INVITED PAPER

Audiovisual Fusion: Challenges and New Approaches

This paper reviews recent results in audiovisual fusion and discusses main challenges in the area with a focus on desynchronization of the two modalities and the issue of training and testing where one of the modalities might be absent from testing.

By Aggelos K. Katsaggelos, *Fellow IEEE*, Sara Bahaadini, and Rafael Molina

ABSTRACT | In this paper, we review recent results on audiovisual (AV) fusion. We also discuss some of the challenges and report on approaches to address them. One important issue in AV fusion is how the modalities interact and influence each other. This review will address this question in the context of AV speech processing, and especially speech recognition, where one of the issues is that the modalities both interact but also sometimes appear to desynchronize from each other. An additional issue that sometimes arises is that one of the modalities may be missing at test time, although it is available at training time; for example, it may be possible to collect AV training data while only having access to audio at test time. We will review approaches to address this issue from the area of multiview learning, where the goal is to learn a model or representation for each of the modalities separately while taking advantage of the rich multimodal training data. In addition to multiview learning, we also discuss the recent application of deep learning (DL) toward AV fusion. We finally draw conclusions and offer our assessment of the future in the area of AV fusion.

KEYWORDS | Audiovisual (AV) fusion; deep learning (DL); machine learning; multimodal analysis; multiview learning; stream asynchrony

I. INTRODUCTION

Multimodal integration is the synergistic use of the information provided by multiple modalities to assist in the completion of a task by a system. Multimodal fusion refers to any stage in the integration process where there is an actual combination of different sources of information [1]. Integration and fusion of data is meaningful when the data provide redundant and complementary information. It can reduce overall uncertainty and thus serve to increase the accuracy with which the features are perceived by the system. Redundancy can also serve to increase reliability in the case of error or failure in some sources. Complementary information from multiple modalities allows for perceiving features in the environment that are impossible to perceive using just the information from each individual modality operating separately. More timely information may also be provided by multiple modalities due to either the actual speed of operation of each modality, or the processing parallelism that may be possible to achieve as part of the integration process.

There are numerous application areas calling for the integration and fusion of multimodal data. Example areas are biomedical applications (e.g., critical care monitoring and medical images), transportation systems (e.g., intelligent vehicle and highway systems), and multimedia analysis [e.g., person identification from audiovisual (AV) resources, multimodal interaction with robot and multimodal video retrieval].

AV analysis is a specific case of multimodal analysis in which the input sources are audio and video. The two modalities are correlated and convey complimentary information. For example, face visibility benefits speech perception. There has been significant work on investigating the relationship between articulatory movements and vocal tract shape and speech acoustics (e.g., [2]). It has also been shown that there exists a strong correlation among face motion, vocal tract shape, and speech acoustics (e.g.,

Manuscript received January 19, 2015; revised April 8, 2015 and June 25, 2015; accepted July 5, 2015. Date of publication August 13, 2015; date of current version August 20, 2015. This work was supported in part by the U.S. Department of Energy under Grant DE-NA0002520 and by the Spanish Ministry of Economy and Competitiveness under Project TIN2013-43880-R.

A. K. Katsaggelos and **S. Bahaadini** are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu).
R. Molina is with the Departamento de Ciencias de la Computación e I.A., Universidad de Granada, Granada 18071, Spain.

Digital Object Identifier: 10.1109/JPROC.2015.2459017

[3]). Speech production and perception is bimodal. The bimodal integration of AV information in perceiving speech has been demonstrated by the McGurk effect [4].

The basic unit that describes how speech conveys linguistic information is the phoneme. Similarly, the basic visually distinguishable unit, utilized in the AV speech processing and human perception literature [5], [6], is the viseme. Phonemes capture the manner of articulation, while visemes capture the place of articulation [5], [7]. There is no universal agreement about the exact grouping of phonemes into visemes, although some clusters are well defined.

There is a plethora of applications in which audio and video are fused, such as, speech recognition [8]–[15], speaker recognition [16], [17], biometrics verification [18]–[23], event detection [24], concept detection [25]–[27], human or object tracking [28]–[35], active speaker localization and tracking [7], [36]–[40], music content analysis [41], meeting segmentation [42], emotion recognition [43]–[45], monologue detection [46], video retrieval [47], human–computer interaction [48], [49], story segmentation in news video [50], video shot detection [51], voice activity detection (VAD) [52], and source separation [53]–[55]. Clearly, in some of the applications, use of the facial expressions and even the whole body articulation is made, not just the visual articulators. Similarly, in certain applications, the audio (not just the speech signal) is fused with the video signal.

In this paper, we introduce the main concepts and review recent work on the challenging AV information fusion problem. There are a number of review papers on the topic (i.e., [10], [22], and [56]–[64]) and our intention is to continue our review where these papers left off. We present some of the challenges encountered in fusing these two modalities, some of which are encountered in other fusion problems as well. We discuss and compare different ways of addressing such challenges and offer critical perspectives for the field and future research directions in the area. Some of these challenges we address are the effectiveness of each modality in different environmental conditions, in other words, the adaptivity of the AV system to the quality, reliability, and confidence of each modality. We also address the asynchrony issue between the audio and video streams, including different sensing rates and the natural asynchrony between speech and audio clues. We also review the most recent advances and approaches in the field. In particular, we concentrate on the use of deep and multiview learning for AV information fusion.

The paper is organized as follows. In Section II, we describe the feature extraction step and the categories of fusion. In Section III, we discuss some of the dominant fusion techniques, namely, support vector machines (SVMs), dynamic Bayesian networks (DBNs), hidden Markov models (HMMs), and Kalman filters. In Section IV, we describe some of the challenges in fusing audio and video streams. In Section V, we review the approaches followed in addressing some of the challenges in AV fusion and present two recent approaches toward it, namely deep and multiview learning. We draw conclusions and provide our assessment about the future of the field in Section VI.

II. AUDIOVISUAL PROCESSING

Generally, AV analysis encompasses two main steps. In the first step, appropriate features are extracted from each modality. This step is completely dependent on the type of modalities utilized and also the application. In Section II-A, we present an overview of the AV features extracted from these modalities for different applications. In the second step, the information conveyed by the modalities is integrated. Different fusion approaches with their advantages and disadvantages are discussed in Section II-B.

A. Feature Extraction

Representing modalities, i.e., audio and video, in an appropriate and efficient feature space is an important step before their fusion. For audio sources, there are some well-known representative features that have been used widely in the speech and audio research community, such as spectrum-based features, like Mel-frequency cepstral coefficients (MFCCs) [65], [66] and linear predictive coding (LPC) [67], phoneme posterior features [68] and prosodic features [44]. On the other hand, finding appropriate visual features from video sources is challenging [56]. In most AV applications, visual features are extracted from the informative parts of the body such as mouth and eye regions, but, in general, they are application dependent. The approaches in extracting specific information also vary. For example, in AV speech recognition, while MFCC features are typically utilized to represent speech [69], a number of approaches have been considered for extracting visual features that can be categorized into four groups, those of image-based, motion-based, geometry-based, and model-based features [70]. A generic representation of an AV feature extraction system is depicted in Fig. 1. In most cases, a dimensionality reduction step is considered after the visual features are extracted. To capture the temporal dynamics in both the audio and video streams, first- and second-order derivatives (implemented through differences) are taken from new features. Since typically the rates of the audio and video streams differ, an interpolation step is required to represent them at the same rate.

While, in most cases, the information about the modalities is combined only after feature extraction, it is of interest to consider this information combination during feature extraction, as was done, for example, in [71]. We will discuss this more in Section V-C.

B. Fusion Approaches

Fusion can be performed at different levels. Fusion at the feature level is done before the modeling process by integrating or combining features from all modalities;

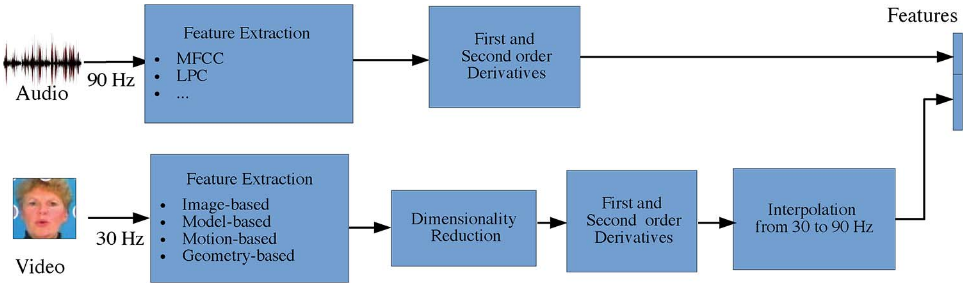


Fig. 1. ***Generic representation of an AV feature extraction system.***

therefore, it is referred to as *early integration* [14]. On the other hand, at the decision level, modeling of each modality is performed separately, and then the outputs or decisions of the models are integrated to produce the final decision [72], therefore it is referred to as *late integration*. Moreover, there is another approach which is in the middle of early and late integrations, and it is referred to as *intermediate integration* (in some literature, it is also considered as early integration). It is also possible to combine two of these fusion approaches by performing fusion at both levels, referred to as *hybrid approach* [73]. In the following, these approaches are described in more detail, focusing on their advantages and disadvantages.

1) Early Integration: An illustration of the early integration approach is shown in Fig. 2(a). As can be seen, appropriate features are first extracted from the two modalities. The extracted features are then combined into one feature set in a process referred to as feature integration. For example, stacking the input feature vectors into a single vector is one of the simplest forms of feature integration. The integrated feature vector will be input to a modeling process, which will produce the final decision or output.

In early integration, the correlation between modalities can be found at the feature level, and only one modeling process is needed, which will result in lower cost and complexity compared to the other fusion techniques which need more modeling process units [57], [72]. However, to be in the same feature space type, the feature vectors need to be converted and probably scaled. Another issue is the size of the integrated feature vector that may result in working in a high-dimensional feature space. It can make the modeling process harder and reduce the scalability of the system. Some techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) can be used to tackle this problem [57]. Additionally, there can be some sort of asynchrony between different modalities because of their different sensing rates and processing times. The feature vectors that are combined together should be from the same time, therefore some considerations should be taken to address this issue [57]. It may be worthwhile to note that while the feature integration is the most common way of early integration, sometimes, one modality can be used to do a particular initialization or preparation, and the rest of the task is performed exploiting just the other modality. For example, Barnard *et al.* [40], for the application of visual tracking of multiple human speakers, use the audio source for the initialization to constrain the search space of the visual face detector.

