR levels and tuned the function parameters such that difference between the global values and converted weights were minimized.

Note that the methods described so far all follow the same strategy that the stream reliability is first estimated from the observed data and then converted into λ_t^{A}. However, there are other methods that did not use the strategy. In the rest of this section, we will review these methods. Instead of dynamically choosing the stream weights in Eq. (11), Kolossa et al. [46] focused on selecting the robust features at the running time. For visual features, they trained two Gaussian distributions P_m and P_n on the features extracted from the correctly and incorrectly detected mouth images, respectively. At time t, if $p_m(\mathbf{x}_t^{\mathrm{V}}) < p_n(\mathbf{x}_t^{thrmV})$, $\mathbf{x}_t^{\mathrm{V}}$ would not be considered by the system, i.e., $\lambda_t^{\mathrm{V}} = 0$. The audio features were compared with a background noise estimate. An audio feature was deemed unreliable and therefore, discarded if the value of the corresponding feature extracted from the background noise exceeded 90% of its value. In their experiments, λ_t^{A} was set globally according to different levels of noise in audio data.

Stewart et al. [111] proposed the maximum weighted stream posterior model. They defined the posterior of the synchronized HMM state q_t given a weight w_t as:

$$p\left(q_t | w_t, \mathbf{x_t}^{\mathrm{A}}, \mathbf{x}_t^{\mathrm{V}}\right) = p \frac{\left(\mathbf{x}_t^{\mathrm{A}} | q_t\right)^{w_t} p\left(\mathbf{x}_t^{\mathrm{V}} | q_t\right)^{1-w_t}}{\sum_{q_t'} p\left(\mathbf{x}_t^{\mathrm{A}} | q_t'\right)^{w_t} p\left(\mathbf{x}_t^{\mathrm{V}} | q_t'\right)^{1-w_t}} \tag{15}$$

and the optimal posterior as

$$p\left(q_t | \mathbf{x}_t^{\mathrm{A}}, \mathbf{x}_t^{\mathrm{V}}\right) = \max_{w_t} p\left(q_t | w_t, \mathbf{x}_t^{\mathrm{A}}, \mathbf{x}_t^{\mathrm{V}}\right). \tag{16}$$

In their work, w_t was quantized to have a finite set of values within [0,1].

Note that none of the fusion methods we have described so far explicitly modeled the actual noise, or in other words, the uncertainties in the observed stream features. As an exception, Papandreou et al. [86] considered the observed *noisy* features $\mathbf{x}_t^M$ where $M = \{A,V\}$, as the sum of the underlying *clean* features $\mathbf{z}_t^M$ and some Gaussian noise

ϵ_t^M with mean $\mu_{n_t}^M$ and covariance $\Sigma_{\epsilon_t}^M$. The likelihood $p(\mathbf{x}_t^M|q_t^M)$ could then be expressed as:

$$p\left(\mathbf{x}_t^M|q_t^M\right) = \int p\left(\mathbf{x}_t^M|\mathbf{z}_t^M\right)p\left(\mathbf{z}_t^M|q_t^M\right)\mathrm{d}\mathbf{z}_t^M. \tag{17}$$

Given that $p(\mathbf{z}_t^M|q_t^M)$ was modeled by a GMM, they obtained a close-form expression for $p(\mathbf{x}_t^M|q_t^M)$ which was also a GMM.

They introduced uncertainty compensation fusion model (UCFM) as:

$$p\left(q_t|\mathbf{x}_t^{\mathrm{A}},\mathbf{x}_t^{\mathrm{V}}\right) = p(q_t)p\left(\mathbf{x}_t^{\mathrm{A}}|q_t\right)p\left(\mathbf{x}_t^{\mathrm{V}}|q_t\right). \tag{18}$$

They also introduced a hybrid model that gave separate weights for the two modalities, i.e.,

$$p\left(q_t|\mathbf{x}_t^{\mathrm{A}},\mathbf{x}_t^{\mathrm{V}}\right) = p(q_t)p\left(\mathbf{x}_t^{\mathrm{A}}|q_t\right)^{\lambda_t^{\mathrm{A}}}p\left(\mathbf{x}_t^{\mathrm{V}}|q_t\right)^{\lambda_t^{\mathrm{V}}}. \tag{19}$$

AAM was used to extract visual features. They set $\mu_{\epsilon_t}^{\mathrm{V}} = 0$ and $\Sigma_{\epsilon_t}^{\mathrm{V}}$ as the covariance matrix obtained as the by-product of the least-squares AAM fitting process [87]. On the audio side, they adopted the method described in [19] to measure the noise mean and covariance. In the experiments, the proposed models were compared with the stream weighting fusion model (SWFM) in Eq. (11). It turned out that the hybrid model achieved the best performance and SWFM significantly outperformed UCFM especially in the low SNR area. They also found out that the noise adapted likelihood in Eq. (17) could substantially boost the audio-only ASR performance.

5. Audio-visual speech databases

One major obstacle to the current research on visual speech decoding is the lack of suitable databases. In contrast to the richness of audio speech corpora, only few databases are publicly available for visual-only or audio-visual ASR [97]. Most of them include a limited number of speakers and a small vocabulary (e.g., digits, phrases or short sentences), and therefore, are not suitable for training LVCSR systems. Those eligible for LVCSR are often publicly unavailable, making it difficult for researchers to develop a large-scale visual-based ASR system. Furthermore, very few databases contain the multi-view visual speech data which are important for constructing systems that take head-pose variation into account. In this section, we provide details of the English databases that have been recently or can be potentially used for research in visual speech decoding. Fig. 6 shows some sample images from those databases.

