Recent Advances in the Automatic Recognition of Audio-Visual Speech

Gerasimos Potamianos, *Member, IEEE,* Chalapathy Neti, *Member, IEEE,* Guillaume Gravier, Ashutosh Garg, *Student Member, IEEE,* and Andrew W. Senior, *Senior Member, IEEE*

(Invited Paper)

***Abstract*— Visual speech information from the speaker's mouth region has been successfully shown to improve noise robustness of automatic speech recognizers, thus promising to extend their usability in the human computer interface. In this paper, we review the main components of audio-visual automatic speech recognition and present novel contributions in two main areas: First, the visual front end design, based on a cascade of linear image transforms of an appropriate video region-of-interest, and subsequently, audio-visual speech integration. On the latter topic, we discuss new work on feature and decision fusion combination, the modeling of audio-visual speech asynchrony, and incorporating modality reliability estimates to the bimodal recognition process. We also briefly touch upon the issue of audio-visual adaptation. We apply our algorithms to three multi-subject bimodal databases, ranging from small- to large-vocabulary recognition tasks, recorded in both visually controlled and challenging environments. Our experiments demonstrate that the visual modality improves automatic speech recognition over all conditions and data considered, though less so for visually challenging environments and large vocabulary tasks.**

***Index Terms*— Audio-visual speech recognition, speechreading, face tracking, visual feature extraction, audio-visual fusion, hidden Markov models, multi-stream HMM, product HMM, stream reliability, adaptation, multimedia databases.**

I. Introduction

AUTOMATIC *speech recognition* (ASR) is viewed as an integral part of future human-computer interfaces, that are envisioned to use speech, among other means, to achieve natural, pervasive, and ubiquitous computing. However, although ASR has witnessed significant progress in well-defined applications like dictation and medium vocabulary transaction processing tasks in relatively controlled environments, its performance has yet to reach the level required for speech to become a truly pervasive user interface. Indeed, even in "clean" acoustic environments, state of the art ASR system performance lags human speech perception by up to an order of magnitude [1]. Moreover, its lack of robustness to channel and environment noise continues to be a major hindrance [2], [3]. Clearly, non-traditional approaches, that use sources of information orthogonal to the audio input, are needed to achieve ASR performance closer to the human speech perception level, and robust enough to be deployable in field applications. *Visual speech* constitutes a promising such source, clearly not affected by acoustic noise.

Both human speech production and perception are *bimodal* in nature [4], [5]. The visual modality benefit to speech intelligibility in noise has been quantified as far back as in 1954 [6]. Furthermore, bimodal integration of audio and visual stimuli in perceiving speech has been demonstrated by the McGurk effect [7]: When, for example, the spoken sound /ga/ is superimposed on the video of a person uttering /ba/, most people perceive the speaker as uttering the sound /da/. In addition, visual speech is of particular importance to the hearing impaired: Mouth movement is known to play an important role in both sign language and simultaneous communication between the deaf [8]. The hearing impaired speechread well, and possibly better than the general population [9].

There are three key reasons why vision benefits human speech perception [10]: It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The latter is due to the partial visibility of articulators, such as the tongue, teeth, and lips. Place of articulation information can help disambiguate, for example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) from the nasal alveolar /n/ [11]. All three pairs are highly confusable on basis of acoustics alone. In addition, jaw and lower face muscle movement is correlated to the produced acoustics [12–14], and its visibility has been demonstrated to enhance human speech perception [15], [16].

The above facts have motivated significant interest in automatic recognition of visual speech, formally known as *automatic lipreading*, or *speechreading* [4]. Work in this field aims at improving ASR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to *audio-visual automatic speech recognition* (AV-ASR) systems. Compared to audio-only speech recognition, AV-ASR introduces new challenging tasks, that are highlighted in the block diagram of Fig. 1: First, in addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker's face. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by extraction of suitable visual features. In contrast to audio-only recognizers, there

Manuscript received December 20, 2002; revised March 13, 2003.

G. Potamianos, C. Neti, and A. W. Senior are with the Human Language Technologies Department, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA. Emails: {gpotam,cneti,aws}@us.ibm.com

G. Gravier is with CNRS at IRISA/INRIA Rennes, 35042 Rennes Cedex, France. Email: ggravier@irisa.fr

A. Garg is with the IBM Almaden Research Center, San Jose, CA 95120, USA. Email: ashutosh@us.ibm.com

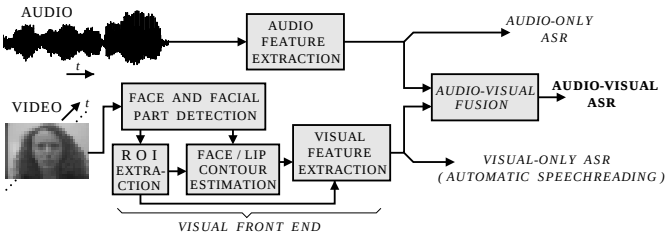

Fig. 1. The main processing blocks of an audio-visual automatic speech recognizer. The visual front end design and the audio-visual fusion modules introduce additional challenges, compared to traditional audio-only ASR.

are now *two* streams of features available for recognition, one for each modality. The combination of the audio and visual streams should ensure that the resulting system performance is better than the best of the two single modality recognizers, and hopefully, significantly outperform it. Both issues, namely the *visual front end design* and *audio-visual fusion*, constitute difficult problems [17] and have generated much research work by the scientific community.

The first automatic speechreading system was reported in 1984 by Petajan [18]. Given the video of the speaker's face, and by using simple image thresholding, he was able to extract binary (black and white) mouth images, and subsequently, mouth height, width, perimeter, and area, as visual speech features. He then developed a visual-only recognizer based on dynamic time warping [19] to rescore the best two choices of the output of the baseline audio-only system. His method improved ASR for a single-speaker, isolated word recognition task on a 100-word vocabulary that included digits and letters.

Since then, over a hundred articles have concentrated on AV-ASR, with the vast majority appearing during the last decade. The reported systems differ in three main aspects [17]: The visual front end design, the audio-visual integration strategy, and the speech recognition method used. Unfortunately, the diverse algorithms suggested in the literature are difficult to compare, as they are rarely tested on a common audio-visual database. Nevertheless, the majority of systems outperform audio-only ASR over a wide range of conditions. Such improvements have been typically demonstrated on databases of small duration, and, in most cases, limited to a very small number of speakers and to small vocabulary tasks [17], [20]. Common tasks typically include recognition of non-sense words [21], [22], isolated words [18], [23–29], connected digits [30], [31], letters [30], or of closed-set sentences [32], mostly in English, but also in French [21], [33], German [34], [35], and Japanese [36], among others. Recently however, significant improvements have also been demonstrated for *large vocabulary continuous speech recognition* (LVCSR) [37], as well as cases of speech degraded due to speech impairment [38] or Lombard effects [28]. These facts, when coupled with the diminishing cost of quality video capturing systems, make automatic speechreading tractable for achieving robust ASR in certain scenarios and tasks [17].

In this paper, we provide a brief overview of the main techniques for AV-ASR that have been developed over the past two decades, with emphasis on the algorithms investigated in our own research. In addition, we present recent improvements in the visual front end of our automatic speechreading system and a number of new contributions in the area of audio-visual integration. Furthermore, we benchmark the discussed methods on three data sets, reporting AV-ASR of both small and large vocabularies, as well as of data recorded in visually challenging environments.

In more detail, Section II of the paper concentrates on the visual front end, first summarizing relevant work in the literature, and subsequently discussing its three main blocks in our system. It also reports recent improvements in the extraction and normalization of the visual region of interest. Section III presents issues in visual speech modeling, that are relevant to audio-visual fusion, and also serves to introduce the notation used in the remainder of the paper. Section IV is devoted to an overview of audio-visual fusion, considering three classes of algorithms, i.e., feature, decision, and hybrid fusion. In particular, it introduces a novel technique within the last category, and also discusses the issue of audio-visual asynchrony modeling. Section V concentrates on a very important aspect of decision fusion based AV-ASR, namely modeling the reliability of the audio and visual stream information. A number of local stream reliability indicators are considered, and a function that maps their values to appropriate decision fusion parameters is introduced. Section VI is devoted to audio-visual adaptation, necessary for improving recognition performance on datasets of small duration, or for particular subjects. Section VII discusses our audio-visual databases, and Section VIII reports experimental results on them. Finally, Section IX concludes the paper with a summary and a brief discussion on the current state and open problems in AV-ASR.

II. The Visual Front End

The first major issue in audio-visual ASR is the visual front end design (see also Fig. 1). Over the past 25 years, a number of such designs have been proposed in the literature. Given the video input they produce visual speech features that in general fit in one of the following three categories: *Appearance* based features, *shape* based ones, or combination of both [17].

Appearance features assume that all video pixels within a *region-of-interest* (ROI) are informative about the spoken utterance. To allow speech classification, they consider mostly linear transforms of the ROI pixel values, resulting in feature vectors of reduced dimensionality that contain most relevant speech information [23], [26], [29], [34], [37], [39–44]. In contrast, shape based feature extraction assumes that most speechreading information is contained in the contours of the speaker's lips, or more generally in the face contours, e.g., jaw and cheek shape, in addition to the lips [37]. Within this category belong geometric type features, such as mouth height, width, and area [18], [21], [25], [27], [28], [31], [32], [45–47], Fourier and image moment descriptors of the lip contours [27], [48], statistical models of shape, such as active shape models [37], [49], or other parameters of lip-tracking models [41], [50–52]. Finally, features from both categories can be concatenated into a joint shape and appearance vector [26], [41], [53], or a joint statistical model can be learned on such vectors, as is the case of the active appearance model [54], used for speechreading in [37].

Clearly, a number of video pre-processing steps are required before the above mentioned visual feature extraction techniques can commence. One such step is face and facial part detection, followed by ROI extraction (see also Fig. 1). Of course, the pre-processing depends on the type of visual data provided to the AV-ASR system, being unnecessary for example, when a properly head-mounted video camera is used [55]. In case shape-based visual features are to be extracted, the additional step of lip and possibly face shape estimation is required. Some popular methods that are used in this task are snakes [56], templates [57], and active shape and appearance models [54], [58]. Alternative image processing and statistical image segmentation techniques can also be employed [59–61], possibly making use of the image color information, especially if the speaker's lips are marked with lipstick [21], [47], [61].

Our AV-ASR system extracts solely appearance based features, and operates on full face video with no artificial face markings. As a result, both face detection and ROI extraction are required. All stages of the adopted visual front end algorithm are described below.