2) Intermediate Integration: Intermediate integration techniques are very similar to the early integration ones [57]. With these approaches, audio and video features are provided jointly to one modeling process unit. The main difference is that the exploited modeling process unit is especially designed for handling several modalities. It tries modeling each modality separately while considering the interaction between them. Compared to early integration which does not differentiate between features from different modalities, intermediate approaches consider the difference between them. This enables these approaches to handle some degree of asynchrony between modalities and also consider weights for them in different situations. The main difficulty with intermediate integration is the

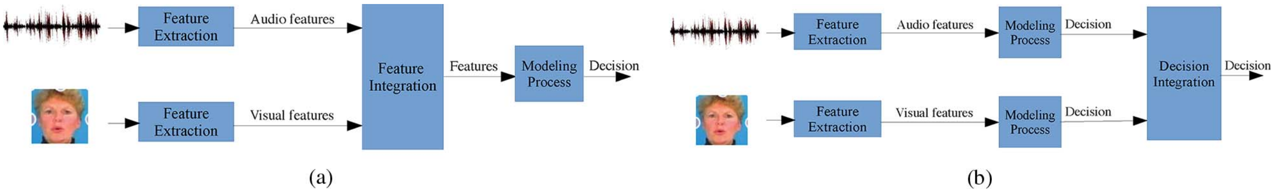


Fig. 2. ***(a) Early integration. (b) Late integration.***

limitation in selecting the modeling techniques because they should be designed specifically for the intermediate integration process [57].

3) Late Integration: The overall process of the late integration approach is illustrated in Fig. 2(b). With this approach, for each modality, a separate modeling process is exploited which receives the features of one modality as input and produces an output decision. These are then integrated to form the final result by the decision integration unit. The most straightforward techniques used in this step are weighting, summation, and voting [57]. More advanced machine learning algorithms such as Adaboost [74] can also be used as was done in [52].

In late integration, the outputs of the modeling processes have the same representation, and combining them is easier than combining feature vectors, as is done in the early integration. Additionally, handling asynchrony is easier at the decision level, and also the system is scalable with the number of modalities compared to early integration techniques. Another advantage of this approach is that for each modality a specific technique appropriate for that modality can be used. For example, in an AV speech recognition task, SVMs represent the preferable modeling process for visual features, while HMMs are used for speech signals [57].

The main disadvantage of late integration is that it is not possible to benefit from the correlation of modalities at the feature level. Moreover, due to the need of separate modeling for each modality, late integration is more challenging compared to early integration.

As discussed above, each type of integration has its own pros and cons. Some studies have suggested combining these approaches to benefit from the advantages of both [57]. Such an approach is typically referred to as *hybrid* integration. With this, both early (possibly intermediate) and late integration are employed, and then the decisions from both systems are combined using a decision integration unit to produce the final decision. In this way, we can have the advantages of both early and late integrations at the same time.

C. Data Sets

Although there are many AV databases, there is still a great need to produce appropriate databases for AV applications. None of the existing data sets have all desirable characteristics such as adequate data size, realistic variability, standard experimental settings, and evaluation measures. Additionally, there is no commonly accepted standard evaluation which makes the comparison of different features and fusion methods difficult [22], [57]. Some of the available AV data sets which have been used in the literature are PETS [75] (multimodal analysis tasks, for example, object tracking), AV16.3 [76] (audio-only, video-only, and AV speaker localization and tracking) TRECVID [77] (used in different applications like video retrieval, semantic video analysis, video segmentation, concept detection), BIOMET [78] (contains face, speech, fingerprint, hand, and signature modalities), M2VTS [79] (audio and video recordings of some subjects uttering digits, used in biometric verification applications), XM2VTS [80] (extended M2VTS), VidTIMIT [81] (video recordings of people reciting sentences from the TIMIT [82] corpus), DAVID [83], VALID [84], AVICAR [85] (AV speech corpus in a car environment), BANCA [86] (biometric access control for networked and e-commerce applications), and CUAVE [87].

III. FUSION TECHNIQUES

There are a number of techniques which have been used for the modeling and fusion steps in AV processing, such as SVMs, graphical models, e.g., DBNs and HMMs, neural networks, and estimation algorithms, for example, Kalman filtering. Generally, these are modeling techniques which are applicable to various parts of an AV system; for example, they can be used as a technique in the modeling process units (see Fig. 2). As the focus of the current study is on fusion, we will not cover such works that have used these modeling techniques in other parts of the system and simply focus on the integration units. In the following, we provide a short description of some of the most commonly used techniques for fusion in AV applications.

A. Support Vector Machines

SVMs represent popular modeling techniques that have been used widely in many classification problems. In most of the AV works that have utilized SVMs, they have been used for modeling a single modality independently. However, there are studies, particularly in late integration, which have used SVMs as a fusion technique to integrate the decisions obtained from the other components of the system. For example, many studies have been performed on AV concept detection in videos [26], [27], where some audio, visual and textual clues are modeled separately, and the corresponding scores are produced. The obtained scores are then concatenated to form a feature vector that will be the input to an SVM for detecting semantic concepts. The same idea is utilized in other applications such as biometric identification. Bredin and Chollet [19] combine the scores obtained from three components including a face recognition system, a speaker verification system, and a synchrony (correlation) estimation module using an SVM as the decision integration unit.

B. Dynamic Bayesian Networks

Bayesian networks are probabilistic graphical models that represent a set of random variables with their conditional dependencies. The graphical representation of a Bayesian network is done with acyclic directed graphs in which a vertex represents each variable, and the conditional dependency between two variables is represented by an edge between the corresponding vertices. DBNs are Bayesian networks that model sequences of observations. DBNs and their variants are used widely in AV applications especially where temporal sequencing should be considered, such as speech processing and video analysis.

Noulas and Kröse [88] have suggested a two-layer DBN modeling approach to be used in video analysis applications to address the problem of assigning clues to the person that created them. In the first layer, each modality, in this case audio and visual, is modeled independently with a separate DBN. In the second layer, another DBN is employed to model the interaction between these two modalities. The use of expectation–maximization (EM) is proposed to estimate the parameters of the DBNs. Other researchers have also suggested using multistream DBNs to model the interactions between modalities. For example, Dielmann and Renals [42] used an automatic meeting segmentation system to analyze the meeting videos based on multistream DBNs. The goal is to automatically structure a meeting recorded with several microphones and cameras into sequences of group meeting actions, such as monologue, discussion, and presentation. They have proposed to model the AV clues jointly with a multistream DBN that relates low-level features to more complex group behaviors.

In the general multistream DBN model structure given by Bilmes and Bartels [89], both in the AV streams, each word is composed of a fixed number of states, and each state is associated with observation vectors. The number of training parameters is very large, especially for the task of large vocabulary speech recognition. To reduce the training parameters, each word is composed of its corresponding phones sequence, and each phone is associated with an observation vector. Since phones are shared by all words, the training parameters are enormously reduced [this is referred to as the multistream asynchrony DBN (MS–ADBN) model]. This model however is a word model whose recognition basic units are words. Based on the MS–ADBN model, Lv *et al.* [90] introduced an extra hidden node level state between the phone node level and the observation variable level in both streams, resulting in multistream multistates asynchrony DBN (MM–ADBN) model. In it, each phone is composed of a fixed number of states, and each state is associated with an observation vector; besides word, a dynamic pronunciation process of phone is also described. Terry and Katsaggelos [11] introduced an extension to this model. In [89], the AV streams are modeled independently with each consisting of phones as subword units (referred to as phone/phone model). AV speech, however, is not composed of the same subword units. The visual speech units, visemes, are related to audio speech units, phones, in a many-to-many mapping [91]. The approach in [11] reflects this and models the audio stream as consisting of phones, while the video stream consists of visemes (phone/viseme model).

DBNs have been employed in various AV fusion tasks that need to model the multiple dependencies between their random variables. Furthermore, they can efficiently deal with the time-series data [92]. These advantages make them suitable for many multimedia analysis tasks. The main disadvantage of DBNs is the difficulty in determining their correct state [57], [93].

C. Hidden Markov Models

An HMM can be considered as a simple form of a DBN that represents probability distributions over sequences of observations. Like DBNs, HMMs have been used widely in speech and video processing tasks. In some works, a single HMM is exploited to model AV features jointly, without discriminating between them. These works can be categorized as early integration approaches [51], [94]. For example, an HMM is used by Wang *et al.* [51] to model the extracted AV features from each frame to perform video shot detection. On the contrary, several variations of HMMs are proposed as intermediate integration techniques that try to model the modalities separately while considering their interactions at the same time.

Multistream HMMs (MSHMMs) employ two separate streams for the audio and video observations. They couple these observations at every frame. The complexity of the decoding algorithm is linear in the number of streams [95]. This modeling has been used extensively in AV speech recognition applications [9], [12], [13]. Instead of coupling observations at every frame, which may be too tight, in state-asynchronous HMMs [96], two standards HMMs are tied at the boundaries of the modalities. In this way, the asynchrony between modalities and also their alignments can be maintained. In another variant of HMM, called coupled hidden Markov models (CHMMs) [97], parallel streams are modeled using concurrent HMMs where each HMM state can transit within the asynchronous region, but should remain tied at the model boundaries [98]. The main problem with CHMMs and state-asynchronous HMMs is that their exact training algorithms become intractable when using more than two streams [95]. The readers may refer to [98] for a complete explanation of the various types of HMMs, their advantages, and disadvantages.

In addition to DBNs and HMMs, other types of graphical models, such as conditional random fields (CRFs) and their variations [99], have been used for multimodal fusion [100].

D. Estimation-Based Methods

The estimation-based techniques for fusion of multiple sources include variants of the Kalman and particle filter methods [57]. Kalman filtering is a technique to estimate a state–space model from a sequence of noisy observations over time. It is able to retain the history of its previous states without extra memory. The Kalman filter represents an optimal estimator for 1-D linear systems with additive Gaussian noise [101]. A nonlinear version of the Kalman filter referred to as extended Kalman filter (EKF) [102] is used to model nonlinear systems.

Particle filters are used to model stochastic dynamical systems from a series of observations over a period of time. These methods are also known as sequential Monte Carlo (SMC) methods [103]. While the Kalman filter is typically used to model linear systems, and the extended Kalman filter can be used for nonlinear systems, particle filters are more suitable for nonlinear and non-Gaussian models, particularly with sufficiently large number of samples.

These are popular techniques in object localization, people tracking, and also data fusion. They can be employed at both feature and decision levels of fusion. For example, Loh *et al.* [39] combined the audio data from three microphones and the video data from one camera to estimate the position of the speaker, and then a Kalman filter is employed to estimate her/his velocity and acceleration. Gehring *et al.* [37] provided the recognized faces from different cameras and the time delay of arrival (TDOA) between different microphones as audio and video features, respectively, to an EKF to detect the active speaker location. A hierarchical Kalman filter structure has been proposed by Talantzis *et al.* [30] to track people in a 3-D space using multiple microphones and cameras. First, two separate local Kalman filters for audio and video streams are considered. Then, the outputs of these two local filters are fused employing one global Kalman filter. Kilic *et al.* [104] proposed a novel approach for integration of audio and video information for tracking multiple moving speakers using particle filtering. They reshape the traditional Gaussian noise distribution of particles in the propagation step and reweight the observation model in the measurement step by exploiting the audio information and the direction of arrival (DOA) angle.