5.1. AVICAR

Lee et al. [51] introduced the *Audio-Visual speech In a Car* (AVICAR) database that was recorded in a moving car. They employed four cameras in a lateral array on the dashboard for video recording, resulting in four synchronized video streams with different views. Due to the limited space in the car, the angles between the views relative to the speaker were modest and the actual degrees unknown. There are 100 speakers (50 male and 50 female) involved in the recording and data of 86 of them are available for downloading. There were 5 noise conditions set up during recording. Under each condition, each speaker was asked to first speak isolated digits and letters twice. It was followed by 20 phone numbers with 10 digits each and 20 sentences randomly chosen out of 450 TIMIT sentences [127]. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio sampled at 16 kHz, 16-bit resolution.

5.2. AVLetters

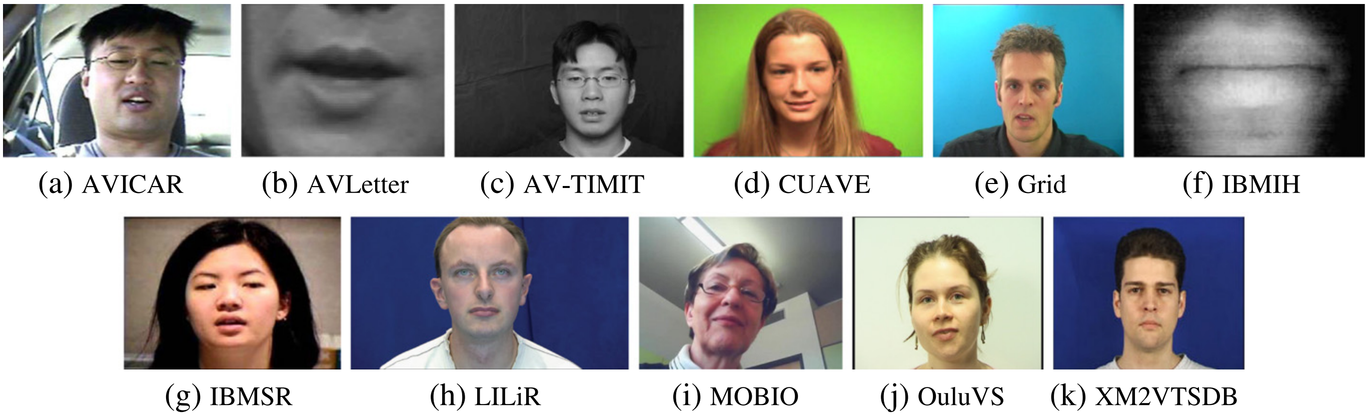
The AVLetters database [68] consists of 10 speakers (5 male and 5 female) uttering isolated letters A-Z. Each letter was repeated three times by each speaker during recording. Video was recorded at 25 fps with a resolution of 376 × 288 pixels and audio at 22.5 kHz with a 16-bit resolution. Image data were processed such that a 80 × 60 full-face region was cropped based on the manually located center of the mouth in the middle frame of each utterance. Moreover, they temporally segmented each utterance such that it began and ended with the speaker's mouth in the closed position.

5.3. AVLetter2

Cox et al. [15] collected a higher definition version of the AVLetter database, named 'AVLetter2'. The corpus includes 5 speakers uttering 26 isolated letters seven times. Video was recorded at 50 fps with a resolution of 1920 × 1080 pixels and audio as 16-bit 48 kHz mono.

5.4. AV-TIMIT

Hazen et al. [36] produced the AV-TIMIT database for their studies of speaker-independent AV-ASR. The corpus contains 4 h of AV data collected from 233 speakers (117 male and 106 female). The spoken utterances were chosen from the phonetically balanced TIMIT sentences [127]. Each speaker was asked to read 20 sentences and each sentence read by at least 9 different speakers. They chose one sentence that was uttered by all the speakers. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio sampled at 16 kHz. Unfortunately, the corpus was not made accessible for public research.



(a) AVICAR (b) AVLetter (c) AV-TIMIT (d) CUAVE (e) Grid (f) IBMIH

(g) IBMSR (h) LILiR (i) MOBIO (j) OuluVS (k) XM2VTSDB

Fig. 6. Sample images from various audio-visual databases.

5.5. CUAVE

Patterson et al. [90] recorded the Clemson University Audio-Visual Experiments (CUAVE) database that included speaker movement and simultaneous speech from multiple speakers. It consists of two major sections. In the first section, 36 speakers (17 male and 19 female) were involved in the recording. Each speaker was asked to utter 50 isolated digits while standing naturally and another 30 isolated digits while moving side-to-side, back-and-forth, or tilting the head. After that, the speaker was framed from both profile views while uttering 20 isolated digits. The individual then uttered 60 connected digits while facing the camera again. The second section of the database includes 20 pairs of speakers. For each pair, one speaker was asked to utter a connected-digit sequence, followed by the other speaker and vice versa a second time. For the third time, both speakers uttered their own digit sequences simultaneously. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio at 16-bit, mono rate of 16 kHz. The data was fully labeled manually at the millisecond level.

5.6. Grid

The Grid AV corpus was collected by Cooke et al. [11]. There are 34 speakers (18 male and 16 female) involved in its recording. Due to some technical oversight, video data for speaker 21 are not available according to the information provided on the downloading webpage. The utterances are sentences with the form <*verb*> + <*color*> + <*preposition*> + <*digit*> + <*letter*> + <*adverb*>. There are multiple words that could be chosen at each position, resulting 1000 sentences per speaker in total. Video was recorded at 25 fps with a resolution of 720 × 576 pixels (a lower quality version (360 × 288) also available) and audio down-sampled to 25 kHz with the peak SNR varying across speakers from 44 to 58 dB. During recording, subjects were asked to speak sufficiently quickly to fit each sentence into a 3-second time window.

5.7. IBMIH

The *IBM Infrared Headset* (IBMIH) Database [42] contains the infrared videos instead of those recorded under idea visual conditions. The speaker was asked to wear a headset containing an infrared video camera in front of the mouth and a microphone. Only the mouth-chin area was framed in the videos. The corpus consists of 79 speakers uttering continuous digit strings and another 113 speakers reading ViaVoice [95] dictation scripts (sentences). There are a total of 4011 utterances of digit strings and 12186 utterances of sentences. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio at 22 kHz.