Fig. 2. Face, facial part detection, and ROI extraction for example video frames of two subjects recorded at a controlled studio environment (upper row) and in a typical office (lower row). The following are depicted for each set (left to right): Original frame with eleven detected facial parts super-imposed; face-area enhanced frame; size-normalized mouth-only ROI (upper); and size-, rotation-, and lighting-normalized, enlarged ROI (lower).

A. Face and Facial Part Detection

Face and facial part detection has attracted significant interest in the literature [59], [62–64], and it constitutes a difficult problem, especially in cases where the background, head pose, and lighting are varying. Many reported systems use traditional image processing techniques, such as color segmentation, edge detection, image thresholding, template matching, or motion information [59], while others consider a statistical modeling approach, employing neural networks for example [62]. Our system belongs to the second category, using the algorithm reported in [64].

In more detail, given a video frame, face detection is first performed by searching for face candidates that contain a relatively high proportion of skin-tone pixels over an image "pyramid" of possible locations and scales. Each face candidate is size-normalized to a chosen template size (here, an 11×11 square), and its greyscale pixel values are placed into a 121-dimensional face candidate vector. Every vector is given a score based on a two-class (face versus non-face) Fisher linear discriminant [65], as well as its "distance from face space" (DFFS), i.e., the face vector projection error onto a lower, 40-dimensional space, obtained by means of *principal components analysis* (PCA) [66]. All candidate regions exceeding a threshold score are considered as faces. Among such faces at neighboring scales and locations, the one achieving the maximum score is returned by the algorithm as a detected face [64].

Once a face has been detected, an ensemble of facial feature detectors are used to estimate the locations of 26 facial features, including the lip corners and centers (eleven such facial features are marked on the frames of Fig. 2). Each feature location is determined by using a score combination of prior feature location statistics, linear discriminant, and "distance from feature space" (similar to the DFFS discussed above), based on the chosen feature template size.

A training step is required to estimate the Fisher discriminant and PCA eigenvectors for face detection and facial feature estimation, as well as the facial feature location statistics. Such training requires a number of frames manually annotated with the faces and their visible features (see Section VIII).

B. Region of Interest

In most automatic speechreading systems, the ROI is a square containing the image pixels of the speaker's mouth region, following possible normalization, for example, scale, rotation, and lighting compensation, or windowing with an appropriate mask [29]. The ROI can also include larger parts of the lower face, such as the jaw and cheeks [67], or even the entire face [37]. Often, it can be a three-dimensional rectangle, containing adjacent frame rectangular ROIs, in an effort to capture dynamic speech information [40], [42]. In other systems, the ROI corresponds to a number of image profiles vertical to the lip contour [26], or is a disc around the mouth center [39]. Concatenation of the ROI pixel greyscale [26], [34], [39], [40] or color values [41] results into a high-dimensional ROI vector that captures visual speech information.

In our system, the ROI contains the grey-scale values of a 64×64 size square region, centered around the mouth center, and normalized for variations in mouth scale. Both mouth center and scale parameters are obtained after appropriate temporal smoothing of their frame-level estimates, provided by the face detection algorithm. Originally, the ROI was limited to the mouth region alone [44], but subsequent experiments demonstrated that enlarging it to contain the jaw and cheeks was beneficial [67]. Such enlarged ROIs are used in AV-ASR results reported in Section VIII on the controlled studio environment data. However, recent work on more challenging visual domains, for example on data recorded using a cheap PC video camera in varying lighting conditions, demonstrated that additional ROI processing steps are necessary for improved robustness of the visual front end. We have thus added histogram equalization of the face-region image followed by low-pass filtering, and ROI compensation for head rotation and head height-to-width ratio recording variations. ROI examples

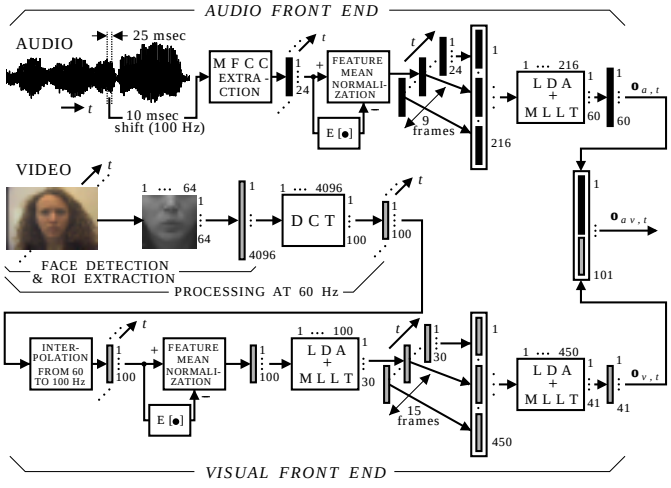

Fig. 3. Block diagram of the front end for AV-ASR. The algorithm generates time-synchronous 60-dimensional audio feature vectors, $\mathbf{o}_{a,t}$, and 41-dimensional visual observations, $\mathbf{o}_{v,t}$, both at a 100 Hz rate.

using the original [44] and newly added processing steps are depicted in Fig. 2.

C. Visual Features and Post-Processing

The dimensionality of the extracted ROI vector ($4096 = 64 \times 64$, in our case) is too large to allow successful statistical modeling [66] of speech classes, by means of a *hidden Markov model* (HMM) for example [19]. The required dimensionality reduction is typically achieved by traditional linear transforms, borrowed from the image compression and pattern classification literatures [65], [66], [68], [69], in the hope that they will preserve most relevant to speechreading information. Most commonly applied transforms are PCA [26], [34], [39–42], [49], [70], [71], the *discrete cosine transform* (DCT) [29], [36], [37], [39], [40], [43], discrete wavelet transform [40], Hadamard and Haar transforms [43], and a *linear discriminant analysis* (LDA) based data projection [39].

The resulting visual features are often post-processed to facilitate and improve AV-ASR. For example, audio and visual feature stream synchrony is required in a number of algorithms for audio-visual fusion, although the modality feature extraction rates typically differ. This can be easily resolved by simple element-wise linear interpolation of the visual features to the audio frame rate. Variations between speakers and recording conditions can be somewhat remedied by visual *feature mean normalization* (FMN), i.e., by subtraction of the vector mean over each utterance [72]. In addition, improved recognition by capturing important visual speech dynamics [73] can be accomplished by augmenting "static" visual features with their first- and second-order temporal derivatives [19], [72]. Finally, when using HMMs with diagonal covariances for ASR, a feature vector rotation by means of a *maximum likelihood linear transform* (MLLT) can be beneficial [74].

Our visual front end system uses a 2-dimensional, separable, fast DCT, applied to the ROI vector, and retains 100 transform coefficients at specified locations of large DCT energy, as computed over a set of training video sequences. The DCT feature vectors are extracted for each de-interleaved half-frame (field) of the video, available at 60 Hz, and are immediately up-sampled to the audio feature rate, 100 Hz, by means of linear interpolation, a process followed by mean normalization (FMN). To further reduce their dimensionality, an intra-frame LDA/MLLT is applied, resulting in a 30-dimensional "static" feature vector. To capture dynamic speech information, 15 consecutive feature vectors (centered at the current frame) are concatenated, followed by an inter-frame LDA/MLLT for dimensionality reduction and improved statistical modeling. The resulting "dynamic" visual features are of length 41. The visual front end block diagram is given in Fig. 3. It is depicted in parallel with the audio front end processing, which produces "static" 24-dimensional *mel frequency cepstral coefficients* (MFCCs) at 100 Hz [19], [72], followed by FMN, LDA, and MLLT, thus providing 60-dimensional "dynamic" audio features [75].

III. Visual Speech Modeling For ASR

Once features become available from the visual front end, one can proceed with automatic recognition of the spoken utterance by means of the video only information (automatic speechreading), or combine them to synchronously extracted acoustic features for audio-visual ASR (see also Fig. 1). The first scenario is primarily useful in benchmarking the performance of visual feature extraction algorithms, with visual-only ASR results typically reported on small-vocabulary tasks [23], [24], [27–30], [34], [38–40], [43], [49], [60], [71], [76–81]. Visual speech modeling is required in this process, its two central aspects being the choice of speech classes, that are assumed to generate the observed features, and the statistical modeling of this generation process. Both issues are important, as they are also embedded into the design of audio-visual fusion (see Section IV), and are discussed next.

A. Speech Classes

The basic unit that describes how speech conveys linguistic information is the *phoneme* [19]. However, since only a small part of the vocal tract is visible, not every pair of these units can be disambiguated by the video information alone. Visually distinguishable units are called *visemes* [4], [5], [11], and consist of phoneme clusters that are derived by human speechreading studies, or are generated using statistical techniques [32], [82]. An example of a phoneme-to-viseme mapping is depicted in Table I [37].

In audio-only ASR, the hidden speech classes, estimated on the basis of the observed feature sequence, typically consist of *context-dependent* sub-phonetic units, that are obtained by decision tree based clustering of the possible phonetic contexts [19], [72]. For automatic speechreading, it seems appropriate to use sub-*visemic* classes, obtained by decision tree clustering of visemic contexts on the basis of visual feature observations [37]. Naturally, visemic based speech classes are often considered in the literature [44], [79], [82]. However, having different speech classes in its audio- and visual-only components complicates audio-visual integration. Typically therefore, identical classes for both modalities are used. Here, such classes are defined over eleven-phone contexts (see Section VIII).

TABLE I
A 44 PHONEME TO 13 VISEME MAPPING OF THE HTK PHONE SET [72].

Viseme class	Phonemes in cluster
Silence	/sil/, /sp/
Lip-rounding based vowels	/ao/, /ah/, /aa/, /er/, /oy/, /aw/, /hh/ /uw/, /uh/, /ow/ /ae/, /eh/, /ey/, /ay/ /ih/, /iy/, /ax/
Alveolar-semivowels	/l/, /el/, /r/, /y/
Alveolar-fricatives	/s/, /z/
Alveolar	/t/, /d/, /n/, /en/
Palato-alveolar	/sh/, /zh/, /ch/, /jh/
Bilabial	/p/, /b/, /m/
Dental	/th/, /dh/
Labio-dental	/f/, /v/
Velar	/ng/, /k/, /g/, /w/

B. Speech Classifiers

A number of classification approaches are proposed in the literature for automatic speechreading, as well as audio-visual ASR. Among them: A simple weighted distance in visual feature space [18], artificial neural networks [34], [35], [39], [47], and support vector machines [79], used possibly in conjunction with dynamic time warping [18], [39] or HMMs [47], [79]. By far though, the most widely used classifiers are traditional HMMs that statistically model transitions between the speech classes and assume a class-dependent generative model for the observed features, similarly to HMMs in audio-only ASR [19], [72].