E. Task-Dependent Techniques

Additional AV fusion techniques can be found in the literature that were developed for specific applications, often with no general applicability. These fusion techniques are mostly considered as intermediate approaches. For example, Casanovas *et al.* [105] proposed a method based on sparse representation for blind AV source separation. Two dictionaries are constructed expressing redundant representations of the audio and video modalities. Extending the idea of using two dictionaries to model audio and video observations separately, the "local" information has been exploited by using a unique AV dictionary as in [55].

In their other work [106], an iterative video diffusion technique is proposed that detects regions in the video that are related to the produced sound. A measurement of synchrony between audio and visual modalities is exploited to recognize these regions. The extracted regions can be useful in several AV applications such as audio source localization in videos. This technique has also been employed to extract objects producing sound in a video in an unsupervised manner [6].

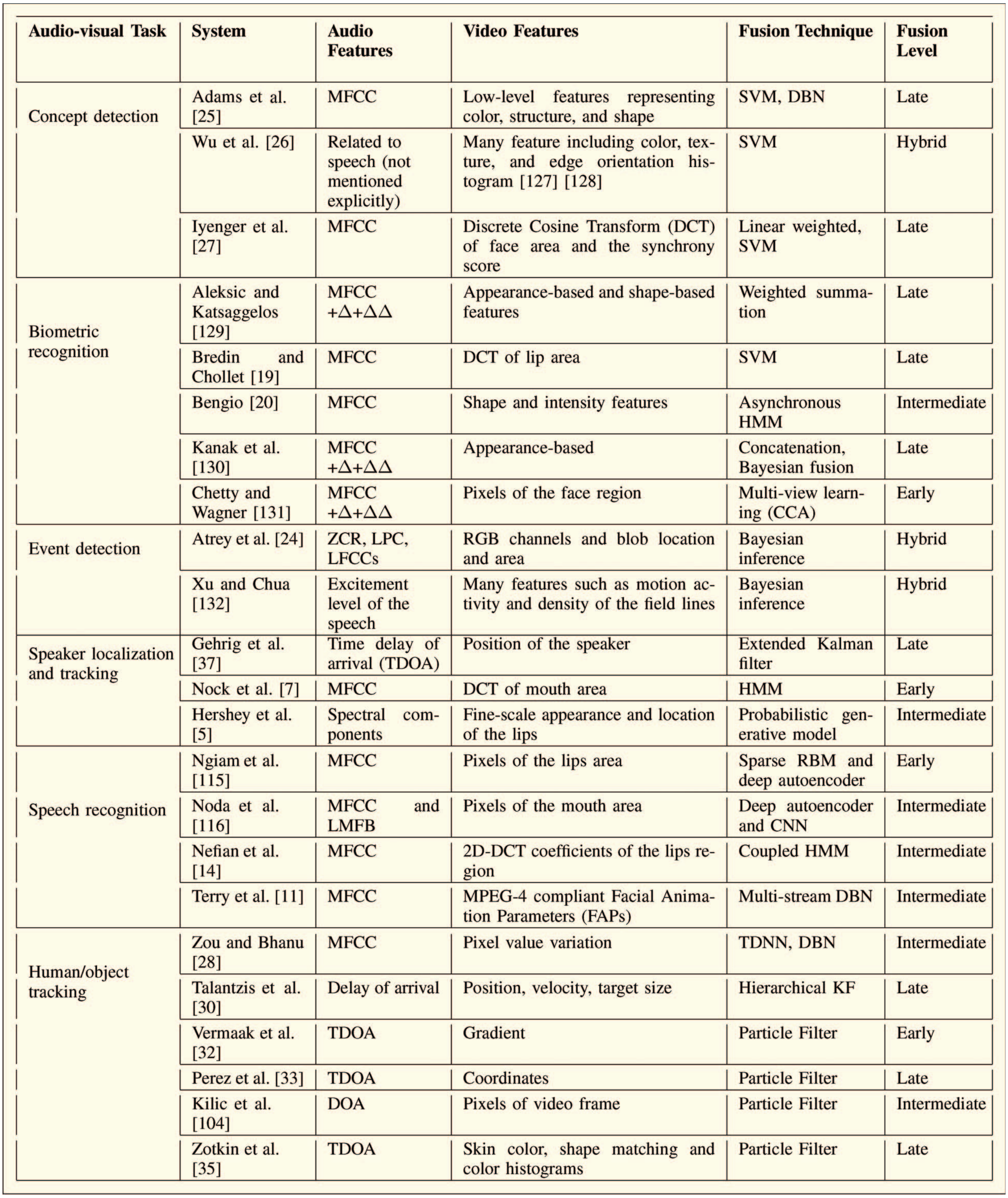
A summary of applications involving AV fusion is shown in Table 1. Representative works under each application are also shown. For each of them the AV features used along with the actual fusion technique and its classification are also shown. This is just a representative list of papers and by no means exhaustive. Additional applications can be found in [10], [22], [57], and the references therein.

IV. CHALLENGES

An important issue in designing an AV system is how to integrate knowledge of diverse modalities (in our case audio and video) to exploit the informative knowledge from each modality while ignoring drawbacks of each one. In the following, some of the main challenges in this area are described.

- The effectiveness of each modality in different environmental conditions is not the same. In some cases, the system should rely more on the audio, for example in a dark scene, while in others it should rely more on the video, such as in acoustically noisy environments. In other words, the system should be adaptive to the quality, reliability, and confidence of modalities. The general approach to this goal is to consider weights for each modality during fusion. Weighting can be done in a dynamic scheme by adjusting the weights constantly according to the quality of test data [101], [107]–[109], or in a static scheme by calculating some constant weights based on only the training data [46], [110], [111]. In cases where the qualities of modalities in the training and test data are different, dynamic weighting is necessary. The problem of estimating the appropriate weights for varying conditions is still an open problem, although many researchers [112], [113] have addressed it.
- Dealing with multiple modalities of different types can cause many synchronization issues. There are two main types of asynchrony in AV fusion. The first type originates from the asynchrony between the audio and video streams. For example, the visual and acoustic signs of speech do not necessarily occur exactly at the same time. As a result, there is a natural asynchrony between speech and visual clues in AV speech recognition referred to as preservatory and anticipatory coarticulation [114]. The other type is related to the difference between sensing rate and processing time of different modalities. Also, the amount of data that is needed to complete a specific task is application dependent; for example, this amount is longer for AV event detection compared to AV speech recognition. Handling asynchrony is an important and critical problem in real-world applications and should be studied and addressed properly.
- These days, large amounts of data are available which are mostly unlabeled. The process of labeling data requires human effort which is time

Table 1 Summary of Audiovisual Applications (Adopted and Extended Based on [22] and [57])



Audio-visual Task	System	Audio Features	Video Features	Fusion Technique	Fusion Level
Concept detection	Adams et al. [25]	MFCC	Low-level features representing color, structure, and shape	SVM, DBN	Late
	Wu et al. [26]	Related to speech (not mentioned explicitly)	Many feature including color, texture, and edge orientation histogram [127] [128]	SVM	Hybrid
	Iyenger et al. [27]	MFCC	Discrete Cosine Transform (DCT) of face area and the synchrony score	Linear weighted, SVM	Late
Biometric recognition	Aleksic and Katsaggelos [129]	MFCC +Δ+ΔΔ	Appearance-based and shape-based features	Weighted summation	Late
	Bredin and Chollet [19]	MFCC	DCT of lip area	SVM	Late
	Bengio [20]	MFCC	Shape and intensity features	Asynchronous HMM	Intermediate
	Kanak et al. [130]	MFCC +Δ+ΔΔ	Appearance-based	Concatenation, Bayesian fusion	Late
	Chetty and Wagner [131]	MFCC +Δ+ΔΔ	Pixels of the face region	Multi-view learning (CCA)	Early
Event detection	Atrey et al. [24]	ZCR, LPC, LFCCs	RGB channels and blob location and area	Bayesian inference	Hybrid
	Xu and Chua [132]	Excitement level of the speech	Many features such as motion activity and density of the field lines	Bayesian inference	Hybrid
Speaker localization and tracking	Gehrig et al. [37]	Time delay of arrival (TDOA)	Position of the speaker	Extended Kalman filter	Late
	Nock et al. [7]	MFCC	DCT of mouth area	HMM	Early
	Hershey et al. [5]	Spectral components	Fine-scale appearance and location of the lips	Probabilistic generative model	Intermediate
Speech recognition	Ngiam et al. [115]	MFCC	Pixels of the lips area	Sparse RBM and deep autoencoder	Early
	Noda et al. [116]	MFCC and LMFB	Pixels of the mouth area	Deep autoencoder and CNN	Intermediate
	Nefian et al. [14]	MFCC	2D-DCT coefficients of the lips region	Coupled HMM	Intermediate
	Terry et al. [11]	MFCC	MPEG-4 compliant Facial Animation Parameters (FAPs)	Multi-stream DBN	Intermediate
Human/object tracking	Zou and Bhanu [28]	MFCC	Pixel value variation	TDNN, DBN	Intermediate
	Talantzis et al. [30]	Delay of arrival	Position, velocity, target size	Hierarchical KF	Late
	Vermaak et al. [32]	TDOA	Gradient	Particle Filter	Early
	Perez et al. [33]	TDOA	Coordinates	Particle Filter	Late
	Kilic et al. [104]	DOA	Pixels of video frame	Particle Filter	Intermediate
	Zotkin et al. [35]	TDOA	Skin color, shape matching and color histograms	Particle Filter	Late

consuming and expensive. It is necessary to have a fusion technique that is able to benefit from such a large amount of unlabeled resource. Exploiting unlabeled data is not considered in most of the conventional AV techniques. However, recently researchers [115], [116] have been working on AV processing in semisupervised or even unsupervised scenarios. They mostly view the multimodal processing problem as a multiview learning problem, and propose new learning techniques to address issues such as missing labels, noisy views (modality), and semisupervised learning.

V. RECENT ADVANCES AND APPROACHES

Having established in the previous section the main AV fusion challenges we choose to focus on, in this section, we first review the recent literature in addressing the asynchrony and dynamic weighting challenges. The nature of the first challenge is specific to the two modalities under consideration: speech and video. The approaches described in addressing it, however, can be applied in handling the asynchrony of other modalities as well. The dynamic weighting challenge on the other hand is generic in some sense, that is, it applies to any fusion application. Subsequently, we describe two recent technologies, deep and multiview learning with both current and future impact on the AV fusion. Although the amount of work utilizing these two technologies toward AV fusion is limited, they have demonstrated improved performance and that they are capable in principle of addressing the challenges of unlabeled, noisy, missing and/or conflicting data.

A. Asynchrony

AV anticipatory asynchrony is a naturally occurring linguistic phenomenon in which the visible gestures (mainly the lip gesture) for a speech segment occur in advance of other articulatory components of the segment, so that the visible gestures are seen before the corresponding phone is heard. A common example of this is the prerounding seen in the word "school." The lips begin to round for the /uw/ sound while the /k/ (or even /s/) is still being produced. This phenomenon is known as anticipatory coarticulation. Preservatory coarticulation is a similar effect, but instead of one gesture beginning in advance, a gesture continues after. Though anticipatory coarticulation is more pervasive in English, the extent and directionality of coarticulation patterns differ across languages [117], [118].