5.8. IBMSR

The *IBM Smart-Room* (IBMSR) Database [60] was collected as part of the European project, CHIL [60]. The corpus consists of 38 speakers uttering continuous digit strings. There were two microphones and three cameras used for AV data collection. The cameras were set to frame the speaker from the frontal and both profile views. Video was recorded at 30 fps with a resolution of 368 × 240 pixels and audio at 22 kHz. There are in total 1661 utterances included in the corpus.

5.9. LILiR

The *Language Independent Lip-Reading* (LILiR) database [1] collected at the University of Surrey consists of 20 speakers uttering 200 sentences from the Resource Management Corpus [99]. The speaker was framed by two HD cameras from the front and profile views and by three SD cameras placed at 30°, 45° and 60°. It is unknown about the video and audio quality.

5.10. MOBIO

The MOBIO database [69] was designed for evaluating face and speaker authentication algorithms on mobile phones. Videos were recorded from a mobile phone held by speakers. Consequently, the microphone and video camera were no longer fixed and were used in an interactive and uncontrolled manner. There are in total 152 speakers each of whom had multiple sessions of video recording. They were asked short-response questions, free-speech questions and to read predefined texts. Video was recorded at 16 fps with a resolution of 640 × 480 pixels.

5.11. OuluVS

Zhao et al. [120] recorded the OuluVS database for visual-only ASR. It consists of 10 daily-use English phrases uttered by 20 speakers (17 male and 3 female). Each utterance was repeated by a speaker up to nine times. Video was recorded at 25 fps with a resolution of 720 × 576 pixels.

5.12. XM2VTSDB

The XM2VTSDB database [73] was collected at the University of Surrey for personal identification. There were 295 subjects involved and the recording consisted of four sessions. In each section, each subject was asked to speak two continuous digit strings and one phonetically balanced sentence. The utterances remained the same in all the four sections.

Table 2 summarizes the AV databases used in the recent work on visual speech decoding. It lists the speaker number, utterance type, head-pose information, recent work conducted on the databases and their accessibility for public research. Table 3 summarizes the visual-only ASR performance recently obtained on those publicly available databases included in Table 2.

6. Discussions

In the above sections, we have reviewed the recent work on visual speech decoding. It can be seen that the research is still at the early stage and we have not yet found the satisfactory answers to the critical questions raised in Section 1. In this section, we discuss the remaining challenges and provide our vision for future development.

6.1. Visual feature extraction

As already mentioned, there are multiple sources of information included in visual speech data. The aim of feature extraction is to represent the speech-related information compactly meanwhile suppressing the other irrelevant information.

In Section 3.1, we have reviewed the effort to tackle speaker dependency. LDA or more generally LGE based methods [96,95,27,26,50] have been widely used to extract visual features. The speech-related information was enforced by the desired geometric relationships defined on data points and the irrelevant information suppressed through penalizing the undesirable relationships. For these methods the underlying assumption is that there exists a low-dimensional linear subspace that can result in satisfactory features. However, it may not be held in practice due to the complexity of visual speech data. One possible extension is to use the 'kernel trick' [107] that first projects visual features into a higher-dimensional feature space and then searches for a suitable linear subspace. Such a technique has been widely used to learn a non-linear solution to feature extraction [108,74,117].

The articulatory features provide an alternative solution to speaker dependency. In [103,102], the problem of feature extraction was transformed into a classification problem where visual observations are categorized into various AF classes. Scores output by the classifiers

Table 2
Summary of the AV speech databases used in the recent work on visual speech decoding. AV: Audio-Visual ASR, VO: Visual-Only ASR, I: Isolated, C: Continuous, F: Frontal view, P: tbfProfile view.

Database	#Sub.	Utt.	Head Pose	Recent work	Accessibility
AVICAR [51]	100	I/C-digits, TIMIT sent.	Four near frontal views	Fu et al. (2007) [27] – VO	Yes
				Fu et al. (2008) [26] – VO	
				Navarathna et al. (2011) [76] – AV	
AVLetters [68]	10	I-letters	F	Zhao et al. (2009) [120] – VO	Yes
				Ngiam et al. (2011) [79] – VO	
				Pei et al. (2013) [91] – VO	
AVLetters2 [15]	5	I-letters	F	Cox et al. (2008) [15] – VO	Yes
				Pei et al. (2013) [91] – VO	
AVTIMIT [36]	233	TIMIT sent.	F	Hazen et al. (2004) [36] – AV	No
				Saenko et al. (2005) [101] – VO	
				Saenko et al. (2009) [102] – VO	
CUAVE [90]	36	I/C-digits	F, P	Gowdy et al. (2004) [31] – AV	Yes
				Lucey and Sridharan (2006) [61] – VO	
				Livescu et al. (2007) [56] – AV	
				Lucey et al. (2008) [62] – VO	
				Gurban and Thiran (2009) [34] – AV	
				Papandreou et al. (2009) [86] – AV/VO	
				Ngiam et al. (2011) [79] – AV/VO	
				Pei et al. (2013) [91] – VO	
Grid [11]	34	Sent.	F	Shao and Barker (2008) [109] – AV	Yes
				Kolossa et al. (2009) [46] – AV	
				Lan et al. (2009) [48] – VO	No
IBMIH [42]	79 + 113	C-digits, sent.	F	Marcheret et al. (2007) [63] – AV	
				Huang and Kingsbury (2013) [41] – AV	
IBMSR [60]	38	C-digits	F, P	Lucey and Potamianos (2006) [58] – VO	
				Lucey et al. (2007) [59] – AV	No
				Lucey et al. (2008) [60] – VO	
LILiR [1]	20	sent.	F, P, 30°, 45°, 60°	Lan et al. (2012) [49] – VO	No
OuluVS [120]	20	phrases	F	Zhao et al. (2009) [120] – VO	Yes
				Zhou et al. (2010) [124] – VO	
				Ong and Bowden (2011) [81,82] – VO	
				Zhou et al. (2011) [125] – VO	
				Pei et al. (2013) [91] – VO	
				Zhou et al. (2014) [123] – VO	
XM2VTSDB [73]	295	C-digits	F	Stewart et al. (2013) [111] – AV	Yes