Let us denote the set of speech classes by $\mathcal{C}$, and the l_s-dimensional feature vector in stream s at time t by $\mathbf{o}_{s,t} \in \mathsf{R}^{l_s}$, where $s = v$ in the case of visual-only features. In generating a sequence of such vectors, the HMM assumes a sequence of hidden states that are sampled according to the *transition* probability parameter vector $\mathbf{a}_s = [\{ Pr[c' \,|\, c''], c', c'' \in \mathcal{C}\}]$. These states subsequently "emit" the observed features with class-conditional probabilities $P(\mathbf{o}_{s,t} \,|\, c)$, $c \in \mathcal{C}$. In the automatic speechreading literature, the latter are sometimes considered as discrete probability mass functions (after vector quantization of the feature space) [83], or non-Gaussian, parametric continuous densities [22]. However, in most cases, they are assumed to be *Gaussian mixture* densities of the form

$$P(\mathbf{o}_{s,t} \,|\, c) = \sum_{k=1}^{K_{s,c}} w_{s,c,k}\, \mathcal{N}_{l_s}(\mathbf{o}_{s,t}; \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}), \quad (1)$$

where the $K_{s,c}$ mixture weights $w_{s,c,k}$ are positive and add to one, and $\mathcal{N}_l(\mathbf{o}; \mathbf{m}, \mathbf{s})$ is the l-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix $\mathbf{s}$. The HMM parameter vector $\mathbf{p}_s = [\mathbf{a}_s, \mathbf{b}_s]$, where

$$\mathbf{b}_s = \big[\{[w_{s,c,k}, \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}]\}, k = 1,\ldots,K_{s,c}, c \in \mathcal{C}\big], \quad (2)$$

is typically estimated iteratively, using the *expectation-maximization* (EM) algorithm [84], as

$$\mathbf{p}_s^{(j+1)} = \arg\max_{\mathbf{p}} Q(\mathbf{p}_s^{(j)}, \mathbf{p} \,|\, \mathbf{O}_s), \quad j = 0, 1, \ldots . \quad (3)$$

In (3), $\mathbf{O}_s = \{\mathbf{o}_{s,t}, t \in \mathcal{T}\}$ consists of all feature vectors in training set $\mathcal{T}$, and $Q(\bullet, \bullet | \bullet)$ represents the EM algorithm *auxiliary function*, defined as in [19]. Alternatively, *discriminative training* methods can be used [85], [86].

In this paper, *single-stream* HMMs with *emission* probabilities (1), and trained as in (3), are exclusively used to model the two single-modality classifiers of interest (audio- and visual-only). Such models are used as the basis of all audio-visual integration techniques, discussed next.

IV. AUDIO-VISUAL INTEGRATION FOR ASR

As already mentioned in Section I, audio-visual integration constitutes a major research topic in AV-ASR, aiming at the combination of the two available speech informative streams into a bimodal classifier with superior performance to both audio- and visual-only recognition. Various information fusion algorithms have been considered for AV-ASR, differing both in their basic design, as well as in the terminology used [17], [21], [26], [30], [34], [37], [46], [78], [80], [82]. In this paper, we adopt their broad grouping into *feature fusion* and *decision fusion* methods. The first are based on training a single classifier (i.e., of the same form as the audio- and visual-only classifiers) on the concatenated vector of audio and visual features, or on any appropriate transformation of it [21], [37], [46]. In contrast, decision fusion algorithms utilize the two single-modality (audio- and visual-only) classifier outputs to recognize audio-visual speech. Typically, this is achieved by linearly combining the class-conditional observation log-likelihoods of the two classifiers into a joint audio-visual classification score, using appropriate weights that capture the reliability of each single-modality classifier, or data stream [17], [26], [30], [33], [37]. In addition to the above categories, there exist techniques that combine characteristics of both. Here, we introduce one such *hybrid fusion* method. The presentation of all techniques initially assumes an "*early*" temporal level of audio-visual integration, namely at the HMM state (see also Fig. 4). So-called "*asynchronous*" models of fusion are discussed at the end of the section. The latter are relevant to decision and hybrid fusion only.

A. Feature Fusion

Audio-visual feature fusion techniques include: Plain feature *concatenation* [21], feature weighting [46], [78], both also known as *direct identification* fusion [46], hierarchical *discriminant* feature extraction [37], as well as the *dominant* and *motor* recording fusion [46]. The latter seek a data-to-data mapping of either the visual features into the audio space, or of both modality features to a new common space, followed by linear combination of the resulting features. Audio feature *enhancement* on the basis of either visual input [13], [87], or concatenated audio-visual features [88–90] falls also within this category of fusion. In this paper, we briefly review two feature fusion methods.

Given time-synchronous audio and visual feature vectors $\mathbf{o}_{a,t}$ and $\mathbf{o}_{v,t}$ respectively, concatenative feature fusion considers

$$\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}, \mathbf{o}_{v,t}] \in \mathsf{R}^{l_{av}}, \text{ where } l_{av} = l_a + l_v, \quad (4)$$

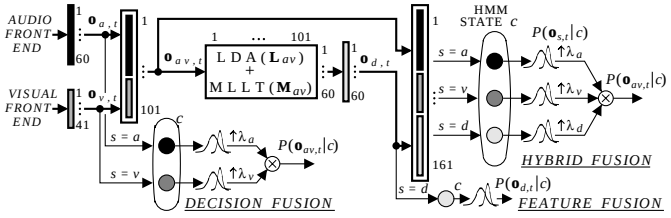

Fig. 4. Representative techniques of the three fusion categories, considered in this paper for AV-ASR. Feature vector dimensionalities are also depicted.

as the joint audio-visual observation of interest, modeled by a single-stream HMM, as in (1). In practice, l_{av} can be large, causing inadequate modeling in (1) due to the curse of dimensionality [66] and insufficient data.

Discriminant feature fusion aims to remedy this, by applying an LDA projection on the concatenated vector $\mathbf{o}_{av,t}$. Such projection results in a lower dimensional representation of (4), while seeking the best discrimination among the speech classes of interest. In [37], LDA is followed by an MLLT rotation of the feature vector to improve statistical data modeling by means of Gaussian mixture emission probability densities with diagonal covariances, as in (1). The transformed audio-visual feature vector then becomes

$$\mathbf{o}_{d,t} \;=\; \mathbf{o}_{av,t}\,\mathbf{L}_{av}\,\mathbf{M}_{av} \;\in\; \mathsf{R}^{l_d}\;, \qquad (5)$$

where $\mathbf{L}_{av}$ denotes the LDA matrix of size size $l_{av}\times l_d$, and $\mathbf{M}_{av}$ is the MLLT matrix of size $l_d\times l_d$. In this work, $\mathbf{o}_{d,t}$ is assumed to be of the same dimension as the audio observation, i.e., $l_d = l_a$. Both concatenative and discriminant feature fusion techniques are implementable in most existing ASR systems with minor changes, due to their use of single-stream HMMs.

B. Decision Fusion

Although many feature fusion techniques result in improved ASR over audio-only performance [37], they cannot explicitly model the reliability of each modality. Such modeling is extremely important, due to the varying speech information content of the audio and visual streams. The decision fusion framework, on the other hand, provides a mechanism for capturing these reliabilities, by borrowing from classifier combination theory, an active area of research with many applications [91–93].

Various classifier combination techniques have been considered for audio-visual ASR, including for example a cascade of fusion modules, some of which possibly using only rank-order classifier information about the speech classes of interest [18], [82]. However, by far the most commonly used decision fusion techniques belong to the paradigm of audio- and visual-only classifier combination using a parallel architecture, adaptive combination weights, and class score level information. These methods derive the most likely speech class or word sequence by linearly combining the log-likelihoods of the two single-modality classifier decisions, using appropriate weights [21], [26], [27], [30], [37], [45], [46]. This corresponds to the adaptive product rule in the likelihood domain [94], and it is also known as the *separate identification* model for audio-visual fusion [46], [82].

In the case where single-stream HMMs, with the same set of speech classes (states), are used for both audio- and visual-only classification, as in (1), this type of likelihood combination can be considered at a frame (HMM state) level, and modeled by means of the *multi-stream* HMM. Such an HMM has first been introduced for audio-only ASR [72], [95], [96], and, subsequently, its two-stream variant was deemed suitable for AV-ASR [26], [30], [36], [37], [45], [97]. For a two-stream HMM, the state-dependent emission of the audio-visual observation vector $\mathbf{o}_{av,t}$ is governed (see also (1) and (4)) by

$$P\,(\,\mathbf{o}_{av,t}\,|\,c\,) \;=\; P\,(\,\mathbf{o}_{a,t}\,|\,c\,)^{\lambda_{a,c,t}}\,P\,(\,\mathbf{o}_{v,t}\,|\,c\,)^{\lambda_{v,c,t}}\;, \qquad (6)$$

for all HMM states $c\in\mathcal{C}$. Notice that (6) corresponds to a linear combination in the log-likelihood domain, however it does not represent a probability distribution in general, and will therefore be referred to as a "*score*". In (6), $\lambda_{s,c,t}$ denote the stream exponents (weights), that are non-negative, and model stream reliability as a function of modality s, HMM state $c\in\mathcal{C}$, and utterance frame (time) t. In this paper, we also constrain the exponents to add up to one, and for the remainder of this section, we assume that they are set to global, modality-only dependent values, i.e., $\lambda_s \leftarrow \lambda_{s,c,t}$, for all c and t.

The multi-stream HMM parameters are (see also (1), (2), and (6))

$$\overline{\mathbf{p}}_{av} \;=\; [\,\mathbf{p}_{av}\,,\lambda_a\,,\lambda_v\,]\,,\;\; \text{where}\;\; \mathbf{p}_{av} \;=\; [\,\mathbf{a}_{av},\mathbf{b}_a,\mathbf{b}_v\,] \qquad (7)$$

consists of the HMM transition probabilities $\mathbf{a}_{av}$ and the emission probability parameters $\mathbf{b}_a$ and $\mathbf{b}_v$ of its single-stream components. The parameters of $\mathbf{p}_{av}$ can be estimated *separately* for each stream component using the EM algorithm, namely (3) for $s=a,v$, and subsequently, by possibly setting the joint HMM transition probability vector equal to the audio-one, i.e., $\mathbf{a}_{av}=\mathbf{a}_a$, or to the product of the transition probabilities of the two HMMs, i.e., $\mathbf{a}_{av}=\mathsf{diag}\,(\mathbf{a}_a^{\top}\mathbf{a}_v)$. The alternative is to *jointly* estimate parameters $\mathbf{p}_{av}$, in order to enforce state synchrony in training. In the latter scheme, the EM based parameter re-estimation becomes [72]

$$\mathbf{p}_{av}^{(j+1)} \;=\; \arg\max_{\mathbf{p}}\, Q\,(\,\overline{\mathbf{p}}_{av}^{(j)}\,,\,\mathbf{p}\,|\,\mathbf{O}_{av}\,) \qquad (8)$$

(see also (3)). The two approaches thus differ in the E-step of the EM algorithm. In both separate and joint HMM training, in addition to $\mathbf{p}_{av}$, the stream exponents λ_a and λ_v need to be obtained. The issue is deferred to Section V.