Anticipatory coarticulation has been studied since at least the 1930s under the assumption that coarticulation occurs in part because segments may lack inherent specification for particular articulations [119]. In 1966, Henke proposed a computational model of the articulation of English stop + vowel sequences under the assumption that segments need not always have complete articulatory targets, and, thus, are open to coarticulation effects [120]. This work, known for its "look-ahead" mechanism for anticipatory coarticulation, proposed that once the stop contact is made, the stop looks ahead to the vowel's targets for other articulators, such as the lip-rounding present in "school." In the speech recognition literature, it was shown by Bregler and Konig [121] that, on average, acoustic features are maximally correlated with visual features 120 ms in the past. This was also reported in psychological experiments by Benoit [122]. In the case of AV biometrics, Aleksic and Katsaggelos [22] cite these asynchrony effects as one of the major open problems.

One of the many problems in AV processing is the lack of sufficient corpora for system development [123]. A good database is an essential component of a research plan and must contain the phenomena one is trying to model. The GRID corpus [124] contains many linguistic contexts in which one may find AV asynchrony and served as the primary database for the work in [114]. To aid in the labeling and analysis tasks, an AV data display (AVDDisplay) tool was developed in [113], which provides interfaces for both human annotation and display and manipulation of automatically produced alignments and recognition hypotheses. Using AVDDisplay human labeled data were collected and were utilized to establish the ground truth [113], [114].

In analyzing the human labeled data it was concluded that the cross-annotator synchrony characterizations were very consistent [114]. The overall asynchrony data conformed to our linguistic expectations that the data should be skewed toward early video onsets. The histogram of the amount of asynchrony at each onset, measured as the video mark minus the audio mark, is shown in Fig. 3. The histogram is centered near the boundary between synchrony and early video (20 ms) and is significantly skewed toward early video.

Currently, a typical approach to modeling asynchrony in AV speech modeling is the coupled HMM (CHMM) [125], in which state transitions in each modality depend on the state of the other modality (an alternative approach is also represented by asynchronous multistream HMM). In CHMM, asynchrony is typically allowed only within the boundaries of each phone/viseme, whereas observed asynchrony often crosses multiple phone boundaries. In contrast, the asynchronous dynamic Bayesian network model of Saenko and Livescu [112] and Saenko *et al.* [126] allows asynchrony across multiple phones/visemes within a word,

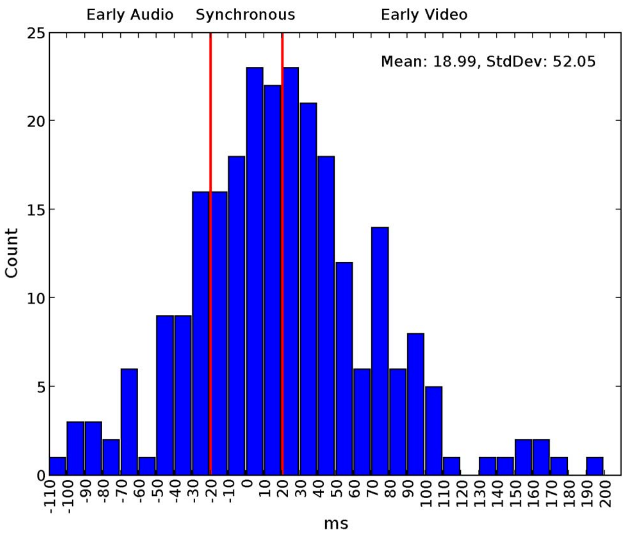


Fig. 3. ***Histogram of asynchrony distribution (video mark/audio mark) in ground truth data for all words. Red lines indicate boundaries between early audio, synchronous, and early video cases [114].***

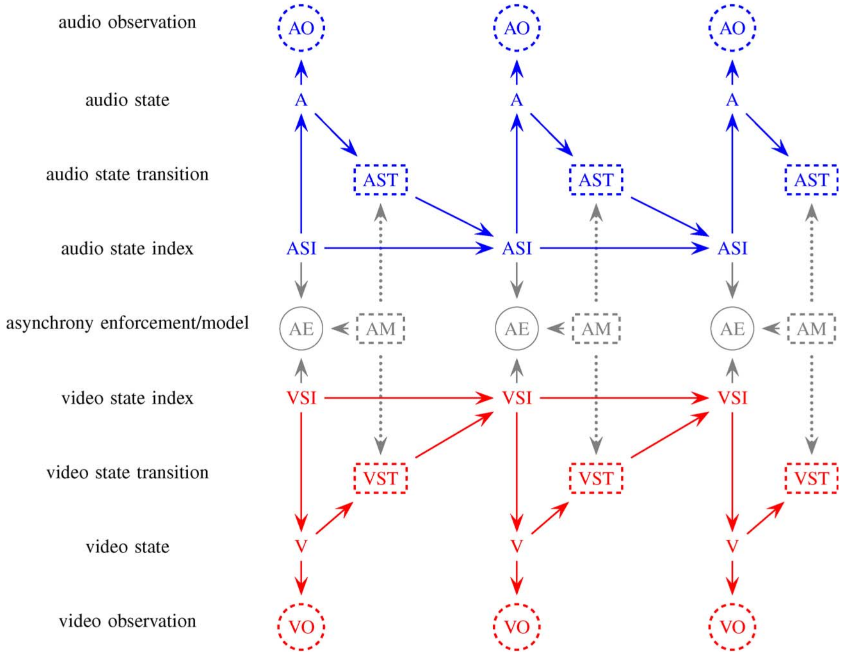


Fig. 4. *Word-synchronous SD and ADT models for training/alignment. All variables are the same for both models and dotted edges are excluded in the SD model and included in the ADT model. Diagram is simplified for clarity and is conditioned on word-level variables that are not shown [114].*

but does not account for the asymmetry that is typical to AV asynchrony.

Terry [114] developed a model of asynchrony that allows for explicit modeling of anticipatory coarticulation, while spanning multiple phones/visemes. Furthermore, as speech generally appears synchronous with sporadic bursts of asynchrony, it is posited that an AV speech system would benefit from modeling these two regimes, synchrony and asynchrony, in different manners. Based on the assumption that AV speech in a state of asynchrony will tend to revert back to synchrony, it was hypothesized that the state transitions of each modality will differ based on the amount of asynchrony. To model this, asynchrony-dependent state transitions are introduced. The state transitions of each modality in this new model depend on both the current degree of asynchrony and the modality's current state.

The model in [114] is based on the word-synchronous DBN used in [11] with the addition of a synchrony control mechanism based on [112], [126]. This model also takes inspiration from CHMMs [125] in that it allows state transitions to depend on variables other than just the current modality's state. In this case, however, the dependency is on the instantaneous asynchrony rather than the state itself, which reduces the number of parameters. In [112] and [126], the amount of asynchrony is defined as the absolute value of the difference between the state indices of the streams. The work in [114] drops the absolute value, which increases the number of parameters in the model but allows to more correctly model the difference between audio lead and audio lag. This asynchrony model is learned during training. In addition to the asynchrony model itself, an extra parameter, the asynchrony model weight, is added to control the relative importance of the asynchrony model.

A model with asynchrony-dependent transitions is denoted as the "ADT" model and a model with the standard transitions and asynchrony mechanism as the state-differences or "SD" model. Therefore, aside from AV stream weights, there are three main tuning parameters of these models: the maximum number of states of audio lag, the maximum number of states of video lag, and the weighting of the asynchrony model. Fig. 4 shows the SD and ADT models as DBNs. For clarity, state and phone/viseme level variables have been collapsed into single nodes in the graph. Also, some common elements, such as pronunciation variants and stream weighting, are not shown. Blue nodes and edges represent the audio modality, while red nodes and edges represent the video. The gray nodes and edges denote the asynchrony model and its links to the AV modalities. Nodes with no border are deterministic and hidden, while nodes with a solid circular border are deterministic and observed. Dashed rectangular borders denote hidden, stochastic nodes and dashed circular borders denote observed, stochastic nodes. The observed audio and video input nodes have Gaussian mixture distributions conditioned on their respective state.

- AV state index (ASI, VSI): The index of the current state relative to the most recent word boundary, used to determine the current phone/viseme and for measuring asynchrony.

- AV state (A, V): The current AV phone/viseme and subphone/viseme state.
- AV state transition (AST, VST): A binary variable indicating whether an AV state transition has occurred. For the SD model, the distribution is conditioned only on the AV state, while for the ADT model, the distribution is conditioned on the AV state and the value of the asynchrony model (AM, described below). These distributions are learned during training. If a state transition occurs, the state index at the next time instant will increment or reset to zero in the case of a word transition.
- AV observation (AO, VO): Acoustic and visual feature vectors, distributed according to state-specific Gaussian mixture models.
- Asynchrony model (AM): The instantaneous degree of asynchrony (difference between the audio and video state indices). Its probability mass function over the set of allowed asynchrony values represents the probability of a given number of states of audio or video lag.
- Asynchrony enforcement (AE): A binary variable with observed value always equal to one that enforces the asynchrony constraints by ensuring that $\mathrm{ASI}(t) - \mathrm{VSI}(t) = \mathrm{AM}(t)$, where t denotes time. As explained in [126], this variable is not needed for decoding, but is needed for training the asynchrony model distribution with standard EM.

The AV speech modeling system in [114] was evaluated in the context of a forced alignment task using the GRID [124] corpus. It was found that the state transition probabilities for /uw/ and /r/ share similar characteristics, and as expected, the probability of transitioning varies greatly depending on the asynchrony state. For situations when the audio is lagging the video, the video is very unlikely to transition until the audio has caught up (i.e., the asynchrony state returns to synchronous). Similarly, when the video is lagging, the video is very likely to transition when it has caught up with the audio and returned to synchrony.

Regarding the partitioning of the data that was used in [114], ten speakers were selected from the GRID corpus: speakers 2, 3, 4, 10, 15, 18, 19, 20, 22, 24. These speakers were selected for more neutral accent as well as ease of tracking for visual feature extraction. Utterances were pooled into three mutually exclusive sets, one for training, one for development, and one for testing. For each speaker, 700 of the 1000 total utterances were randomly selected for the training set, 100 randomly selected for development, and the remaining 200 were set aside for testing. Thus, the total sizes of the training, development, and testing sets are 7000, 1000, and 2000, respectively.

Besides the AV forced alignment the ADT system was also used for speech recognition. It was found that the overall word recognition rate improvement is rather small, but, interestingly, there is a significant improvement in the first word recognition.