were converted into probabilities and fed into DBNs for speech recognition. Since only a relatively small number of AFs are required to describe a speech observation, it is likely to gather sufficient training data for each AF class to counter speaker dependency. The major challenge for constructing an AF-based ASR system is the difficulty of obtaining ground-truth AF values for training. The labeling work can be done either manually or automatically. The former requires expertise in the visual vocal tract and is time-consuming for a large corpus. The latter first detects phoneme boundaries and labels visual data following some fixed rules. Although it can be done automatically, such labeled AF data could contain more noise, as pointed out by Saenko et al. [102], resulting in poorer ASR performance. A bootstrap strategy may be adopted to find the trade-off between the accuracy and efficiency. In addition to AF labeling, we need to ask ourselves whether the probabilities converted from classification scores are the proper input to DBNs. A probabilistic model [44,84] built upon visual observations and AFs may be more suitable for this purpose.

The generative latent variable model has recently been proposed by Zhou et al. [123] to explicitly model different sources of variations in visual speech data. Within the model, two low-dimensional latent variables are defined to represent the variations of the visual appearances among speakers and those caused by uttering in an image. Due to the generative nature, the model allows us to 'see' the corresponding variations represented by the features. In spite of its superior performance shown in the paper, their method was tested in a relatively easy task of recognizing ten different phrases. The challenge is how to adapt the method to cope with continuous speech.

We have seen two very different strategies to deal with head-pose variation. The first one focuses on one particular camera view and tries to extract pose-dependent features [118,58,47,60,62,104] while the other normalizes features with respect to a canonical view [59,89,24,49]. Due to the pose dependency, the former strategy requires to collect sufficient training data, design features and train classifiers for each view. Consequently, building a system that deals with multiple poses would be inevitably expensive. In comparison, the second strategy is much less costly. For each non-canonical view, we only need to collect sufficient amount of data for learning some pose-normalization functions.

Its major challenge remains in the design of the normalization mechanism. So far, the most widely used model is the simple linear regression model (see Eq. (10)). Experimental results in various studies have shown that ASR performance on the normalized features is significantly worse than the one on the features extracted from the canonical view. It indicates that the speech-related information might not be sufficiently preserved by the normalization governed by the simple model and a more sophisticated model could be needed. In [123], Zhou et al. provided their vision of using generative latent variable models to handle pose variation motivated by the work done by Li et al. [54]. In their proposal, visual data with various poses are assumed to be generated by different processes. Moreover, the speech-related latent variables share the same latent space and those corresponding to the same language unit are tied up to enforce the pose invariance in visual features.

In Section 3.3, we have described methods concerning the extraction of temporal information. It is an important problem since such information is able to significantly boost ASR performance. So far, the most commonly used method is the inter-frame LDA strategy that first concatenates feature vectors calculated from consecutive frames and then apply LDA to reduce the dimensionality [95,59,60,62,50,49,24]. Among the more sophisticated methods, Zhao et al. [120], Ong and Bowden [82] and Pachoud et al. [85] explored the local pixel-level spatial-temporal structures to extract temporal information. Zhou et al. [125,123] introduced the path-graph representation and Pei et al. [91] constructed the random forest manifold to enforce the frame-level temporal structures. The latter achieved the state-of-the-art performance, however, its sensitivity to the AAM tracking error was unknown. Note

Table 3
Summary of the visual-only ASR performance on those publicly available AV speech databases included in Table 2. utt: utterance type, exp: experimental setting including speaker **D**ependent, **M**ulti-speaker and speaker **I**ndependent, sub: number of subject used in **TR**ai**N**ing and **Te**S**T**ing, CV: **C**ross **V**alidation, sample: number of video samples used in training and testing, pose: head poses including **F**rontal, **P**rofile and **M**ultiple near frontal views. § result reported in terms of the word accuracy which is defined as $(H - I) / N$ where H is the number of correctly recognized word instances, I the number of insertion errors and N the total number of word instances to be recognized. † facial landmarks manually labeled to extract ROIs.

Database	Experiments						Acc.(%)
	utt	exp	ref	sub	sample	pose	
AVICAR	C–digits	I	Fu et al. [26]	TRN:21 TST:13	TRN:851 TST:490	M	37.87
AVLetter	I–letters	D	Pei et al. [91]		n/a	F	69.6
		M	Matthews et al. [68]		TRN:520 TST:260		41.9
			Zhao et al. [120]				58.85
			Ngiam et al. [79]				64.4
		I	Zhao et al. [120]	TRN:9 TST:1 CV			43.46
AVLetter2	I–letters	D	Cox et al. [15]		TRN:156 TST:26 CV	F	≈ 85
			Pei et al. [91]		n/a		91.8
		M	Cox et al. [15]		TRN:780 TST:130 CV		≈ 85
		I		TRN:4 TST:1 CV			< 10
CUAVE	I–digits	D	Pei et al. [91]		n/a	F	100
		I	Lucey & Sridharan [61]	TRN:28 TST:8			77.08
			Papandreou et al. [86]	TRN:30 TST:6			≈ 83
			Ngiam et al. [79]	TRN:18 TST:18			68.7
			Lucey et al. [62]	TRN:25 TST:8		F,P	38.8
Grid	sent.	I	Lan et al. [48]	TRN:14 TST:1 CV		F	65§
OuluVS	phrases	D	Zhao et al. [120]		TST:1 CV	F	70.2
			Zhou et al. [124]				90.6†
			Ong & Bowden [82]				86.5
			Zhou et al. [125]				96.5†, 85.1
			Pei et al. [91]		n/a		97.3
		I	Zhao et al. [120]	TRN:19 TST:1 CV			62.4
			Ong & Bowden [82]				65.6
			Zhou et al. [125]				81.3
			Pei et al. [91]				89.7
			Zhou et al. [123]				85.6†, 76.6

that these methods were all developed for the problem of recognizing a limited number of phrases. Therefore, the main challenge is how to incorporate these techniques with other tools for continuous visual-based ASR.