C. Hybrid Fusion

Certain feature fusion techniques, for example discriminant fusion by means of (5), outperform audio- and visual-only ASR (see [37] and Section VIII). It therefore seems natural to consider (5) as a stream in multi-stream based decision integration (6), thus combining feature and decision fusion within the framework of the latter. In this paper, we propose two such hybrid approaches, by generalizing the two-stream HMM of (6) into

$$P\,(\,\mathbf{o}_{av,t}\,|\,c\,) \;=\; \prod_{s\,\in\,\mathcal{S}} P\,(\,\mathbf{o}_{s,t}\,|\,c\,)^{\,\lambda_s}\;, \qquad (9)$$

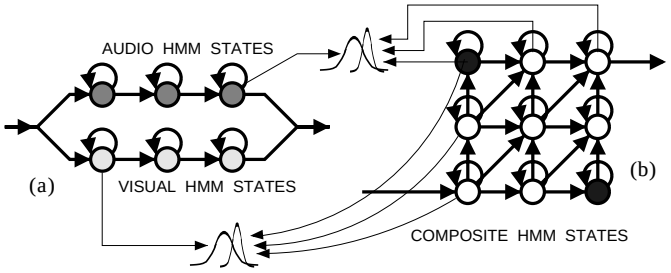

Fig. 5. (a) Phone-synchronous (state-asynchronous) two-stream HMM with three states per phone and modality. (b) Its equivalent product (composite) HMM; black circles denote states that are removed when limiting the degree of within-phone allowed asynchrony to one state. The single-stream emission probabilities are tied for states along the same row (column) to the corresponding audio (visual) state probabilities of form (1), according to (10).

where $S = \{a, v, d\}$, or $S = \{a, d\}$. In the first case, we obtain a three-stream HMM, with the added stream of discriminant features $\mathbf{o}_{d,t}$ of (5). In the second case, we retain the two-stream HMM, however after replacing the less speech-informative visual stream with its superior stream (5). As discussed above, stream exponents are constrained by $\lambda_s \geq 0$ and $\Sigma_{s \in \mathcal{S}} \lambda_s = 1$, whereas parameter estimation of the HMM components can be performed separately, or jointly. A schematic representation of the two hybrid fusion approaches (9) is depicted in Fig. 4, together with all previously presented fusion algorithms.

D. Audio-Visual Asynchrony in Fusion

In our presentation of decision and hybrid fusion, we have assumed the "*early*" temporal level of HMM states for combining the stream likelihoods of interest (see (6) and (9)). In ASR however, sequences of classes (HMM states or words) need to be estimated, therefore coarser levels for combining stream likelihoods can also be envisioned. One such "*late*" level of integration can be the utterance end, where typically a number of N-best hypotheses (or all vocabulary words, in case of isolated word recognition) are rescored by the stream log-likelihoods, independently computed over the entire utterance. An example of late fusion is the discriminative model combination technique [98], applied for AV-ASR in [99]. Alternatively, the phone, syllable, or word boundary can provide an "*intermediate*" level of integration. Such a scheme is typically implemented by means of the *product* HMM [100], or the *coupled* HMM [101], and is discussed next. Notice that both late and intermediate integration permit asynchrony between the HMM state sequences of the streams of interest, thus providing the means to model the actual audio and visual signal asynchrony, observed in practice to be up to the order of 100ms [34], [102].

The product HMM is a generalization of the state-synchronous multi-stream HMM (9) that combines the stream log-likelihoods at an intermediate level, here assumed to be the phone. The resulting phone-synchronous product HMM allows its single-stream HMM components to be in asynchrony within each phone, forcing their synchrony at the phone boundaries

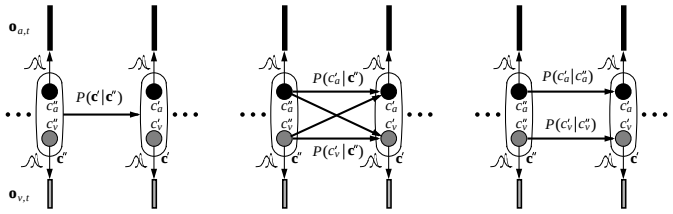

Fig. 6. Three possible schemes for transition probability modeling between the composite states of the product HMM. A two-stream model is depicted.

instead. It consists of *composite* states[1] $\mathbf{c} \in \mathcal{C}^{|\mathcal{S}|}$, with emission scores similar to (9), namely

$$P(\mathbf{o}_{av,t} \,|\, \mathbf{c}) = \prod_{s \in \mathcal{S}} P(\mathbf{o}_{s,t} \,|\, c_s)^{\lambda_s}, \qquad (10)$$

where $\mathbf{c} = \{c_s, s \in \mathcal{S}\}$. An example of such a model is depicted in Fig. 5, for the typical case of one audio and one visual stream, and three states per phone and stream. Notice that in (10), the stream components correspond to the emission probabilities of certain single-stream states, tied as demonstrated in Fig. 5. Therefore, compared to its corresponding state-synchronous multi-stream HMM, the product HMM utilizes the same number of mixture weight, mean, and variance parameters (see also (1) and (2)). On the other hand, additional transitions $\{P(\mathbf{c}'|\mathbf{c}''), \mathbf{c}', \mathbf{c}'' \in \mathcal{C}^{|\mathcal{S}|}\}$ between its composite states are required. Such probabilities are often factored as $P(\mathbf{c}'|\mathbf{c}'') = \Pi_{s \in \mathcal{S}} P(c'_s|\mathbf{c}'')$, in which case the resulting model is typically referred to in the literature as the coupled HMM [29], [80], [101], [103]. A further simplification of this factorization is sometimes employed, namely $P(\mathbf{c}'|\mathbf{c}'') = \Pi_{s \in \mathcal{S}} P(c'_s|c''_s)$, thus requiring the same number of parameters as the original state-synchronous multi-stream HMM. The latter factorization is employed in the product HMM of [104], as well as the factorial HMM of [29]. The three schemes are depicted in Fig. 6.

It is worth mentioning, that the product HMM allows the restriction of the degree of asynchrony between the two streams, by excluding certain composite states in the model topology (see also Fig. 5). In the extreme case, when only the states that lie in its "diagonal" are kept, the model becomes equivalent to (9).

Similar to the state-synchronous model, product HMM parameter estimation can be performed either separately for each stream component, or jointly for all streams at once. The latter scheme is preferable, as it consistently models asynchrony at both training and testing. Notice however that proper stream parameter tying is required, as lack of tying leads to models with significantly more parameters compared to (9) (exponential on the number of streams). This would not allow a fair comparison between (9) and (10), and could easily lead to poor statistical modeling due to insufficient data. On the other hand, and as experiments reported in [104] indicate, transition probability tying (i.e., factorization, discussed above) does not seem necessary. In the audio-visual ASR literature,

[1] For example, in the case of two streams (e.g., $\mathcal{S} = \{a, v\}$), the cardinality of $\mathcal{S}$ equals 2 ($|\mathcal{S}| = 2$), and the composite states are defined over the Cartesian product $\mathcal{S} \times \mathcal{S}$.

product (or, coupled) HMMs have been considered in some small-vocabulary recognition tasks [26], [28], [29], [70], [80], [105], where synchronization is sometimes enforced at the word level, and recently for LVCSR [37], [104]. However, with few exceptions [29], [104], proper parameter tying is usually not enforced.

V. Stream Reliability Modeling For AV-ASR

We now address the issue of stream exponent (weight) estimation, when combining likelihoods in the audio-visual decision and hybrid fusion models of the previous section (see (6), (9), and (10)). There, such exponents are set to constant stream-dependent values, to be computed for a particular audio-visual environment and database, based on the available training, or more often, held-out data. Due to the form of the emission scores, the stream exponents cannot be obtained by maximum likelihood estimation [30], [105]. Instead, discriminative training techniques are used.

Some of these methods seek to minimize a smooth function of the word classification error by the resulting audio-visual model on the data, and employ the generalized probabilistic descent algorithm [86] for stream exponent estimation [30], [36], [97], [106]. Other techniques use maximum mutual information training [85], as in [45]. A different approach minimizes the frame classification error, by using the maximum entropy criterion [106]. Alternatively, one can seek to directly minimize the word error rate of the resulting audio-visual ASR system on a held-out data set. In the case of two global exponents, constrained to add to a constant, the problem reduces to one-dimensional optimization of a non-smooth function, and can be solved using simple grid search [97], [106]. In the case where additional streams are present, as in (9), or when class-dependent stream exponents are desired, the problem becomes of higher dimension, and the downhill simplex method can be employed [107]. In general however, class dependency has not been demonstrated to be effective in AV-ASR [45], [106], with the exception of the late integration, discriminative model combination technique [99]. Therefore, in this paper, class dependence of global stream exponents is not considered. These global exponents are then simply estimated by grid search on a held-out set.

Although the use of such exponents has led to significant improvements in decision based fusion AV-ASR, it is not very suitable for practical systems. There, the quality of captured audio and visual data, and thus the speech information present in them, typically varies over time. For example, possible noise bursts, face occlusion, or other face tracking failures can greatly change the reliability of the affected stream. Utterance-level or even frame-level dependence of the stream exponents is clearly desirable. This can be achieved by first obtaining an estimate of the local environment conditions, using for example signal based approaches like the audio channel signal-to-noise ratio [21], [24], [27], [46], [47], [76], or the *voicing* index [108], as in [99]. Alternatively, statistical indicators of classifier confidence on the stream data, can be used. These indicators capture the reliability of each stream at a local, *frame* level, and have the advantage of not depending on the properties of the underlying signal. They are therefore used in this paper, as discussed next. Following their presentation, a method of exponent estimation based on these indicators is introduced.

A. Stream Reliability Indicators

A number of functions have been proposed in the literature as a means of assessing the reliability of the class information that is contained in an observation, assumed to be modeled by a particular classifier [21], [24], [33], [47], [109]. Following prior work [109], we select two reliability indicators for each stream of interest. Given the stream observation $\mathbf{o}_{s,t}$, both indicators utilize the class-conditional observation likelihoods of their N-best most likely generative classes, denoted by $c_{s,t,n} \in \mathcal{C}$, $n = 1, \ldots, N$. These are ranked according to descending values of $P(\mathbf{o}_{s,t}|c)$, $c \in \mathcal{C}$ (see also (1)).