B. Dynamic Weighting

It is a well-known fact that the performance of automatic speech recognition (ASR) systems degrades heavily in the presence of noise. Consequently, the problem of weighting AV modalities for speech classification naturally arises at the description or observation levels. The weight assigned to each modality should be related to its reliability to perform classification. For instance, in a quiet environment with ideal AV signals, a larger weight should be given to the audio stream, reflecting the fact that the audio modality is more reliable than the video one when it comes to recognizing speech. In general, when one of the modalities is degraded (due, for instance, to background noise in the audio channel or an occlusion of the speaker's mouth in the visual signal) the importance assigned to it should decrease and reflect the confidence we have on that modality in such circumstances. Let us now examine how the weighting of the contribution of the audio and video signals in various scenarios has been approached in the literature. It should be kept in mind that stronger constraints must be imposed on the weights other than their sum being equal to one [133]. Often the weights are tuned on heldout data (e.g., [112] and [113]). It is also interesting to note that there is often a mismatch between the implicit weighting used during training and the weights applied at test time. Terry *et al.* [113] report that the best performance of their system was achieved by tuning the training and testing weights separately.

One of the earliest and most cited papers on weighting is by Potamianos and Graf [134]. The authors utilize synchronized AV features to train respectively audio-only and visual-only single stream HMMs of identical topology by maximum likelihood. A two-stream HMM is obtained by combining the two single stream HMMs; exponents that weight the log-likelihood of each stream are then introduced. They use the minimum classification error discriminative criterion to estimate the exponents. However other criteria can also be used; see, for instance, [135] for the use of maximum mutual information to perform the same task and [136] for the use of maximum entropy principles.

The approach of Potamianos and Graf was adopted by various researchers. For example, Garg *et al.* [137], also use MSHMMs and they propose two reliability indicators of the class information contained in an observation which are then calculated for the AV streams. The exponents are modeled as the sigmoid of a weighted function of the four calculated reliability indicators. The weights associated with each indicator are calculated using maximum conditional likelihood of the training data labels.

Advancing the approach introduced in [138] and [137], and utilizing the same model, Marcheret *et al.* [139] concentrate on feature selection for capturing the reliability of the AV streams, and weight estimation based on such features. They consider likelihood, as was done in previous works, and also analyze acoustic signal based features. To estimate the weight, sigmoid functions are used and two variants of the Gaussian mixture model (GMM) estimation are proposed.

The approach followed by Gurban *et al.* [140] is also based on finding estimators of stream reliability and mapping them dynamically to stream weights. The authors estimate stream confidence directly from each classifier. If a clear peak emerges in the posterior distribution, the stream is reliable; otherwise ambiguity is strong and the modality is unreliable. It uses entropy to measure the stream reliability. Several mappings from entropy to weights are proposed. Lee and Park [141] discuss and compare different definitions of the reliability of a modality. Rajavel and Sathidevi [142] propose a genetic-algorithm-based reliability measure and the final weights are proportional to the reliability measure of the outputs of the acoustic and visual HMMs. They describe a neural-network-based fusion method which uses the reliability measures of the two modalities and produces noise-robust recognition performance over various noise conditions.

Terry *et al.* [143] propose a video reliability metric based on the extracted video features rather than the video sequence itself. The features are extracted from clean data and are sent through a vector quantizer with memory so that the conditional probability mass function (PMF) of a video state given an audio state is estimated during training. This conditional PMF along with an audio stream reliability metric, such as audio signal-to-noise ratio (SNR), are utilized to determine the AV stream weights at any given time.

In a multispeaker environment and to make the system robust to acoustic noise, Shao and Barker [144] replace the state likelihoods with a score based on a weighted combination of the AV likelihood components and the weights are allowed to change from frame to frame. The proposed weighting process learns the SNR from the complete likelihood data using an artificial neural network (ANN). The SNR is also used as a reliability measure in the work of Estellers *et al.* [145]. They propose a dynamic scheme in which weights are derived from instantaneous measures of the stream reliability. The authors propose a confidence measure on the audio stream and study how to map it to weights in order to obtain minimum word error rate in a noisy training data set.

Various approaches for determining the stream weights have been followed when CHMMs are utilized toward AV ASR. For example, Nefian *et al.* [146] modified the probability for each observation conditional likelihood to handle different levels of noise. The weights assigned to each modality are obtained experimentally to maximize the average recognition rate for a specific acoustic SNR level. The use of a multilayer perceptron with a CHMM is investigated by Abdelaziz and Kolossa [109]. Finally, Addelaziz *et al.* [147] used the EM algorithm to estimate the dynamic stream weights in the context of a CHMM.

Terry and Katsaggelos [11] introduced a new model for AV automatic speech recognition with DBNs. Stream weighting is directly incorporated into the graphical model and a phone/phone model is transitioned into a phone/viseme model. The system is evaluated and compared against one of the recently proposed systems utilizing a large vocabulary continuous speech recognition (LVCSR) task with noisy audio. By modeling the visual stream more accurately through the use of visemes, the system provides a higher recognition rate. The integration of the information provided by AV signals is carried out by Heckmann *et al.* [138] at the posterior probability level, using the so-called separate integration model. They analyze different weighting schemes and their coefficients are learned using an ANN/HMM on noiseless environments (see also [148]).

C. Deep Learning

A definition of deep learning (DL) is [149]: "A class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification." It lies in the intersection of neural networks, artificial intelligence, graphical modeling, optimization, pattern recognition, and signal processing. Human information processing mechanisms (e.g., vision and audition) suggest the need for deep architectures to extract complex structures and building internal representations from rich sensory inputs. DL has shown very good performance in a number of research areas, such as, object recognition, computer vision, information retrieval, language modeling and natural language processing [149]. It has also been used for multimodal fusion [150]–[153] and representation learning in AV fusion [115]. Ngiam *et al.* [115] introduced three main deep representation learning methods, which we also adopt here in organizing the paper, namely:

- multimodal fusion learning;
- cross-modality learning;
- shared-representation learning.

All three learning methods include the following three phases: 1) unsupervised deep feature learning; 2) supervised training; and 3) testing. Deep networks have been applied to unsupervised feature learning, that is, the network is used as an audio and video feature extractor; the resulting features are in turn utilized in the training and testing phases of all the three learning methods. We review next the literature along the lines of the three learning methods mentioned above.

1) Multimodal Fusion Learning: In the multimodal fusion learning setting, both modalities are available in all three phases, as is the case in most multimodal works. One option is to train deep neural networks separately for the

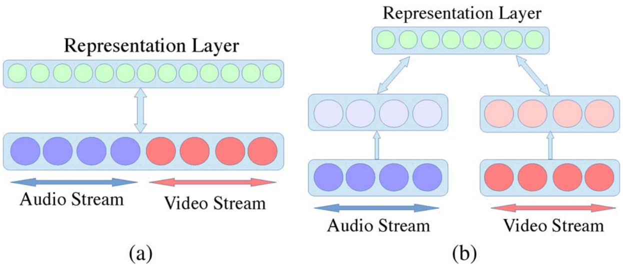


Fig. 5. ***Network architectures used for feature learning (adopted from [115]). (a) Concatenating audio and video vectors and employing a single input network. (b) Two-input network with separate inputs for audio and video streams.***

audio and video streams. Then the output of the trained model, i.e., extracted features, can be used as the new representation for the data. Another alternative is to train a model over the concatenated audio and video data [see Fig. 5(a)]. A third alternative is to greedily train a deep model over the pretrained layers for each modality. In AV speech recognition, the outputs obtained from the two separate models for AV inputs can be considered informally as phonemes and visemes, respectively. The outputs are then given to another layer to model the relationship between the modalities [see Fig. 5(b)]. This model is motivated by the stacking idea in DL modeling.

Several researchers have adopted this strategy in AV fusion. An example is represented by the work of Ngiam *et al.* [115] for speech classification. They use restricted Boltzmann machines (RBMs[1]) [154] for feature learning and investigate many learning architectures. They train separate RBMs for audio and video, one shallow RBM model for the concatenated audio and video features [see Fig. 5(a)], and also a bimodal deep belief network (DeBN[2]) model [see Fig. 5(b)].

In another work, Kim *et al.* [155] used some DeBN models, similar to the ones introduced by Ngiam *et al.* [115], for an emotion detection task. They test several feature selection techniques performed either before providing the modalities to the input layer or after obtaining features from the output layer. They have also investigated the effect of feature reduction by adding a new layer with lower number of nodes to the last layer of a DeBN. Better performance is achieved compared to the baseline system especially for nonprototypical data for which there is not complete agreement.

Noda *et al.* [116] have also proposed a noise robust AV ASR system by utilizing two different models to extract noise robust features from audio and video. They employ a deep denoising autoencoder and a convolutional neural network (CNN) encoder to represent AV features, respectively. Artificial Gaussian noise with various strengths is added to the audio features, i.e., MFCC and log melscale filter bank (LMFB), to produce noisy features. These features with clean ones are exploited to train a deep denoising autoencoder. To learn a video representation, a CNN is trained with visual features, i.e., pixels from the mouth area, as inputs and phoneme labels as outputs. The outputs of the autoencoder and CNN are modeled with an MSHMM. The system suffers from the static weights for the audio and video streams in the MSHMM. Also, an independent CNN should be trained for each speaker. However, they demonstrate the effectiveness of their approach in providing a noise robust representation for audio and video using DL techniques.

In another work of Huang and Kingsbury [156], AV inputs are provided to two separate DeBNs. The outputs obtained from the two DeBNs have been exploited in two ways, as: 1) scores to estimate the posterior probabilities; these scores are then integrated and used as the state posterior probabilities for the HMM; 2) a midlevel representation; the outputs of two DeBNs are concatenated and given to a third DeBN and then used as the input for a conventional GMM–HMM system. AV continuous digit recognition was the task used in their experiments. It was shown that their two DeBN-based systems perform better in noisy environments compared to conventional GMM/HMM systems, but not in clean conditions.

2) *Cross-Modality Learning:* Compared to multimodal fusion learning, with this method, only a single modality can be presented at training and testing. This technique is beneficial in situations when unlabeled data from other modalities are available for training deep networks for feature learning but they are not available in the next two phases. A deep autoencoder proposed in [115] uses a cross-modality learning method. Initially, a DeBN [the same structure as in Fig. 5(b)] is trained with all modalities. Then, the output of the layer corresponding to the available modality is provided during testing to two networks, e.g., RBMs, to reconstruct both modalities. After training, the output of the middle layer of the deep encoder can be used as a new feature representation. This deep encoder can reconstruct other modalities with only one of the modalities [see Fig. 6(a)] by discovering the correlation between modalities. Ngiam *et al.* [115] could achieve better representation for the video in the case of availability of video and absence of audio in the training and testing phases.

3) *Shared-Representation Learning:* The problem with cross-modality learning is that, for cases with multiple modalities, the number of models that need to be trained increases exponentially. To address this problem, a complete bimodal deep autoencoder is proposed in [115] using artificially noisy data. Motivated by deep denoising autoencoders, examples with one modality set to zero are

[1] An RBM is a generative model and can be used to learn a data representation in an unsupervised manner.