At the end of this subsection, we have to point out that although the three questions raised for visual feature extraction have been tackled separately at the moment, eventually, we would need a solution to all of them. Therefore, a unified framework for feature extraction is required for future ASR systems.

6.2. Dynamic audio-visual speech fusion

Dynamic AV speech fusion is aimed to develop a fusion scheme capable of automatically adjusting the fusion rule according to the measured reliabilities of the observed AV streams. Such a scheme is essential for a practical ASR system that uses both acoustic and visual speech information.

Most of the work we have reviewed uses the adaptive stream weighting model described in Eq. (11). The joint likelihood is decomposed as the product of the weighted likelihoods of the two individual modalities. The weights are directly related to the stream reliabilities and their sum often assumed to be constant to simplify the problem. There have been various methods proposed to measure the reliability [2,72,16,100,6,112,94,30,38,109,23], most of which rely on the GMM-component likelihoods emitted by the single-stream HMMs trained on some noisy-free data of one modality (mostly audio). After obtaining the reliability measure, we need to convert it to the corresponding weight. The conversion was often done by some predefined simple function (e.g., the piece-wise linear function [72,30] or sigmoid function [29]), or by a lookup table [35,109].

The main drawback of the above fusion model, however, is that it lacks the mechanism to explicitly model the stream reliability or equivalently, the stream noise. Although stream weighting allows us to carry out dynamic fusion, choosing optimal values for the weights is not straightforward. The rule that converts a reliability measure to the corresponding weight is often defined heuristically and learned heuristically through optimizing a global measure (e.g., the WER) upon some noisy dataset. Moreover, from a theoretical point of view, the product of the weighted likelihoods no longer has the probabilistic interpretation of the joint likelihood of the observed AV streams.

In [86], Papandreou et al. proposed a dynamic fusion scheme that explicitly modeled the stream noise/uncertainties. With the common assumptions of Gaussian noise and GMM likelihood of the inaccessible clean stream features, they found the likelihood of the observed noisy features also a GMM with the distribution adapted with respect to the estimated noise distribution. Mathematically, the proposed probabilistic model provides us a systematic way of dealing with uncertainties. Despite its theoretical advantage, the main challenge of the method is how to reliably estimate the noise distribution especially for the visual stream. It requires the feature-extraction method to also take uncertainties into account such that the noise distribution can be estimated naturally together with visual features.

6.3. Audio-visual speech databases

The availability of representative databases is utterly important for any vision based application including visual speech decoding. After reviewing the AV speech databases that have been recently used in Section 5, we feel that the amount of available data is far from sufficient and there certainly needs more effort in data collection for future research.

Among the factors that determine the suitability of a database, the number of subjects is perhaps the most important one. A large number could benefit research in both feature extraction and AV speech fusion. As can be seen in Table 2, there are four databases collected with a large number (≥100). Among them, only AVICAR and XM2VTSDB are publicly available with the latter designed for personal identification and therefore including very few utterances per speaker.

For the research concerning pose variation, there are four accessible databases that contain videos recorded from different camera views. The AVICAR dataset includes a large subject number, but a relatively small range of pose variations due to the limited space in the car. The others contain only frontal and profile views. Certainly, a database that consists of a sufficiently large number of both subjects and camera views is necessary for such research.

To test the fusion technology, the common practice is to add artificial noise to audio and evaluate the performance with respect to SNR. However, the actual noise encountered in a real world situation might be very different from the artificial one. There are two databases, AVICAR and CUAVE, that provide realistic noise. The former recorded the natural wind sound heard in a moving car and latter the voice from other speakers in the background, which as pointed out by Shao and Barker [109] could mislead some systems that perform well on the artificial noise.

A comprehensive AV speech database is urgently needed at this stage. It should consist of a large number of subjects and a rich pool of utterances for various ASR tasks, be produced in an indoor controlled environment with multiple cameras set up for capturing multi-pose visual speech data and most importantly, be available for public research. Such a high-quality database is crucial for making breakthroughs in research of visual speech decoding. In addition, a more realistic dataset may be collected for testing practical ASR systems. It could include natural human conversations, spontaneous facial expressions and large pose variations. Its audio is expected to be corrupted by environmental noise and videos by factors such as camera motion, illumination changes or moving background objects. One possible way to collect such a dataset is to exploit videos available on the internet.

7. Conclusion

We have reviewed the recent studies in visual speech decoding. They are categorized and described with respect to the four important research questions in this area. Among them, three questions are directly related to the extraction of visual features, concerning speaker dependency, pose variation and temporal information. The fourth question considers the dynamic change of modality reliabilities when audio and visual features are fused in practice. In addition, we have introduced the recent advances in facial landmark localization which can be used to improve ROI detection, but have been largely ignored by researchers working on visual speech decoding. We have also provided details of audio-visual speech databases. Finally, we have discussed the remaining challenges and provided our insights into future research.

Acknowledgment

This research was supported by the Academy of Finland 250155 and Infotech Oulu.