The first reliability indicator is the N-best log-likelihood difference, defined as

$$\mathcal{L}_{s,t} \;=\; \frac{1}{N-1} \sum_{n=2}^{N} \log \frac{P\,(\,\mathbf{o}_{s,t}\,|\,c_{s,t,1}\,)}{P\,(\,\mathbf{o}_{s,t}\,|\,c_{s,t,n}\,)}\,, \tag{11}$$

for a stream $s \in \mathcal{S}$. This is chosen, since it is argued that the likelihood ratios between the first N classification decisions are informative about the class discrimination. The second selected reliability indicator is the N-best log-likelihood dispersion. This is defined as

$$\mathcal{D}_{s,t} \;=\; \frac{2}{N\,(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} \log \frac{P\,(\,\mathbf{o}_{s,t}\,|\,c_{s,t,n}\,)}{P\,(\,\mathbf{o}_{s,t}\,|\,c_{s,t,n'}\,)}\,. \tag{12}$$

The main advantage of (12) over (11) lies on the fact that (12) captures additional N-best class likelihood ratios, not present in (11). In our analysis, we choose N to be 5. As desired, both reliability indicators, averaged over the utterance, are well correlated to the utterance word error rate of their respective classifier. This is demonstrated in Section VIII.

B. Reliability Indicators For Stream Exponents

The next stage is to obtain a mapping of the chosen reliability indicators to the frame-dependent stream exponents. We use a sigmoid function for this purpose, due to the fact that it is monotonic, smooth, and bounded within zero and one. For simplicity, let us assume that only two streams $s \in \{a, v\}$ are available, thus requiring the estimation of exponent $\lambda_{a,t}$ and its derived $\lambda_{v,t} = 1 - \lambda_{a,t}$, on basis of the vector of the four selected reliability indicators, $\mathbf{d}_t = [\,d_{1,t}\,, d_{2,t}\,, d_{3,t}\,, d_{4,t}\,] = [\,\mathcal{L}_{a,t}\,, \mathcal{L}_{v,t}\,, \mathcal{D}_{a,t}\,, \mathcal{D}_{v,t}\,]$. Then, the mapping is defined as

$$\lambda_{a,t} \;=\; \frac{1}{1 \,+\, \exp\,(\,-\sum_{i=1}^{4} \mathrm{w}_i\, d_{i,t}\,)}\,, \tag{13}$$

where $\mathbf{w} = [\,\mathrm{w}_1\,, \mathrm{w}_2\,, \mathrm{w}_3\,, \mathrm{w}_4\,]$ is the vector of the sigmoid parameters. In the following, we propose two algorithms to estimate $\mathbf{w}$, given frame-level labeled audio-visual observations $\{(\mathbf{o}_{av,t}, c_t\,), t \in \mathcal{T}\,\}$, for a training set of time instants $\mathcal{T}$. Notice that the required labels $c_t \in \mathcal{C}$, $t \in \mathcal{T}$, can be obtained by a forced alignment of the training set utterances.

The first algorithm seeks *maximum conditional likelihood* (MCL) estimates of parameters $\mathbf{w}$ in (13), under the observation model (e.g., (6)). Given an audio-visual vector $\mathbf{o}_{av,t}$, we first represent the conditional likelihood of class $c \in \mathcal{C}$ by

$$P(c \mid \mathbf{o}_{av,t}) = \frac{P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}}{\sum_{c \in \mathcal{C}} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}}, \quad (14)$$

under the assumption of a uniform class prior $P(c)$ (see also (6)). We then seek parameters $\mathbf{w}$ of (13) as

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} \sum_{t \in \mathcal{T}} \log P(c_t \mid \mathbf{o}_{av,t}). \quad (15)$$

The above optimization problem can be solved iteratively, by

$$\mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} + \eta \sum_{t \in \mathcal{T}} \frac{\partial \log P(c_t|\mathbf{o}_{av,t})}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(j)}}, \quad (16)$$

for $j = 0,1,...$, where the gradient vector elements are

$$\frac{\partial \log P(c_t|\mathbf{o}_{av,t})}{\partial w_i} = \lambda_{a,t}(1-\lambda_{a,t}) d_{i,t} \Bigg[\log \frac{P(\mathbf{o}_{a,t}|c_t)}{P(\mathbf{o}_{v,t}|c_t)} - \frac{\sum_{c \in \mathcal{C}} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}} \log \frac{P(\mathbf{o}_{a,t}|c)}{P(\mathbf{o}_{v,t}|c)}}{\sum_{c \in \mathcal{C}} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}} \Bigg],$$

for $i = 1,2,3,4$ (see also (13)-(15)). In (16), we choose $\mathbf{w}^{(0)} = [1,1,1,1]$. Parameter η controls the convergence speed, and since (15) is not a convex optimization problem, it needs to be kept relatively small. In our experiments, when choosing $\eta = 0.01$, convergence is typically achieved within few tens of iterations.

The second technique adopted in this work for estimating the sigmoid parameters $\mathbf{w}$ is the *minimum classification error* (MCE) approach, seeking $\hat{\mathbf{w}}$ that maximizes the frame-level classification performance on the training data $\mathcal{T}$. We choose to solve this problem by a brute-force grid search over the four-dimensional parameter space. To simplify the search, we utilize the MCL parameter estimates, thus obtaining an approximate parameter dynamic range and limiting the search within it. Then, for each parameter vector value over the reduced grid, we compute the frame error. The weight assignment that results in the best performance is chosen as the output.

VI. Audio-Visual Adaptation

Adaptation techniques are traditionally used in audio-only ASR to improve system performance across speakers, tasks, or environments, when the available data in the condition of interest are insufficient for appropriate HMM training [110–114]. In the audio-visual ASR domain, adaptation is of great importance, especially since audio-visual corpora are scarce and their collection expensive, as discussed in Section VII.

A number of audio-only adaptation algorithms can be readily extended to audio-visual ASR. Given few adaptation data of a new subject or environment, some of these techniques re-estimate the parameters of an HMM that has been originally trained in a speaker-independent fashion and/or under different conditions. Two popular such methods are *maximum likelihood linear regression* (MLLR) [111] and *maximum-a-posteriori* (MAP) adaptation [110]. Other adaptation techniques proceed to transform the extracted features instead, so that they are better modeled by the available HMMs with no retraining [114].

In this paper, we briefly discuss MLLR and MAP audio-visual adaptation in the case of discriminant feature fusion by means of (1) and (5). Extensions to multi-stream HMM based fusion techniques (e.g., (6), (9), or (10)) can be easily considered, as in [115]. We also discuss a simple feature adaptation technique that transforms the LDA and MLLT matrices used in both modality front ends, as well as in discriminant feature fusion (5), in a manner akin to the MAP algorithm. Adaptation experiments utilizing these techniques, as well as their combination, are reported in Section VIII.

A. MLLR Adaptation

MLLR obtains a maximum likelihood estimate of a linear transformation of the HMM means, while leaving unchanged the covariance matrices, mixture weights, and transition probabilities of the original HMM parameter vector $\mathbf{p}^{(\mathrm{ORIG})}$. MLLR is most appropriate when a small amount of adaptation data is available (rapid adaptation).

Let $\mathcal{P}$ be a partition (obtained by K-means clustering [19], for example) of the set of all Gaussian mixture components of single-stream HMM (1), defined over the discriminant stream $s = d$, and let $p \in \mathcal{P}$ denote any member of this partition. Then, we seek MLLR adapted HMM parameters (see also (2))

$$\mathbf{p}_d^{(\mathrm{AD})} = \big[\mathbf{a}_d, \{ [w_{d,c,k}, \mathbf{m}_{d,c,k}^{(\mathrm{AD})}, \mathbf{s}_{d,c,k}], k = 1,...,K_{d,c}, c \in \mathcal{C} \} \big], \quad (17)$$

where the HMM means are linearly transformed as

$$\mathbf{m}_{d,c,k}^{(\mathrm{AD})} = [1, \mathbf{m}_{d,c,k}] \mathbf{W}_p, \quad (18)$$

where $(c,k) \in p$, and $\mathbf{W}_p$, $p = 1,...,|\mathcal{P}|$ are matrices of dimension $(l_d+1) \times l_d$. The transformation matrices are estimated on basis of the adaptation data $\mathbf{O}_d^{(\mathrm{AD})}$ [111], by means of the EM algorithm solving, similarly to (3),

$$\mathbf{p}^{(\mathrm{AD})} = \arg \max_{\mathbf{p} \text{ satisfy (17), (18)}} Q(\mathbf{p}^{(\mathrm{ORIG})}, \mathbf{p} \mid \mathbf{O}_d^{(\mathrm{AD})}).$$

Closed form solutions for the unknown matrices exist, since the HMM covariances are diagonal, due to (2) [111].

B. MAP Adaptation

In contrast to MLLR, MAP follows the Bayesian paradigm for estimating the adapted HMM parameters. These eventually converge to their EM-obtained estimates as the amount of adaptation data becomes large. Such convergence is slow however, thus MAP is not suitable for rapid adaptation. In practice, MAP is often used in conjunction with MLLR [112].

Our MAP implementation is similar to the *approximate* MAP adaptation algorithm (AMAP) [112]. AMAP interpolates the "counts" of the original training data, $\mathbf{O}_d^{(\mathrm{ORIG})}$, with the adaptation data. Equivalently, the training data $\mathbf{O}_d$ of the adapted HMM are

$$\mathbf{O}_d = [\mathbf{O}_d^{(\mathrm{ORIG})}, \underbrace{\mathbf{O}_d^{(\mathrm{AD})}, ..., \mathbf{O}_d^{(\mathrm{AD})}}_{m \text{ times}}]. \quad (19)$$

TABLE II
TYPICAL CORPORA USED IN THE LITERATURE FOR AV-ASR, GROUPED ACCORDING TO RECOGNITION TASK COMPLEXITY. DATABASE NAME, OR COLLECTING INSTITUTION, ASR TASK (ISOLATED (I) OR CONNECTED (C) WORD RECOGNITION IS SPECIFIED, WHEREVER APPROPRIATE), NUMBER OF SUBJECTS, LANGUAGE, AND SAMPLE REFERENCES ARE LISTED.