[2] It is a type of deep neural network, composed of multiple layers of hidden units. It can be viewed as a composition of simple, unsupervised networks such as RBMs

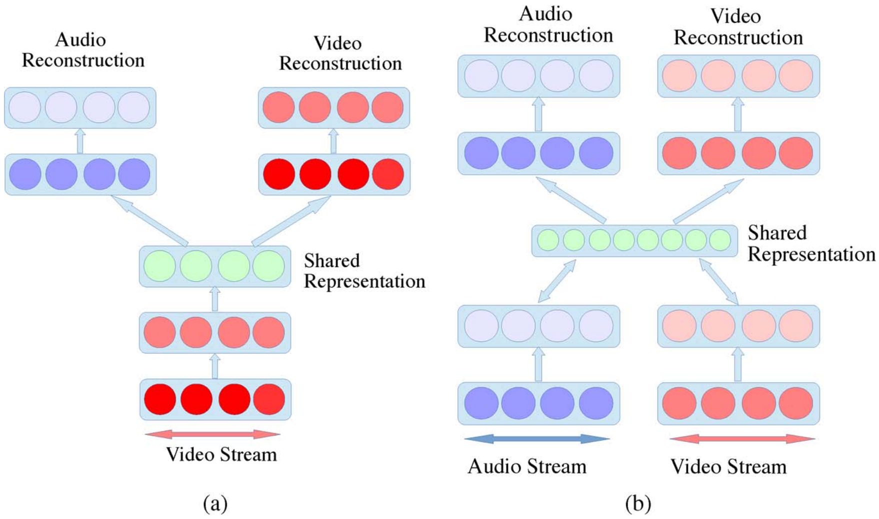


Fig. 6. *Deep autoencoders in cross-modality and shared-representation learning (adopted from [115]). (a) Cross-modality learning by using a single input autoencoder. (b) Shared-representation learning by using two-input autoencoder.*

added to the training data. This way, the learned model will be robust to missing modalities, and different combinations of modalities can be utilized in the supervised learning and testing phases [see Fig. 6(b)].

Here are some remarks on DL methods.

- The multimodal fusion learning method is the most widely used deep AV fusion method. Various deep networks architectures can be used with this method, which provides the possibility of adaptation based on the task at hand and the available resources. However, the particular architecture should be chosen very carefully. For example, since the correlations between the raw audio and video data are highly nonlinear, it is hard for a network to learn these correlations from concatenated features [Fig. 5(b)] without using a sufficiently *deep* network.
- The main issue with the multimodal fusion learning method is that all modalities should be available during all three phases: feature learning, training, and testing. This is certainly not always possible. On the other hand, these days, large amounts of unlabeled data are available. It would be very beneficial to have a method to make use of these data for feature learning. This can be done with cross-modality and shared-representation learning.
- Multimodal learning is closely related to the concept of multitask learning, a machine learning approach that learns to solve several related problems at the same time, using a shared representation [149]. The learning domains or tasks cut across several modalities. Multitask learning is often applied to conditions where no or very little training data are available for the target task domain. It is evident that multitask learning naturally fits the paradigm of DL where the shared representations and statistical strengths across tasks (e.g., those involving separate modalities of audio and video) are expected to greatly facilitate many machine learning scenarios under low- or zero-resource conditions.
- Although in many of the DL-based AV fusion techniques, RBM, DeBN, and CNN are used, other variants of models can also be employed depending on the task and resources. For example, Shah *et al.* [157] proposes a multimodal emotion recognition framework using an energy-based variant of RBMs, known as replicated softmax model (RSM). The effectiveness of the approach toward emotion recognition was tested with facial expressions, speech, and language as source data.

D. Multiview Learning

Multimodal tasks such as speech processing are natural applications for techniques from the area of multiview learning. Multiview learning is a set of techniques that leverage relationships between views (here, audio and video) to learn better models than would be learned from each view separately or from a simple concatenation of the two views. Multiview learning techniques typically produce models that can be used even if only one of the two views is available at test time. This property is quite useful, since it may be possible to collect AV training data while only having access to either audio or video at test time. One of the views may be completely missing, or may be very corrupted by noise, and it is desirable to be able to handle this situation gracefully.

The application of multiview learning techniques to AV speech processing is still in its infancy, with most work focusing on small data sets and simple tasks. In this section, we review the work done thus far in this category, as well as some natural extensions that may be fruitful avenues for future work.

1) Cotraining: One classic technique from multiview learning is cotraining [158], a semisupervised approach for learning a pair of classifiers, one for each view. In cotraining, there is a small amount of "seed" labeled data, which is used to learn an initial pair of classifiers, and a large amount of unlabeled data. The two classifiers then alternate: 1) labeling unlabeled data points on which they are most confident; and 2) retraining the classifiers. The motivation is that the multiple views are leveraged to label the unlabeled data and therefore to effectively increase the amount of training data and improve performance.

An approach related to cotraining is developed by Christoudias *et al.* [159] which is dubbed coadaptation. In this approach, an initial pair of classifiers are used to label data from a new speaker or domain, and the most confident labels are kept and used as a seed set for applying cotraining. They apply this idea to train AV viseme classifiers, where models for a new speaker can be trained without embarking on a lengthy annotation effort. While cotraining is used where little labeled but a lot of unlabeled data are available for a single domain or scenario, coadaptation is beneficial for the cases that sufficient labeled data from a domain, e.g., a set of speakers or environmental conditions, is available, but there is no labeled data for a new domain or scenario, e.g., new speaker.

2) Multiview Feature Learning: Multiview learning can also be used to learn improved representations, or features, by taking advantage of relationships between the views. In the case of AV speech processing, it is of course possible to use any combination of standard acoustic features and image features. However, it might be possible to improve on these standard features. Multiview techniques for feature learning typically take advantage of the fact that the sources of noise in the two views (or, more generally, nuisance parameters) are independent or at least uncorrelated. For example, the acoustic view may include background noise while the video may include lighting variations. Therefore, by looking for features that are in some sense common to the two views, multiview feature learning techniques can eliminate or reduce such noise. In addition, if the audio and video views can be represented in a truly common feature space, this makes it possible to directly compare acoustic and visual signals for cross-modal retrieval or for training on one modality and testing on another.

One typical approach for multiview feature learning is to use canonical correlation analysis (CCA) to learn transformations of each view [160], [161]. In particular, CCA finds pairs of projections, one for each view, such that the projected features are as highly correlated as possible. Theoretical results (e.g., [162]) show that CCA projections can improve class separation under certain conditions, such as uncorrelated noise in the two views. This is experimentally demonstrated in [162] by clustering audio or video frames from AV speech recordings into speaker clusters; they find that clustering CCA-based features greatly improves the speaker cluster quality, and makes it more robust to noise, over clustering in the original acoustic or visual space. The same CCA-projected features are exploited by Livescu and Stoehr [163] to improve speaker recognition in noise. AV speaker identification is improved in [164], [165] by combining visual (lip) features, audio features, and correlated audio-lip features discovered via CCA. Using CCA, they also find the best time shift for synchronizing the audio and video relative to each other, which also helps to boost identification performance.

CCA has been extended to the case of nonlinear projections via kernels [166] and deep neural networks [167], but to our knowledge nonlinear CCA has not yet been applied for AV speech processing. On the other hand, nonlinear feature learning approaches with other objectives have recently been developed and used for AV speech, typically using deep networks. For example, as already mentioned, improved representations for audio and/or video are suggested by Nagim *et al.* [115] using deep autoencoders with various structures, which learn to reconstruct both audio and video from both inputs together or from video alone, and use the learned representations to classify spoken digits/letters given only video or both audio and video, as described in the previous section. They find that the learned representations do better than the original features and than single-modality autoencoders. Also, by applying CCA to a hidden layer of the learned audio/video autoencoders, they can obtain further improvements. In addition, they are able to learn a joint representation such that they train a classifier using one modality data and test it using the other modality. The results in [115] are further improved in [168] using deep Boltzmann machines with a similar structure. The deep Boltzmann machines, unlike autoencoders, learn a generative model that can explicitly generate data from a missing modality.

3) Measuring Audiovisual Asynchrony: The idea of using cross-modal correlation has been applied beyond multiview feature learning, to detecting and measuring AV synchrony or asynchrony. For example, in [169] and [170], the correlation/canonical correlation between audio and video signals is exploited as a measure of AV synchrony. Similar measures (most successfully, pixelwise Gaussian mutual information) are used in [7] to locate the speaker in a video and identify the active speaker in a pair. In [171], audio and video signals are mapped through the single-layer perceptrons trained to maximize the mutual information between their outputs, and use the resulting mappings to localize a speaker as well as to enhance the speech of a desired speaker in the case of multiple simultaneous speakers.

VI. CONCLUSION

We conclude the paper by summarizing our views on where AV fusion stands and where it is probably heading. After an analysis of recent publications it could probably be argued that the research area has not really progressed much in the recent past. This is not to imply that the published results are not worthwhile but that although the main ideas have been very successful, it seems that they have not been pursued that much after their initial successes. Despite all the successful work on some of the challenges we addressed in this paper, i.e., stream weighting and asynchrony, there is still much to be done on these topics, in the sense that it is very hard to model reliability well and handle asynchrony properly. There has not been much discriminative structured modeling done for AV (structured SVMs, CRFs, etc.) and we expect that the various graphical models that have been used for asynchrony should benefit from that.

DL will undoubtedly improve AV fusion performance, as it has in every other area it has touched. It has only started to be used for AV but the already obtained initial results are highly promising and encouraging. Another possible future change is that multimodal work might start to become agnostic to what the specific modalities are. DL is having this effect in some areas, where basic domain-specific work is being replaced by deep networks that learn from the input signal. This does not mean domain knowledge is not needed, but maybe multimodal applications will start to care less about what the modalities are as a result of this trend.

Multiview learning for AV speech is emerging as a promising approach. Recent work has only begun to take advantage of multiview techniques. As mentioned above, certain techniques, such as nonlinear CCA, have yet to be applied to problems in this domain. In addition, there is much room to explore the use of multiview techniques for dealing with AV noise, beyond the very initial work described above. We believe that multiview learning is really barely off the ground and we expect that it will be a very fruitful area for future research.

As mentioned earlier, although there exist a number of AV databases, probably none of them has all desirable characteristics such as adequate data size, realistic variability, standard experiment settings, and evaluation measures. This limits progress in the filed. Maybe by taking better advantage of data that exist "in the wild", e.g., YouTube, the community might be helped to deal with realistic noisy data. Since most of these data are unlabeled, deep and multiview learning can be effective. With DL, representation of the data can be learned in an unsupervised fashion without the need to hand-engineer new sets of features. With cotraining, unlabeled data, on which the classifiers are most confident, can be labeled.

Finally, and to conclude, one can probably think that AV fusion is a very special area, but one thing that makes it particularly special is that there is so much AV data out there, e.g., YouTube videos, as opposed to other multimodal data. They will contribute to the booming and taking off of AV fusion that we all envisage. ■

Acknowledgment

The authors would like to thank Prof. K. Livescu for the fruitful discussions regarding the paper and Dr. L. Terry for making material available from his Ph.D. dissertation. They would also like to thank the anonymous reviewers for helping them improve the paper.

REFERENCES

[1] R. C. Luo and M. G. Kay, "Multisensor integration and fusion in intelligent systems," *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 5, pp. 901–931, Sep./Oct. 1989.