References

[1] http://www.ee.surrey.ac.uk/Projects/LILiR/index.html.
[2] A. Adjoudani, C. Benôit, On the integration of auditory and visual parameters in an HMM-based ASR, in: D. Stork, M. Hennecke (Eds.), Speech reading by Humans and Machines, Springer, Berlin, Germany, 1996, pp. 461–471.
[3] P. Aleksic, J. Williams, Z. Wu, A. Katsaggelos, Audio-visual speech recognition using MPEG-4 compliant visual features, EURASIP J. Appl. Signal Process. 2002 (1) (2002) 1213–1227.
[4] A. Asthana1, S. Zafeiriou1, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, Proc. IEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2013.
[5] P. Belhumeur, D. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2011, pp. 545–552.
[6] F. Berthommier, H. Glotin, A new SNR-feature mapping for robust multistream speech recognition, Proc. Int. Congr. Phonetic Sci, 1999, pp. 711–715.
[7] V. Blanz, P. Grother, P. Phillips, T. Vetter, Face recognition based on frontal views generated from non-frontal images, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), vol. 2, 2005, pp. 454–461.
[8] C. Bregler, Y. Konig, Eigenlips' for robust speech recognition, Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1994, pp. 669–672.
[9] N. Brooke, Using the visual component in automatic speech recognition, Proc. Int. Conf. Spoken Language Process, 1996, pp. 1656–1659.
[10] H. Çetingül, Y. Yemez, E. Erzin, A. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Trans. Image Process. 15 (10) (2006) 2879–2891.
[11] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2008) 2421–2424.
[12] T. Cootes, G. Edwards, C. Taylor, Active appearance models, Proc. European Conf. Comput. Vis. (ECCV), 1998, pp. 484–498.
[13] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. (2001) 681–685.
[14] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models — their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[15] S. Cox, R. Harvey, Y. Lan, J. Newman, B. Theobald, The challenge of multispeaker lip-reading, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2008, pp. 179–184.
[16] S. Cox, I. Matthews, A. Bangham, Combining noise compensation with visual information in speech recognition, Proc. European Tutorial Workshop Audio-Visual Speech Process, 1997, pp. 53–56.
[17] D. Cristinacce, T. Cootes, Automatic feature localisation with constrained local models, Pattern Recogn. 41 (10) (2008) 3054–3067.
[18] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), vol. 1, 2005, pp. 886–893.
[19] L. Deng, J. Droppo, A. Acero, Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion, IEEE Trans. Speech Audio Process. 13 (3) (2005) 412–421.
[20] L. Ding, A. Martinez, Features versus context: an approach for precise and detailed detection and delineation of faces and facial features, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2022–2038.