Name/Instit.	ASR Task	Sub.	Lan.	Sample references
ICP	vowel/conson.	1	FR	[21]
ICP	vowels	1	FR	[46]
Tulips1	I-4 digits	12	US	[23], [43], [49], [71], [77], [79]
M2VTS	I-digits	37	FR	[26], [97], [116]
XM2VTS	C-digits	295	UK	[103], [117]
UIUC	C-digits	100	US	[31]
CUAVE	I/C-digits	36	US	[27], [81]
IBM	C-digits	50	US	[38]
U.Karlsruhe	C-letters	6	D	[34], [39], [76]
U.LeMans	I-letters	2	FR	[33], [82]
U.Sheffield	I-letters	10	UK	[24]
AT&T	C-letters	49	US	[30], [40], [115]
UT.Austin	I-500 words	1	US	[83]
AMP-CMU	I-78 words	10	US	[28], [29], [60], [78], [80]
ATR	I-words	1	JP	[36], [105]
AV-TIMIT	150-sent.	1	US	[32]
Rockwell	100-C&C sent.	1	US	[25]
AV-ViaVoice	LVCSR(10.4k)	290	US	[37], [67], [99], [104], [106]

Then, (3) is used to estimate the adapted HMM parameters, $\mathbf{p}^{(\mathrm{AD})}$. HMM parameters of all mixtures are re-estimated, provided the adaptation data contain instances of the mixture component in question. Here, we use $m = 15$, in (19).

C. Front End Adaptation

In addition to updating HMM parameters, one may seek to adapt the front end, so as to better capture the speech information in the adaptation data. For the audio-visual front end of Section II and discriminant feature fusion (5), a simple form of *front end adaptation* is to re-estimate all appropriate LDA and MLLT matrices. Here, we simply compute such matrices using the combination of the original and adaptation data, given by (19). HMM parameters for the updated front end are then estimated using (3) on training data (19).

VII. AUDIO-VISUAL DATABASES

In contrast to the abundance of audio-only corpora, there exist only a few databases suitable for audio-visual ASR research. This is because the field is relatively young, but also due to the fact that audio-visual corpora pose additional challenges concerning database collection, storage, and distribution. A number of audio-visual datasets, commonly used in the literature, are listed in Table II. Notice that most are the product of efforts by few university groups or individual researchers with limited resources, and as a result, they suffer from one or more shortcomings [17], [20]: They contain a single or small number of subjects, affecting the generalizability of developed methods to the wider population; they typically have small duration, often resulting in undertrained statistical models, or non-significant performance differences between various proposed algorithms; and finally, they mostly address simple recognition tasks, such as small-vocabulary ASR of isolated or connected words.

Fig. 7. Example video frames of five subjects from each of the three audio-visual datasets considered in this paper for AV-ASR (top to bottom: studio-LVCSR, studio-DIGIT, office-DIGIT). The sides of the 704×480 size frames are cropped in the upper two rows. Clearly, the office-DIGIT database presents more challenges to the visual front end.

To help bridge the growing gap between audio-only and AV-ASR corpora, we have recently collected the IBM ViaVoice$^{\mathrm{TM}}$ audio-visual database, a large corpus suitable for speaker-independent audio-visual LVCSR (see also Table III). The corpus consists of full-face frontal video and audio of 290 subjects, uttering ViaVoice$^{\mathrm{TM}}$ training scripts, i.e., continuous read speech with mostly verbalized punctuation, dictation style. The data are collected using a teleprompter in a quiet studio environment. In more detail, the video is of a 704×480 pixel size, interlaced, captured in color at a rate of 30 Hz (60 fields per second are available at a resolution of 240 lines), and it is MPEG2 encoded at the relatively high compression ratio of about 50:1. Example video frames are depicted in Fig. 7. Notice that the lighting conditions, background, and head pose are quite uniform in the set, thus simplifying the visual front end processing. In addition to the video, high quality wideband audio is synchronously collected at a rate of 16 kHz and a *signal-to-noise ratio* (SNR) of 19.5 dB. The duration of the entire database is approximately 50 hours, from which about 44 hours (21k utterances) with a 10.4k vocabulary are used in the experiments reported in the next section.

To be able to study the visual modality benefit to a popular small-vocabulary ASR task, we have also collected a 50-subject connected digit database, in the same studio environment as the LVCSR data just described. This DIGIT corpus contains about 6.7k utterances (10 hrs) of 7- and 10-digit strings (both "zero" and "oh" are used).

Finally, we are interested in studying AV-ASR in domains and conditions that pose greater challenges to the visual front end processing, compared to the controlled studio environment of the previous sets. For this purpose, we have been collecting data in two visually challenging domains: The first set is recorded inside moving automobiles, where dramatic variations in lighting due to shadows are observed [118]. The second corpus is a DIGIT set (7- or 10-digit strings), collected in typical offices using a cheap camera that is connected via a

TABLE III

THE IBM AUDIO-VISUAL DATABASES USED IN OUR EXPERIMENTS. THEIR PARTITIONING INTO TRAINING, CHECK (HELD-OUT), ADAPTATION, AND TEST SETS IS DEPICTED (NUMBER OF UTTERANCES, DURATION (IN HOURS), AND NUMBER OF SUBJECTS ARE SHOWN FOR EACH SET). BOTH LARGE-VOCABULARY CONTINUOUS SPEECH (LVCSR) AND CONNECTED DIGIT (DIGIT) RECOGNITION ARE CONSIDERED FOR DATA RECORDED AT A STUDIO ENVIRONMENT. FOR THE LOW QUALITY OFFICE-DIGIT DATA, DUE TO THE LACK OF SUFFICIENT TRAINING DATA, ADAPTATION OF HMMS TRAINED ON THE STUDIO-DIGIT SET IS CONSIDERED.

Environ.	Task	Set	Utter.	Dur.	Sub.
Studio	LVCSR	Train	17111	34:55	239
		Check	2277	4:47	25
		Adapt	855	2:03	26
		Test	670	2:29	26
Studio	DIGIT	Train	5490	8:01	50
		Ch/Adapt	670	0:58	50
		Test	529	0:46	50
Office	DIGIT	Adapt	1007	1:15	10
		Check	116	0:09	10
		Test	200	0:15	10

USB-2.0 interface to a portable PC. In this paper, we discuss experiments on this second set, as a means of also showcasing audio-visual adaptation algorithms across datasets. Example video frames of this set are depicted in the third row of Fig. 7. Notice the variation in lighting conditions, background, and pixel ratio compared to the previous databases that have been collected in a studio-like environment. The frame rate and size are also inferior, as only 320×240 color pixels are now available at 30 Hz. As expected, all these factors pose challenges to the visual front end.

The details of all three databases, as well as their partitioning into various subsets used in our experimental framework, are given in Table III.

VIII. EXPERIMENTS

We now proceed to report a number of ASR experiments on the three databases of Table III, using the algorithms discussed in the previous sections. We first briefly introduce the experimental paradigm adopted, followed by a more detailed presentation of our results.

A. The Experimental Paradigm

For all single-stream recognition tasks considered, we use 3-state, left-to-right phone HMMs, with context-dependent subphonetic classes (states). These classes are obtained by means of decision trees that cluster contexts spanning up to 5 phones to each side of the current phone, in order to better model coarticulation and improve ASR performance. For both studio quality databases, the DIGIT and LVCSR decision trees are estimated using the clean audio of the corresponding database *training* set, by bootstrapping on a previously developed audio-only HMM (and its corresponding front end), which provides data class labels by forced alignment [75]. Subsequently, K-means clustering is used to estimate audio-only HMMs, that correspond to the newly developed trees. It is by bootstrapping on these models, that the parameters of all HMMs considered in this paper are estimated (on their required front ends). The total number of the resulting context-dependent HMM states are 159 for the DIGIT task (corresponding to 22 phones) and approximately 2.8k for LVCSR (for 52 phones). Note that all single-stream HMMs have identical number of Gaussian mixture components, namely about 3.2k and 47k for the DIGIT and LVCSR tasks, respectively. Since the amount of data available in the visually challenging office-DIGIT task does not suffice to properly train new decision trees and initial audio models, we just use the ones estimated on the studio-DIGIT data.

TABLE IV

VISUAL-ONLY WER, %, ON THE TEST SETS OF THE THREE DATABASES OF TABLE III. PER-SPEAKER, MLLR-ADAPTED PERFORMANCE IS ALSO SHOWN FOR THE TWO STUDIO SETS. FOR THE OFFICE-DIGIT SET, RESULTS USING THE IMPROVED VISUAL FRONT END OF FIG. 2 ARE DEPICTED AT THE RIGHT-MOST COLUMN (*).

Recognition mode	s-LVCSR	s-DIGIT	o-DIGIT	o-DIGIT *
Speaker-Independent	93.52	38.53	83.94	65.00
Multi-Speaker	——	23.58	71.12	35.00
Speaker-Adapted	82.51	16.77	——	——

Once decision trees and initial DIGIT and LVCSR audio HMMs are developed, we proceed to estimate the parameters of single-stream HMMs that model visual-only, as well as audio-only and audio-visual feature sequences at a number of audio channel conditions. Both the original clean database audio at approximately 19.5 dB SNR, as well as noisy conditions, where speech babble noise is artificially added at various SNRs, are considered. We use three EM algorithm iterations for training, with the E-step of the first iteration employing the initial audio-only HMM (for bootstrapping). The resulting models may be further adapted on the *adaptation* sets of Table III. Appropriate single-stream HMMs are also joined to form the decision and hybrid fusion models of Section IV, i.e., (6), (9), and (10), with the stream exponents set to global values, estimated on the *held-out* sets of Table III. Joint stream HMM training is also considered.

With the exception of the reliability modeling experiments (where the SNR level is not assumed known), all results on the studio-DIGIT and -LVCSR tasks are reported on recognition of matched *test* data (same SNR as in training). For the DIGIT task, decoding is based on a simple digit-word loop grammar (with unknown string length), whereas for LVCSR, a trigram language model is used. In both cases, a two-stage stack decoding algorithm is employed, that uses a fast match followed by a detailed match [119]. Unless otherwise noted, LVCSR results are speaker-independent, whereas DIGIT recognition is multi-speaker (due to the small number of subjects), as implied by Table III.