[2] S. Narayanan and A. Alwan, "Articulatory-acoustic models for fricative consonants," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 328–344, May 2000.

[3] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1, pp. 23–43, 1998.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[5] J. Hershey, H. Attias, N. Jojic, and T. Kristjansson, "Audio-visual graphical models for speech processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2004, vol. 5, p. V-649.

[6] A. Llagostera Casanovas and P. Vandergheynst, "Audio-visual object extraction using graph cuts," Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, Tech. Rep., 2012.

[7] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study *Image and Video Retrieval*. New York, NY, USA: Springer-Verlag, 2003, pp. 488–499.

[8] Y. Zhao, H. Wang, and Q. Ji, "Audio-visual Tibetan speech recognition based on a deep dynamic Bayesian network for natural human robot interaction," *Int. J. Adv. Robot. Syst.*, vol. 9, no. 258, 2012, doi: 10.5772/54000.

[9] L. H. Terry, D. J. Shiell, and A. K. Katsaggelos, "Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1316–1319.

[10] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[11] L. Terry and A. K. Katsaggelos, "A phone-viseme dynamic Bayesian network for audio-visual automatic speech recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2008, doi: 10.1109/ICPR.2008.4761927.

[12] P. S. Aleksic and A. K. Katsaggelos, "Product HMMs for audio-visual continuous speech recognition using facial animation parameters," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2003, vol. 2, p. II-481.

[13] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1213–1227, 2002.

[14] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1274–1288, 1900.

[15] S. Lucey, S. Sridharan, and V. Chandran, "Improved speech recognition using adaptive audio-visual fusion via a stochastic secondary classifier," in *Proc. IEEE Int. Symp. Intell. Multimedia Video Speech Process.*, 2001, pp. 551–554.

[16] P. S. Aleksic and A. K. Katsaggelos, "Lip feature extraction and feature evaluation in the context of speech and speaker recognition *Visual Speech Recognition: Lip Segmentation and Mapping*. Hershey Park, PA, USA: IGI Global, 2009, pp. 39–69.

[17] D. J. Shiell, L. H. Terry, P. S. Aleksic, and A. K. Katsaggelos, "Audio-visual and visual-only speech and speaker recognition: Issues about theory, system design *Visual Speech Recognition: Lip Segmentation and Mapping*. Hershey Park, PA, USA: IGI Global, 2009, pp. 1–38.

[18] D. Dean and S. Sridharan, "Dynamic visual features for audio-visual speaker verification," *Comput. Speech Lang.*, vol. 24, no. 2, pp. 136–149, 2010.

[19] H. Bredin and G. Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *EURASIP J. Appl. Signal Process*, no. 1, pp. 179–179, 2007.

[20] S. Bengio, "Multimodal authentication using asynchronous HMMs," in *Audio-and Video-Based Biometric Person Authentication*. New York, NY, USA: Springer-Verlag, 2003, pp. 770–777.

[21] G. Jaffre and J. Pinquier, "Audio/video fusion: A preprocessing step for multimodal person identification," presented at the Int. Workshop MultiModal User Authentification, Toulouse, France, 2006.

[22] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.

[23] Y. Ivanov, T. Serre, and J. Bouvrie, "Error weighted classifier combination for multi-modal human identification," CSAIL, Tech. Rep., 2005.

[24] P. K. Atrey, M. S. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Multimedia Syst.*, vol. 12, no. 3, pp. 239–253, 2006.

[25] W. Adams *et al.*, "Semantic indexing of multimedia content using visual, audio, text cues," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 2, pp. 170–185, 1900.

[26] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 572–579.

[27] G. Iyengar and H. J. Nock, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 255–258.

[28] X. Zou and B. Bhanu, "Tracking humans using multi-modal fusion," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2005, pp. 4–11.

[29] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 828–836, Jul. 2003.

[30] F. Talantzis, A. Pnevmatikakis, and L. C. Polymenakos, "Real time audio-visual person tracking," in *Proc. IEEE Workshop Multimedia Signal Process.*, 2006, pp. 243–247.

[31] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 22–31, Jan. 2001.

[32] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 741–746.

[33] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *Proc. IEEE Int. Conf. Image Process.*, 2003, vol. 3, p. III-25.

[34] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. 7th Int. Conf. Multimodal Interfaces*, 2005, pp. 61–68.

[35] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[36] J. Cech *et al.*, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *Proc. IEEE Int. Conf. Humanoid Robots*, 2013, doi: 10.1109/HUMANOIDS.2013.7029977.

[37] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005, pp. 118–121.

[38] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2000, vol. 3, pp. 1589–1592.

[39] A. P. Loh, F. Guan, and S. S. Ge, "Motion estimation using audio and video fusion," in *Proc. IEEE Control Autom. Robot. Vis. Conf.*, 2004, vol. 3, pp. 1569–1574.

[40] M. Barnard *et al.*, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 864–880, Apr. 2014.

[41] S. Essid and G. Richard, "Fusion of multimodal information in music content analysis," *Multimodal Music Process.*, vol. 3, pp. 37–52, 2012.

[42] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic Bayesian networks," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 25–36, Jan. 2007.

[43] A. Metallinou *et al.*, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 184–198, Apr.–Jun. 2012.

[44] A. Konar and A. Chakraborty, *Emotion Recognition: A Pattern Analysis Approach*. New York, NY, USA: Wiley, 2014.

[45] M. Glodek *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Affective Comput. Intell. Interaction*, 2011, pp. 359–368.

[46] G. Iyengar, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2003, vol. 1, p. I-329.

[47] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 548–555.

[48] A. Corradini, M. Mehta, N. O. Bernsen, J. Martin, and S. Abrilian, "Multimodal input fusion in human-computer interaction," *Nato Sci. Ser., Comput. Syst. Sci.*, vol. 198, p. 223, 2005.

[49] H. Holzapfel, K. Nickel, and R. Stiefelhagen, "Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures," in *Proc. 6th ACM Int. Conf. Multimodal Interfaces*, 2004, pp. 175–182.

[50] L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, no. 2, pp. 187–208, 2003.

[51] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, Nov. 2000.

[52] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," in *Proc. Sensor Signal Process. Defence*, 2011, pp. 1–5.

[53] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[54] Q. Liu, A. Aubrey, and W. Wang, "Interference reduction in reverberant speech separation with visual voice activity detection," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1610–1623, Oct. 2014.

[55] Q. Liu *et al.*, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5520–5535, Nov. 2013.

[56] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, 2014.

[57] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[58] D. Lahat, T. Adalı, and C. Jutten, "Challenges in multimodal data fusion," in *Proc. Eur. Signal Process. Conf.*, 2014, pp. 101–105.

[59] B. Rivet, W. Wang, S. Naqvi, and J. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.

[60] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[61] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues Vis. Audio-Vis. Speech Process.*, vol. 22, p. 23, 2004.

[62] P. Maragos, A. Potamianos, and P. Gros, *Multimodal Processing and Interaction: Audio, Video, Text*, vol. 33. New York, NY, USA: Springer Science & Business Media, 2008.

[63] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, 2014, Art. ID. e12.

[64] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Anal. Appl.*, vol. 1, no. 1, pp. 18–27, 1998.

[65] C. Chen, *Pattern Recognition and Artificial Intelligence*. New York, NY, USA: Academic, 1976.

[66] A. Waibel and K. Lee, *Readings in Speech Recognition*. San Mateo, CA, USA: Morgan Kaufmann, 1990.

[67] A. Bundy and L. Wallen, "Linear predictive coding," in *Catalogue of Artificial Intelligence Tools*, A. Bundy and L. Wallen, Eds. Berlin, Germany: Springer-Verlag, 1984.

[68] S. Bahaadini, A. Asaei, D. Imseng, and H. Bourlard, "Posterior-based sparse representation for automatic speech recognition," in *Proc. Interspeech*, 2014, pp. 2454–2458.

[69] C. Sui, R. Togneri, S. Haque, and M. Bennamoun, "Discrimination comparison between audio and visual features," in *Conf. Record 46th Asilomar Conf. Signals Syst. Comput.*, 2012, pp. 1609–1612.

[70] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[71] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: A case study on affect detection," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 3–10.

[72] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th*

Annu. ACM Int. Conf. Multimedia, 2005, pp. 399–402.

[73] Z. Wu, L. Cai, and H. Meng, "Multi-level fusion of audio and visual features for speaker identification *Advances in Biometrics*. Berlin, Germany: Springer-Verlag, 2005, pp. 493–499.

[74] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 771–780, p. 1612, 1999.

[75] Performance evaluation of tracking and surveillance, 2009. [Online]. Available: http://www.cvg.rdg.ac.uk/PETS2009/

[76] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking *Machine Learning for Multimodal Interaction*. New York, NY, USA: Springer-Verlag, 2005, pp. 182–195.

[77] TREC Video Retrieval Evaluation (TRECVID). [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[78] S. Garcia-Salicetti *et al.*, "Biomet: A multimodal person authentication database including face, voice," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2688, J. Kittler and M. Nixon, Eds. Berlin, Germany: Springer-Verlag, 2003, pp. 845–853.

[79] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *Audio-and Video-Based Biometric Person Authentication*. Berlin, Germany: Springer-Verlag, 1997, pp. 403–409.

[80] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authenticat.*, 1999, vol. 964, pp. 965–966.

[81] C. Sanderson, "The VidTIMIT database," IDIAP, Tech. Rep., 2002.

[82] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

[83] C. Chibelushi, F. Deravi, and J. Mason, "Bt DAVID databasevinternal rep," Speech Image Process. Res. Grp, Dept. Electr. Electron. Eng., Univ. les Swansea, Swansea, U.K., 1996.

[84] N. A. Fox, B. A. OMullane, and R. B. Reilly, "The realistic multi-modal VALID database and visual speaker identification comparison experiments," in *Proc. 5th Int. Conf. Audio- Video-Based Biometric Person Authenticat.*, 2005.

[85] B. Lee *et al.*, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Interspeech*, 2004, pp. 2489–2492.

[86] K. Messer *et al.*, "The BANCA database and evaluation protocol," in *Audio-and Video-Based Biometric Person Authentication*. New York, NY, USA: Springer-Verlag, 2003, pp. 625–638.

[87] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2002, vol. 2, p. II-2017.

[88] A. K. Noulas and B. J. Kröse, "EM detection of common origin of multi-modal cues," in *Proc. 8th ACM Int. Conf. Multimodal Interfaces*, 2006, pp. 201–208.

[89] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 89–100, Sep. 2005.

[90] G. Lv, D. Jiang, R. Zhao, and Y. Hou, "Multi-stream asynchrony modeling for audio-visual speech recognition," in *Proc. 9th IEEE Int. Symp. Multimedia*, 2007, pp. 37–44.

[91] S. Lee and D. Yook, "Audio-to-visual conversion using hidden Markov models," in *PRICAI 2002: Trends in Artificial Intelligence*. New York, NY, USA: Springer-Verlag, 2002, pp. 563–570.

[92] C. G. Snoek and M. Worring, "A review on multimodal video indexing," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2002, vol. 2, pp. 21–24.