[21] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, Joint IEEE Int. Workshop Visual Surveillance, Performance Evaluation of Tracking Surveillance, 2005, pp. 65–72.
[22] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[23] V. Estellers, M. Gurban, J. Thiran, On dynamic stream weighting for audio-visual speech recognition, IEEE Audio, Speech, Lang. Process. 20 (4) (2012) 1145–1157.
[24] V. Estellers, J. Thiran, Multi-pose lipreading and audio-visual speech recognition, EURASIP J. Adv. Signal Process. (51) (2012).
[25] N. Eveno, A. Caplier, P. Coulon, Accurate and quasi-automatic lip tracking, IEEE Trans. Circuits Syst. Video Technol. 14 (5) (2004) 706–715.
[26] Y. Fu, S. Yan, T. Huang, Classification and feature extraction by simplexization, IEEE Trans. Inf. Forensics Secur. 3 (1) (2008) 91–100.
[27] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, T. Huang, Lipreading by locality discriminant graph, Proc. IEEE Int. Conf. Image Process. (ICIP), vol. 3, 2007, pp. 325–328.
[28] M. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Comput. Speech Lang. 12 (2) (1998) 75–98.
[29] A. Garg, G. Potamianos, C. Neti, T. Huang, Frame-dependent multi-stream reliability indicators for audio-visual speech recognition, Proc. Int. Conf. Multimedia Expo (ICME), vol. 3, 2003, pp. 605–608.
[30] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, J. Luettin, Weighting schemes for audio-visual fusion in speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2001, pp. 173–176.
[31] J. Gowdy, A. Subramanya, C. Bartels, J. Bilmes, DBN based multi-stream models for audio-visual speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 1, 2004, pp. 993–996.
[32] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.
[33] L. Gu, T. Kanade, A generative shape regularization model for robust face alignment, European Conf. Comput. Vis. (ECCV), 2008, pp. 413–426.
[34] M. Gurban, J. Thiran, Information theoretic feature extraction for audio-visual speech recognition, IEEE Trans. Signal Process. 57 (12) (2009) 4765–4776.
[35] S. Gurbuz, Z. Tufekci, E. Patterson, J. Gowdy, Multi-stream product model audio-visual integration strategy for robust adaptive speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 2, 2002, pp. 2021–2024.
[36] T. Hazen, K. Saenko, C. La, J. Glass, A segment-based audio-visual speech recognizer: data collection, development, and initial experiments, Proc. Int. Conf. Multimodal, Interfaces, 2004, pp. 235–242.
[37] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[38] M. Heckmann, F. Berthommier, K. Kroschel, Noise adaptive stream weighting in audio-visual speech recognition, EURASIP J. Appl. Signal Process 2002 (1) (2002) 1260–1273.
[39] X. Hong, H. Yao, Y. Wan, R. Chen, A PCA based visual DCT feature extraction method for lip-reading, Proc. Int. Conf. Intelligent Informat. Hiding Multimedia, Signal Process, 2006, pp. 321–326.
[40] G. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07-49, University of Massachusetts, Amherst, 2008.
[41] J. Huang, B. Kingsbury, Audio-visual deep learning for noise robust speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2013, pp. 7596–7599.
[42] J. Huang, G. Potamianos, J. Connell, C. Neti, Audio-visual speech recognition using an infrared headset, Speech Commun. 44 (2004) 83–96.
[43] O. Jesorsky, K. Kirchberg, R. Frischholz, Robust face detection using the Hausdorff distance, Proc. Int. Conf. Audio, Video-Based Biometric Person Authentication (AVBPA), 2001, pp. 90–95.
[44] A. Katsamanis, G. Papandreou, P. Maragos, Face active appearance modeling and speech acoustic information to recover articulation, IEEE Audio, Speech, Lang. Process 17 (3) (2009) 411–422.
[45] K. Kirchhoff, Robust Speech Recognition Using Articulatory Information, (PhD thesis) University of Bielefeld, Germany, 1999.
[46] D. Kolossa, S. Zeiler, A. Vorwerk, R. Orglmeister, Audiovisual speech recognition with missing or unreliable data, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2009, pp. 117–122.
[47] K. Kumar, T. Chen, R. Stern, Profile view lip reading, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 4, 2007, pp. 429–432.
[48] Y. Lan, R. Harvey, B. Theobald, E. Ong, R. Bowden, Comparing visual features for lipreading, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2009, pp. 102–106.
[49] Y. Lan, B. Theobald, R. Harvey, View independent computer lip-reading, Proc. Int. Conf. Multimedia Expo (ICME), 2012, pp. 432–437.
[50] Y. Lan, B. Theobald, R. Harvey, E. Ong, R. Bowden, Improving visual features for lip-reading, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2010, pp. 142–147.
[51] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang, AVICAR: audio-visual speech corpus in a car environment, Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2004, pp. 380–383.
[52] L. Lee, R. Rose, A frequency warping approach to speaker normalization, IEEE Trans. Speech Audio Process. 6 (1) (1998) 49–60.
[53] R. Leinhart, J. Maydt, An extended set of Haar-like features, Proc. IEEE Int. Conf. Image Process. (ICIP), 2002, pp. 900–903.
[54] P. Li, Y. Fu, U. Mohammed, J. Elder, S. Prince, Probabilistic models for inference about identity, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 144–157.
[55] X. Liu, Discriminative face alignment, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1941–1954.
[56] K. Livescu, Ö. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 JHU summer workshop final report, Technical report, Johns Hopkins Univ., Center for Language and Speech Processing, 2007.
[57] D. Lowe, Object recognition from local scale-invariant features, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), vol. 2, 1999, pp. 1150–1157.
[58] P. Lucey, G. Potaminanos, Lipreading using profile versus frontal views, Proc. IEEE Workshop Multimedia, Signal Process, 2006, pp. 24–28.
[59] P. Lucey, G. Potaminanos, S. Sridharan, An unified approach to multi-pose audio-visual ASR, Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2007, pp. 650–653.
[60] P. Lucey, G. Potaminanos, S. Sridharan, Patch-based analysis of visual speech from multiple views, Proc. Int. Conf. Auditory–Visual Speech Process. (AVSP), 2008, pp. 69–74.
[61] P. Lucey, S. Sridharan, Patch-based representation of visual speech, Proc. IHCSNet Workshop Use Vis. HCI (VisHCI), 2006, pp. 79–85.
[62] P. Lucey, S. Sridharan, D. Dean, Continuous pose-invariant lipreading, Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2008, pp. 2679–2682.
[63] E. Marcheret, V. Libal, G. Potamianos, Dynamic stream weight modeling for audio-visual speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 4, 2007, pp. 945–948.
[64] A. Martinez, R. Benavente, The AR face database, CVC Technical Report 24, 1998.
[65] B. Martinez, M. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[66] K. Mase, A. Pentland, Automatic lipreading by optical flow analysis, Syst. Comput. Jpn. 22 (6) (1991) 67–75.
[67] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[68] I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[69] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matějka, J. Černocký, N. Poh, J. Kittler, A. Larcher, C. Lévy, D. Matrouf, J.-F. Bonastre, P. Tresadern, T. Cootes, Bi-modal person recognition on a mobile phone: using mobile phone data, IEEE ICME Workshop Hot Topics Mobile Multimedia, 2012, pp. 635–640.
[70] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (5588) (1976) 746–748.
[71] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schröder, The Semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent, IEEE Trans. Affective Comput. 3 (1) (2012) 5–17.
[72] U. Meier, W. Hürst, P. Duchnowski, Adaptive bimodal sensor fusion for automatic speechreading, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 2, 1996, pp. 833–836.
[73] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Proc. Int. Conf. Audio, Video-Based Biometrics Person Authentication (AVBPA), 1999.
[74] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, Fisher discriminant analysis with kernels, Proc. IEEE Workshop Neural Netw. Signal Process. IX, 1999, pp. 41–48.
[75] K. Murphy, Machine learning: a probabilistic perspective, Adaptive Computation and Machine Learning Series, The MIT Press, Cambridge, MA, 2012.
[76] R. Navarathna, T. Kleinschmidt, D. Dean, S. Sridharan, P. Lucey, Can audio-visual speech recognition outperform acoustically enhanced speech recognition in automotive environment? Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2011, pp. 27–31.
[77] A. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, Dynamic Bayesian networks for audio-visual speech recognition, EURASIP J. Appl. Signal Process. 2002 (1) (2002) 1274–1288.
[78] J. Newman, S. Cox, Language identification using visual features, IEEE Audio, Speech, Lang. Process. 20 (7) (2012) 1936–1947.
[79] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learning, Proc. Int. Conf. Mach. Learning (ICML), 2011, pp. 689–696.
[80] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray scale and rotation invariant texture analysis with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[81] E. Ong, R. Bowden, Learning sequential patterns for lipreading, Proc. British Mach. Vis. Conf. (BMVC), 2011, pp. 1–10.
[82] E. Ong, R. Bowden, Learning temporal signatures for lip reading, Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), 2011, pp. 958–965.
[83] E. Ong, R. Bowden, Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1844–1859.
[84] İ. Özbek, M. Hasegawa-Johnson, M. Demirekler, Estimation of articulatory trajectories based on Gaussian mixture model (GMM) with audio-visual information fusion and dynamic Kalman smoothing, IEEE Audio, Speech, Lang. Process. 19 (5) (2011) 1180–1195.
[85] S. Pachoud, S. Gong, A. Cavallaro, Macro-cuboid based probabilistic matching for lip-reading digits, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2008, pp. 1–8.
[86] G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition, IEEE Audio, Speech, Lang. Process. 17 (3) (2009) 423–435.