B. Visual-Only Recognition

Given our three datasets, the first task is to extract visual speech features from the available videos. To train the required projections and statistics for face detection and facial feature

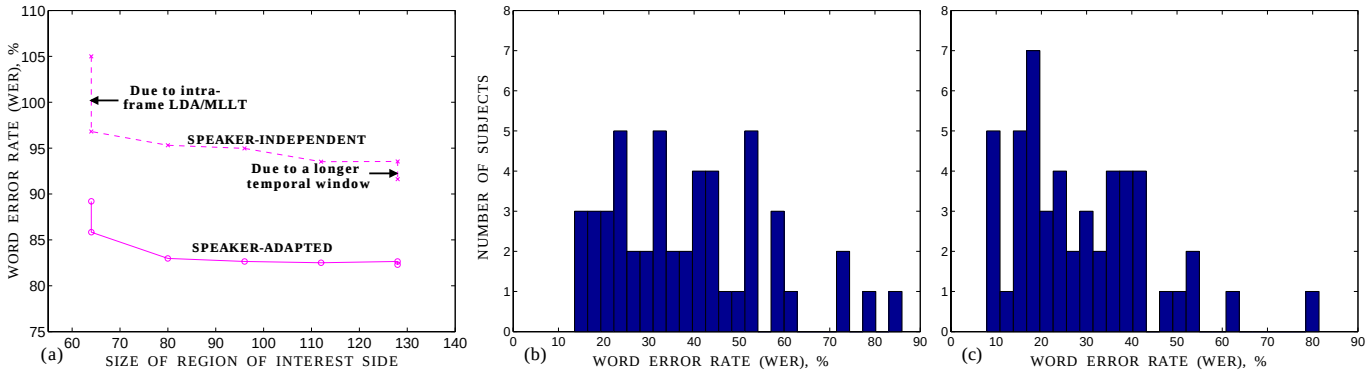

Fig. 8. Visual-only ASR on the studio-LVCSR and -DIGIT datasets. (a) Improvements in speaker-independent and speaker-adapted (by MLLR, per subject) visual-only LVCSR, due to the use of intra-frame LDA/MLLT, larger visual ROIs, and a larger temporal window for inter-frame LDA/MLLT. (b,c) WER histogram of the 50 subjects in the studio-DIGIT dataset, using visual-only HMMs trained in a speaker-independent or multi-speaker fashion.

localization, we annotate 26 facial features on approximately 4k video frames across the three databases (see also Fig. 2). The face detection accuracy is quite high for the two studio quality datasets (in excess of 99.5%), however it degrades to about 95% on the visually more challenging office-DIGIT set. After face detection is completed, ROI extraction and visual speech feature extraction follow, as described in Section II. Visual-only HMMs can subsequently be trained, providing a means to benchmark the visual front end performance.

The visual-only *word error rate* (WER), %, on all three sets is reported in Table IV. Clearly, for LVCSR, the visual features do provide speech information, albeit very weak [37]. DIGIT recognition on the other hand is a visually less confusable task, and the algorithm results in the multi-speaker WER of 23.6%. Per-speaker, MLLR based visual HMM adaptation significantly improves performance in both cases.

Recognition on the visually challenging office-DIGIT set is inferior to the studio-DIGIT task. Indeed, HMMs trained on the latter achieve a poor 83.9% WER on the former, with a small improvement to 71.1% after a cascade of front end / MLLR adaptation on the multi-speaker office-DIGIT adaptation set. These WERs decrease with proper compensation for lighting and head pose, using the improved ROI extraction algorithm of Section II (see also Fig. 2), and reach a 35.0% visual-only WER after adaptation. This still lags when compared to the 23.6% WER achieved on the studio-DIGIT task. Our experiments indicate that approximately half of this difference is due to the inferior quality of the captured image sequences (lower frame rate and resolution), whereas the remaining is most likely due to the visual challenges of the captured data (head pose, lighting, and background variation), that significantly affect face and mouth detection accuracy.

Two additional items of interest are showcased in Fig. 8. The first, in Fig. 8(a), demonstrates the effect of certain blocks of the visual front end of Fig. 3 to ASR performance. In more detail, three aspects of the algorithm are considered: Intra-frame LDA use (as opposed to just obtaining the 30 highest energy DCT coefficients), utilizing larger ROIs (containing successively larger parts of the lower face region), and the use of longer temporal windows for inter-frame LDA (performance for 15 vs. 21 feature frames is depicted). Notice that the resulting improvements due to larger ROIs and temporal windows are consistent with human bimodal speech perception studies [10], [16], [73]. The second point is demonstrated in Figs. 8(b,c), and concerns the variation in visual-only ASR performance across subjects. There, a WER histogram of the 50 studio-DIGIT dataset subjects is depicted, when using speaker-independent or multi-speaker visual-only HMMs. Clearly, there is a large variance in automatic speechreading performance, with some subjects resulting in about a tenth of the WER of others.

C. Audio-Visual ASR

Having demonstrated that the proposed visual front end provides speech informative features, our experiments now shift to quantifying its resulting benefit to ASR, when combined with the acoustic signal. We first apply all audio-visual integration strategies proposed in Section IV to the studio-DIGIT task. Representative techniques are subsequently considered on the studio-LVCSR data. Audio-visual ASR on the office-DIGIT set is deferred to Section VIII.E.

For both studio quality datasets, we consider acoustic conditions at a wide range of SNRs, as discussed in Section VIII.A, and we compare fusion strategies in terms of their resulting *effective SNR gain* in ASR. We measure this gain with reference to the audio-only WER at 10 dB, by considering the SNR value where the audio-visual WER equals the reference audio-only WER.

The performance of all integration algorithms on the studio-DIGIT set is summarized in Fig. 9. In more detail, we first compare AV-ASR by means of the two feature fusion methods of Section IV.A. As it becomes clear from Fig. 9, both concatenative and discriminative feature fusion significantly improve ASR performance at low SNRs, with the latter being somewhat superior, yielding an approximate 6 dB of effective SNR gain. For example, at -2.2 dB SNR, discriminant fusion based AV-ASR results in a 6.3% WER, representing a vast improvement over the audio-only WER of 19.8%. Notice however that feature fusion fails to alter performance at the high end of the SNR range considered. On the other hand, decision based audio-visual integration, by means of the state-synchronous two-stream HMM discussed in Section IV.B, consistently

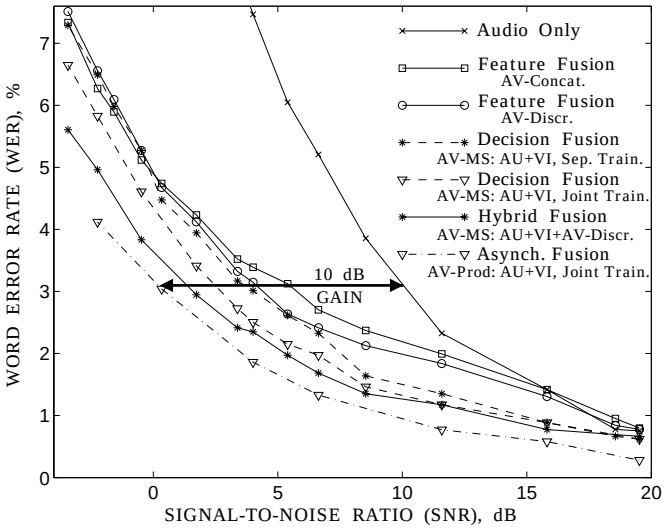

Fig. 9. Audio-only and audio-visual ASR on the studio-DIGIT database test set using a number of integration strategies, discussed in Section IV, namely feature fusion, the state-synchronous two-stream HMM (decision fusion), the state-synchronous three-stream HMM (hybrid fusion), and the state-asynchronous product HMM (asynchronous decision fusion). In all cases, WER, %, is depicted vs. audio channel SNR. The effective SNR gain using the product HMM is also shown, reported with reference to the audio-only WER at 10 dB. All HMMs are trained in matched noise conditions.

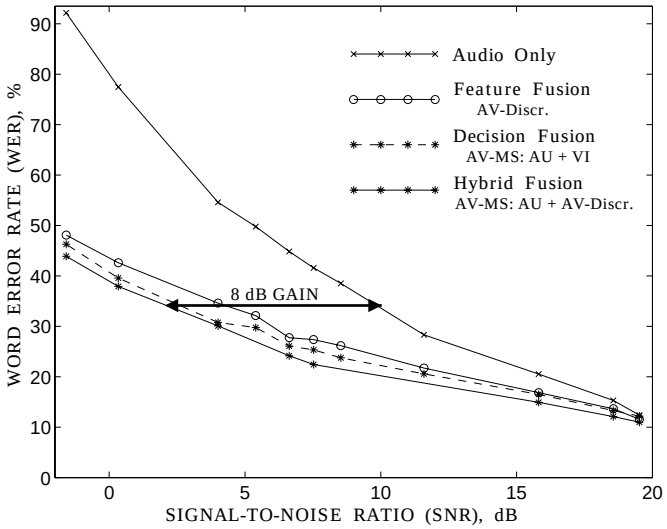

Fig. 10. Audio-only and audio-visual WER, %, on the studio-LVCSR test set using discriminant feature fusion, as well as two-stream HMMs for decision and hybrid fusion. All models are trained in matched noise conditions.

improves performance at all SNRs. In particular, joint stream training of the model seems clearly preferable, outperforming separate stream training and discriminant feature fusion, and yielding a 7.5 dB effective SNR gain. Further improvements can be obtained by using the hybrid fusion approach of Section IV.C that utilizes the discriminant audio-visual features as an additional stream within a three-stream HMM. This technique yields a 9 dB effective SNR gain. Finally, introducing state asynchrony in decision fusion results in further gains. A jointly trained product HMM achieves approximately a 10 dB SNR gain, thus exhibiting at 0 dB the performance of audio-only ASR at the much cleaner acoustic environment of 10 dB. Notice that at -2.2 dB SNR, the product HMM yields a 4.1% WER, which corresponds to a 35% improvement over discriminant feature fusion and 79% over audio-only ASR. But even more remarkably, for the original database audio at 19.5 dB, the audio-visual WER now stands at 0.28%, which represents a 63% WER reduction over the audio-only WER of 0.75% (see also Fig. 9). A large percentage of this gain is due to the joint estimation of all product HMM parameters with appropriate tying, since the composition of a product HMM by separately trained single-stream models achieves an inferior 0.40% WER.

For LVCSR, the performance of a number of the presented fusion techniques is summarized in Fig. 10. Similarly to the results on the studio-DIGIT set, hybrid fusion outperforms decision based integration, which in turn is superior to discriminant feature fusion, as well as audio-only ASR. For simplicity, a two-stream HMM is considered in hybrid fusion, where audio-visual discriminant features are used in place of the less informative visual-only stream. The resulting system achieves approximately an 8 dB effective SNR gain over audio-only ASR at 10 dB.

D. Audio-Visual Reliability Estimation

In the multi-stream HMM based fusion experiments reported above, all stream exponents are kept constant over an entire dataset and for a particular SNR level. In this Section, we investigate the benefit of frame-dependent exponents, estimated on basis of stream reliability indicators. To test the algorithm of Section V over varying stream reliability conditions, we consider the studio-DIGIT task at a mixture of SNR conditions. Babble noise is added to both test and held-out sets of the database, however the audio-only HMMs are trained on the original clean database audio.