[93] M. Makkook, "A multimodal sensor fusion architecture for audio-visual speech recognition," 2007.

[94] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 303–306.

[95] I. McCowan *et al.*, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[96] S. Bengio, "Multimodal speech processing using asynchronous hidden Markov models," *Inf. Fusion*, vol. 5, no. 2, pp. 81–89, 2004.

[97] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 994–999.

[98] D. B. Dean, "Synchronous HMMs for audio-visual speech processing," Ph.D. dissertation, Queensland Univ. Technol., BrisbaneQld.Australia, 2008.

[99] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1852, Oct. 2007.

[100] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonom. Agents Multi-Agent Syst.*, vol. 20, no. 1, pp. 70–84, 2010.

[101] B. D. Anderson and J. B. Moore, *Optimal Filtering*. New York, NY, USA: Courier Dover, 2012.

[102] M. I. Ribeiro, "Kalman and extended Kalman filters: Concept, derivation and properties," *Inst. Syst. Robot.*, p. 43, 2004.

[103] D. Crisan and A. Doucet, "A survey of convergence results on particle filtering methods for practitioners," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 736–746, Mar. 2002.

[104] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 186–200, Feb. 2015.

[105] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, Aug. 2010.

[106] A. L. Casanovas, "Audio-visual fusion: New methods and applications," Ph.D. dissertation. EPFL, Lausanne, Switzerland, 2011.

[107] X. Shao and J. Barker, "Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment," *Speech Commun.*, vol. 50, no. 4, pp. 337–353, 2008.

[108] V. Estellers, M. Gurban, and J. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1145–1157, May 2012.

[109] A. H. Abdelaziz and D. Kolossa, "Dynamic stream weight estimation in coupled-HMM-based audio-visual speech recognition using multilayer perceptrons," in *Proc. Interspeech*, pp. 1144–1148, 2014.

[110] N. Tatbul, M. Buller, R. Hoyt, S. Mullen, and S. Zdonik, "Confidence-based data management for personal area sensor networks," in *Proc. 1st Int. ACM Workshop Data Manage. Sensor Netw./VLDB*, 2004, pp. 24–31.

[111] H. Hsu and S.-F. Chang, "Generative, discriminative, ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2004, vol. 2, pp. 1091–1094.

[112] K. Saenko and K. Livescu, "An asynchronous DBN for audio-visual speech recognition," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2006, pp. 154–157.

[113] L. H. Terry, K. Livescu, J. B. Pierrehumbert, and A. K. Katsaggelos, "Audio-visual anticipatory coarticulation modeling by human and machine," in *Proc. Interspeech*, 2010, pp. 2682–2685.

[114] L. Terry, "Audio-visual asynchrony modeling and analysis for speech alignment and recognition," Ph.D. dissertation, Northwestern Univ., Evanston, IL, USA, 2011.

[115] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[116] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2014.

[117] P. S. Beddor and R. A. Krakow, "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2868–2887, 1999.

[118] P. S. Beddor, J. Harnsberger, and S. Lindemann, "Acoustic and perceptual characteristics of vowel-to-vowel coarticulation in Shona and English," *J. Phonetics*, vol. 30, pp. 591–627, 2002.

[119] P. A. Keating, "Underspecification in phonetics," *Phonology*, vol. 5, no. 2, pp. 275–292, 1988.

[120] W. L. Henke, "Dynamic articulatory model of speech production using computer simulation," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1966.

[121] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1994, vol. 2, p. II-669.

[122] C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces," *J. Commun.*, vol. 43, pp. 32–40, 1992.

[123] J. S. Garofolo, "Overcoming barriers to progress in multimodal fusion research," in *Proc. AAAI Fall Symp. Multimedia Inf. Extraction*, 2008, pp. 3–4.

[124] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech

recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[125] A. V. Nefian *et al.*, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2002, vol. 2, p. II-2013.

[126] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Multistream articulatory feature-based models for visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1700–1707, Sep. 2009.

[127] B. Li, E. Chang, and Y. Wu, "Discovery of a perceptual distance function for measuring image similarity," *Multimedia Syst.*, vol. 8, no. 6, pp. 512–522, 2003.

[128] M. Campbell *et al.*, "IBM research TRECVID-2007 video retrieval system," in *Proc. TRECVID*, 2007.

[129] P. Aleksic and A. Katsaggelos, "An audio-visual person identification and verification system using FAPS as visual features," in *Proc. ACM Workshop Multimodal User Authenticat.*, 2003, pp. 80–84.

[130] A. Kanak, E. Erzin, Y. Yemez, and A. M. Tekalp, "Joint audio-video processing for biometric speaker identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2003, vol. 2, p. II-377.

[131] G. Chetty and M. Wagner, "Audio-visual multimodal fusion for biometric person authentication and liveness verification," in *Proc. NICTA-HCSNet Multimodal User Interaction Workshop*, 2006, vol. 57, pp. 17–24.

[132] H. Xu and T.-S. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 44–67, 2006.

[133] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, pp. 1267–1270.

[134] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1998, pp. 3733–3736.

[135] Y. L. Chow, "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the n-best algorithm," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 1990, vol. 2, pp. 701–704.

[136] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2002, pp. I-853–I-856.

[137] A. Garg, G. Potamianos, C. Neti, and T. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proc. Int. Conf. Multimedia Expo.*, 2003, vol. 3, p. III-605-8.

[138] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, no. 1, pp. 1260–1273, 2002.

[139] E. Marcheret, V. Libal, and G. Potamianos, "Dynamic stream weight modeling for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, vol. 4, pp. IV-945–IV-948.

[140] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition," in *Proc. 10th Int. Conf. Multimodal Interfaces*, 2008, pp. 237–240.

[141] F. Mihelic and J. Zibert, *Speech Recognition: Technologies and Applications, I-Tech*, 2008. [Online]. Available: https://books.google.ch/books?id=9o0zQwAACAAJ.

[142] R. Rajavel and P. S. Sathidevi, "Adaptive reliability measure and optimum integration weight for decision fusion audio-visual speech recognition," *J. Signal Process. Syst.*, vol. 68, no. 1, pp. 83–93, 2012.

[143] L. Terry, D. Shiell, and A. Katsaggelos, "Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1316–1319.

[144] X. Shao and J. Barker, "Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment," *Speech Commun.*, vol. 50, no. 4, pp. 337–353, 2008.

[145] V. Estellers, M. Gurban, and J. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1145–1157, May 2012.

[146] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, no. 1, pp. 1274–1288, 2002.

[147] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "A new EM estimation of dynamic stream weights for coupled-HMM-based audio-visual ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, doi: 10.1109/ICASSP.2014.6853853.

[148] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 175–184, Feb. 2014.

[149] L. Deng and D. Yu, *Deep Learning: Methods and Applications*. New York, NY, USA: Now Publishers, 2014.

[150] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robot. Autonom. Syst.*, vol. 62, no. 6, pp. 721–736, 2014.

[151] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2014, pp. 2141–2149.

[152] H. P. Martínez and G. N. Yannakakis, "Deep multimodal fusion: Combining discrete events and continuous signals," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 34–41.

[153] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2012, pp. 2222–2230.

[154] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2008, pp. 873–880.

[155] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 3687–3691.

[156] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7596–7599.

[157] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2014, pp. 754–757.

[158] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. ACM Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[159] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, "Co-adaptation of audio-visual speech and gesture classifiers," in *Proc. 8th Int. ACM Conf. Multimodal Interfaces*, 2006, pp. 84–91.

[160] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.

[161] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[162] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.

[163] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2009, pp. 82–86.

[164] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, doi: 10.1109/ICASSP.2006.1660095.

[165] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Jul. 2007.

[166] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, Oct. 2000.

[167] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[168] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2012, pp. 2222–2230.

[169] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*. Cambridge, MA, USA: MIT Press, 2000.

[170] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, pp. 814–820.

[171] J. W. Fisher, III, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, pp. 772–778.

ABOUT THE AUTHORS

Aggelos K. Katsaggelos (Fellow, IEEE) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, USA, where he is currently a Professor holder of the AT&T chair. He was previously the holder of the Ameritech Chair of Information Technology (1997–2003). He is also the Director of the Motorola Center for Seamless Communications, a member of the Academic Staff, NorthShore University Health System, an affiliated faculty at the Department of Linguistics and he has an appointment with the Argonne National Laboratory. He has published extensively in the areas of multimedia signal processing and communications (over 230 journal papers, 500 conference papers, and 40 book chapters), and he is the holder of 25 international patents. He is the coauthor of *Rate-Distortion Based Video Compression* (Norwell, MA, USA: Kluwer, 1997), *Super-Resolution for Images and Video* (San Rafael, CA, USA: Claypool, 2007), *Joint Source-Channel Video Transmission* (San Rafael, CA, USA: Claypool, 2007), and *Machine Learning, Optimization, and Sparsity* (Cambridge, U.K.: Cambridge Univ. Press, forthcoming). He has supervised 50 Ph.D. dissertations so far.

Prof. Katsaggelos was the Editor-in-Chief of the IEEE SIGNAL PROCESSING MAGAZINE (1997–2002), a BOG Member of the IEEE Signal Processing Society (1999–2001), a member of the Publication Board of the PROCEEDINGS OF THE IEEE (2003–2007), and he is currently a Member of the Award Board of the IEEE Signal Processing Society. He is a Fellow of the International Society for Optics and Photonics (SPIE; 2009) and the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE ICME Paper Award (2006), an IEEE ICIP Paper Award (2007), an ISPA Paper Award (2009), and a EUSIPCO paper award (2013). He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008).

Sara Bahaadini received the Diploma from the National Organization for Development of Exceptional Talents with the top rank in mathematics and physics, the B.Sc. degree in computer science and engineering from Shiraz University, Shiraz, Iran, in 2008, and the M.Sc. degree in computer science and engineering from Sharif University of Technology, Tehran, Iran, in 2011. Currently, she is working toward the Ph.D. degree at the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA.

Rafael Molina was born in 1957. He received the degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from the University of Granada, Granada, Spain, in 1979 and 1983, respectively.

He became Professor of Computer Science and Artificial Intelligence at the University of Granada, Granada, Spain, in 2000. He is the former Dean of the Computer Engineering School at the University of Granada (1992–2002) and Head of the Computer Science and Artificial Intelligence department of the University of Granada (2005–2007). His research interest focuses mainly on using Bayesian modeling and inference in problems like image restoration (applications to astronomy and medicine), superresolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, and classification.

Dr. Molina serves as an Associate Editor of *Applied Signal Processing* (2005–2007); the IEEE TRANSACTIONS ON IMAGE PROCESSING (2010-present); and *Progress in Artificial Intelligence* (2011-present); and an Area Editor of *Digital Signal Processing* (2011-present). He is the recipient of an IEEE International Conference on Image Processing Paper Award (2007) and an ISPA Best Paper Award (2009). He is a coauthor of a paper awarded the runner-up prize at Reception for early-stage researchers at the House of Commons.