[87] G. Papandreou, P. Maragos, Adaptive and constrained algorithms for inverse compositional active appearance model fitting, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2008, pp. 1–8.
[88] J. Papcun, T. Hochberg, F. Thomas, J. Larouche, J. Zacks, S. Levy, Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data, J. Acoust. Soc. Am. 92 (1992) 688–700.
[89] A. Pass, J. Zhang, D. Stewart, An investigation into features for multi-view lipreading, Proc. IEEE Int. Conf. Image Process. (ICIP), 2010, pp. 2417–2420.
[90] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE: a new audio-visual database for multimodal human-computer interface research, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 2, 2002, pp. 2017–2020.
[91] Y. Pei, T. Kim, H. Zha, Unsupervised random forest manifold alignment for lipreading, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2013, pp. 129–136.
[92] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[93] G. Potamianos, H. Graf, E. Cosatto, An image transform approach for hmm based automatic lipreading, Proc. IEEE Int. Conf. Image Process. (ICIP), 1998, pp. 173–177.
[94] G. Potamianos, C. Neti, Stream confidence estimation for audio-visual speech recognition, Proc. Int. Conf. Spoken Language Process, 2000, pp. 746–749.
[95] G. Potamianos, C. Neti, G. Gravier, Recent advances in the automatic recognition of audio-visual speech, Proc. IEEE 91 (9) (2003) 1306–1326.
[96] G. Potamianos, C. Neti, G. Iyengar, A. Senior, A. Verma, A cascade visual front end for speaker independent automatic speechreading, Int. J. Speech Technol. 4 (2001) 193–208.
[97] G. Potamianos, C. Neti, I. Matthews, Audio-visual automatic speech recognition: an overview, in: G. Bailly, E. Vatikiotis-Bateson, P. Perrier (Eds.), Issues in audio-visual speech processing, MIT Press, Cambridge, MA, 2004.
[98] G. Potamianos, P. Scanlon, Exploiting lower face symmetry in appearance-based automatic speechreading, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2005, pp. 79–84.
[99] P. Price, W. Fisher, J. Bernstein, D. Pallett, Resource Management RM2 2.0, Linguistic Data Consortium, Philadelphia, 1993.
[100] A. Rogozan, P. Deléglise, M. Alissali, Adaptive determination of audio and visual weights for automatic speech recognition, Proc. European Tutorial Workshop Audio-Visual Speech Process, 1997, pp. 61–64.
[101] K. Saenko, K. Livescu, J. Glass, T. Darrell, Production domain modeling of pronunciation for visual speech recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 5, 2005, pp. 473–476.
[102] K. Saenko, K. Livescu, J. Glass, T. Darrell, Multistream articulatory feature-based models for visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1700–1707.
[103] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, T. Darrell, Visual speech recognition with loosely synchronized feature streams, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2005, pp. 1424–1431.
[104] R. Saitoh, R. Konishi, Profile lip reading for vowel and word recognition, Proc. IEEE Int. Conf. Pattern Recognition (ICPR), 2010, pp. 1356–1359.
[105] J. Saragih, R. Goecke, Learning AAM fitting through simulation, Pattern Recogn. 42 (11) (2009) 2628–2636.
[106] J. Saragih, S. Lucey, J. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2011) 200–215.
[107] B. Schölkopf, A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2002.
[108] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[109] X. Shao, J. Barker, Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment, Speech Commun. 50 (2008) 337–353.
[110] J. Sohn, N. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3.
[111] D. Stewart, R. Seymour, A. Pass, J. Ming, Robust audio-visual speech recognition under noisy audio-video conditions, IEEE Trans. Cybern. 44 (2) (2013) 175–184.
[112] P. Teissier, J. Robert-Ribes, J. Schwartz, Comparing models for audiovisual fusion in a noisy-vowel recognition task, IEEE Trans. Speech Audio Process. 7 (6) (1999) 629–642.
[113] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, Proc. Int. Conf. Language Resources Evaluation, Workshop, EMOTION, 2010, pp. 65–70.
[114] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), vol. 1, 2001, pp. 511–518.
[115] S. Wang, A. Liew, W. Lau, S. Leung, An automatic lipreading system for spoken digits with limited training data, IEEE Trans. Circuits Syst. Video Technol. 18 (12) (2008) 1760–1765.
[116] X. Xiong, F. De la Torre, Supervised descent method and its application to face alignment, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2013, pp. 532–539.
[117] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[118] T. Yoshinaga, S. Tamura, K. Iwano, S. Furui, Audio-visual speech recognition using lip movement extracted from side-face images, Proc. Int. Conf. Auditory–Visual Speech Process. (AVSP), 2003, pp. 117–120.
[119] S. Young, A review of large-vocabulary continuous-speech, IEEE Signal Process. Mag. 13 (5) (1996) 45–57.
[120] G. Zhao, M. Barnard, M. Pietikäinen, Lipreading with local spatiotemporal descriptors, IEEE Trans. Multimedia 11 (7) (2009) 1254–1265.
[121] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[122] X. Zhao, S. Shan, X. Chai, X. Chen, Cascaded shape space pruning for robust facial landmark detection, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2013, pp. 1033–1040.
[123] Z. Zhou, X. Hong, G. Zhao, M. Pietikäinen, A compact representation of visual speech data using latent variables, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 181–187.
[124] Z. Zhou, G. Zhao, M. Pietikäinen, Lipreading: a graph embedding approach, Proc. Int. Conf. Pattern Recognition (ICPR), 2010, pp. 523–526.
[125] Z. Zhou, G. Zhao, M. Pietikäinen, Towards a practical lipreading system, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2011, pp. 137–144.
[126] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition (CVPR), 2012, pp. 2879–2886.
[127] V. Zue, S. Sene, J. Glass, Speech database development: TIMIT and beyond, Speech Commun. 9 (4) (1990) 351–356.