We first argue that the selected indicators (11) and (12) do capture the reliability of the speech class information, available in the two streams of interest. Indeed, as depicted in Table V, the values of these indicators, averaged at the utterance level, are significantly correlated to the utterance WER using the corresponding single-stream HMM, with low correlation present across streams. In addition, as the audio channel becomes corrupted by increasing levels of noise, the speech information present in it is expected to degrade. Fig. 11 demonstrates that both $\mathcal{L}_{a,t}$ and $\mathcal{D}_{a,t}$ successfully convey such degradation, since they are monotonic on the SNR, similarly to the optimal global audio-stream exponent. The observations above argue favorably for using audio and visual stream reliability indicators in AV-ASR.

TABLE V
CORRELATION BETWEEN THE STREAM RELIABILITY INDICATORS (11) AND (12) AND THE AUDIO-ONLY AND VISUAL-ONLY WERS.

Reliability Indicator	Correlation with audio-only WER	Correlation with visual-only WER
$\mathcal{L}_a$	-0.7434	0.0183
$\mathcal{L}_v$	0.1041	-0.2191
$\mathcal{D}_a$	-0.7589	0.0126
$\mathcal{D}_v$	0.1014	-0.2066

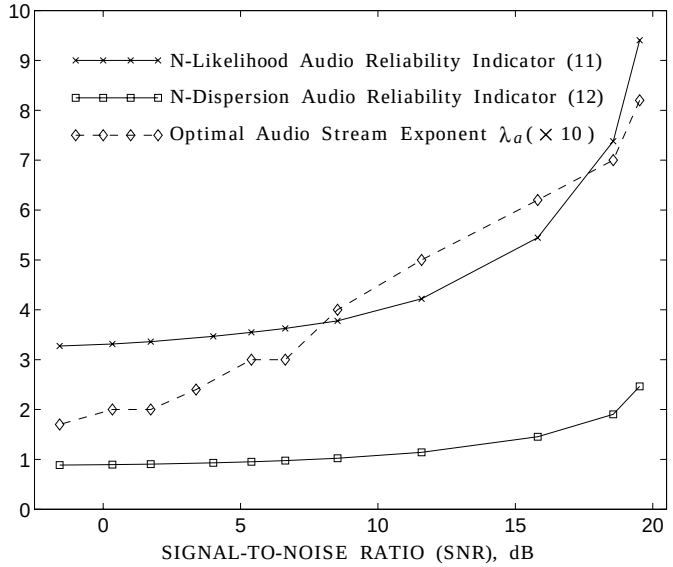

Fig. 11. Mean values of the audio reliability indicators $\mathcal{L}_{a,t}$ and $\mathcal{D}_{a,t}$, depicted as a function of the audio channel SNR. The corresponding "optimal" audio exponents λ_a in two-stream HMM based AV-ASR are also shown, scaled by a factor of 10. Results are reported on the studio-DIGIT database, with all HMMs trained in clean audio.

We now proceed to estimate stream exponents by means of the four selected reliability indicators and the sigmoid mapping of (13). The obtained results are summarized in Table VI, where we report both WER, as well as *frame* classification *error rate* (FER), assuming 22 DIGIT phone classes of interest. As an AV-ASR baseline, we first estimate a global audio exponent, constant over the entire dataset and all SNRs. The resulting two-stream HMM is labeled "AV-Global" in Table VI, and clearly outperforms audio-only ASR. We subsequently use the MCL and MCE algorithms to estimate the sigmoid parameters in (13). Both approaches further improve FER and WER, with the MCE based estimation resulting in a 17% relative WER reduction, over the use of global fusion weights. It is interesting to compare these WERs to the scenario that uses utterance-dependent exponents, and assumes a-priori knowledge of the SNR (a best case scenario for SNR-dependent exponent estimation). Such exponents are estimated on held-out data matched to the noise level, and are also depicted in Fig. 11 (scaled by a factor of 10). Even in this "cheating" case, the resulting 9.08% WER is worse than the WER achieved by frame-dependent exponents with MCE estimation of the sigmoid parameters. In conclusion, the proposed scheme of Section V is beneficial to AV-ASR.

E. Adaptation

In the final set of experiments, we apply the adaptation techniques of Section VI to the office-DIGIT set. As already indicated, the small amount of such data collected (see Table III) is not sufficient for HMM training, thus adaptation techniques are required to improve performance. A number of such methods are used for adapting audio-only, visual-only, and audio-visual HMMs using discriminant feature fusion. Two SNR conditions are considered, and the results are depicted in Table VII. To simplify experiments, the original HMMs are trained on the studio-DIGIT corpus at the clean audio condition, and adapted on the 15 dB office-DIGIT data.

TABLE VI

FRAME MISCLASSIFICATION ERROR RATE (FER) AND WER, %, FOR TWO-STREAM HMM BASED AV-ASR ON THE STUDIO-DIGIT TASK, USING GLOBAL VS. FRAME-DEPENDENT EXPONENTS ESTIMATED BY MEANS OF MAPPING (13). AUDIO-ONLY RECOGNITION RESULTS ARE ALSO DEPICTED. NOISE AT A NUMBER OF SNRS IS ADDED TO THE AUDIO UTTERANCES, BUT ALL HMMS ARE TRAINED ON THE ORIGINAL DATA.

Condition	FER	WER
Audio-Only	58.80	30.29
AV-Global	31.80	10.35
AV-Frame, MCL	31.53	10.13
AV-Frame, MCE	31.18	8.64

TABLE VII

SINGLE-MODALITY AND AUDIO-VISUAL ASR PERFORMANCE ON THE OFFICE-DIGIT TEST SET AT TWO AUDIO CHANNEL CONDITIONS (ORIGINAL DATA AT 15 DB SNR AND ARTIFICIALLY CORRUPTED AT 8 DB). HMMS TRAINED ON THE STUDIO-DIGIT DATASET ARE ADAPTED TO THE OFFICE DATA USING VARIOUS ALGORITHMS. ALL HMMS ARE TRAINED / ADAPTED ON THE ORIGINAL DATA AUDIO.

Method	Visual	AU-15dB	AV-15dB	AU-8dB	AV-8dB
Unadapted	65.00	8.71	7.18	35.00	31.06
MLLR	62.71	4.59	4.24	25.94	19.47
MAP	45.65	2.24	1.65	19.47	10.00
MAP+MLLR	45.18	2.00	1.71	19.24	10.41
FE	36.24	2.12	1.76	20.35	9.41
FE+MLLR	35.00	2.06	1.76	20.29	9.64

As it is clear from Table VII ("Unadapted" entries), the original studio-DIGIT HMMs perform poorly on the new set. This is due to the inferior quality of the office-DIGIT data. We then consider MLLR and MAP HMM adaptation. Notice that MAP performs better due to the relatively large adaptation set available. Applying MLLR after MAP typically improves results. Front end (FE) adaptation significantly helps visual-only recognition, improving for example performance from 45.6% to 36.2%, or from 45.1% to 35.0% when used in conjunction with MLLR. However, it does not seem to consistently help in neither audio-only nor AV-ASR. In conclusion, adaptation techniques can be successfully applied to bimodal recognition, and bridge performance gaps across datasets.

IX. SUMMARY AND DISCUSSION

In this paper, we provided a brief literature review of the basic techniques necessary in the automatic recognition of audio-visual speech. We mainly concentrated on the two most relevant issues to the design of audio-visual ASR systems, namely first, the visual front end that captures the speech information present in the video signal, and second, the integration of the extracted audio and visual features into the automatic speech recognizer. While presenting these, we focused in the algorithms used in our speechreading system, and we introduced a number of advances in both areas.

In particular, with respect to the visual front end design, we discussed in detail our algorithm for extracting appearance-

type visual features, based on a compressed representation of the image pixel values within a suitably defined region of interest. We demonstrated that it is beneficial for such a region to contain the jaw and cheeks, in addition to the mouth area. Furthermore, by properly compensating this region for simple lighting and head pose variations, we were able to significantly improve robustness to visual data recorded in challenging environments. On basis of our experiments, we concluded that the extracted visual features provide meaningful speech information, although quite weak compared to the traditional acoustic signal.

It is of course by combining the audio and visual features, that the benefit of the visual modality becomes apparent. In this work, we discussed a number of such fusion techniques, based on the popular hidden Markov model framework for speech recognition. We presented methods that integrate speech information at either the feature or the classification score level, and introduced a hybrid fusion algorithm that combines the benefits of both approaches. In addition, we discussed asynchrony modeling in audio-visual fusion, and we argued for the joint training of all properly tied parameters of the resulting model. In a first attempt to capture the varying reliability of the two streams of information, we investigated appropriate indicators of speech information content, and we proposed a trainable mapping from such indicators to time-dependent fusion parameters.

We applied these algorithms on three audio-visual corpora, spanning both small- and large-vocabulary recognition tasks, and containing data collected in visually "clean", as well as in challenging environments. Our best technique, utilizing the product hidden Markov model, resulted in an effective SNR gain of 10 dB for connected-digit recognition, though the best achieved gain on the large-vocabulary task was somewhat inferior, reaching approximately 8 dB. For connected-digit recognition of visually challenging data, our algorithms significantly improved performance compared to audio-only recognition, only after utilizing a number of adaptation techniques discussed in this work.

The paper clearly demonstrates that over the past twenty years, much progress has been accomplished in capturing and integrating visual information into speech recognition. However, the visual modality has yet to become utilized in mainstream ASR systems. This is due to the fact that issues of both practical and research nature remain challenging. On the practical side of things, the high requirements in the captured video frame rate and size, necessary for extracting visual speech information that is capable of enhancing ASR performance, place increased demands on cost, storage, and computer processing. In addition, the lack of common, large audio-visual corpora that address a wide variety of ASR tasks, conditions, and environments, hinders development of audio-visual systems suitable for use in particular applications.

On the research side, key issues in the design of audio-visual ASR systems remain open and subject to more investigation. In the visual front end design, for example, face detection, facial feature localization, and face shape tracking, robust to unconstrained speaker, pose, lighting, and environment variation constitute challenging problems. A comprehensive comparison between face appearance and shape based features for speaker-dependent vs. speaker-independent automatic speechreading is also unavailable. Joint shape and appearance three-dimensional face modeling, used for both tracking and visual feature extraction has not been considered in the literature, although such an approach could possibly lead to the desired robustness and generality of the visual front end. In addition, when combining audio and visual information, a number of issues relevant to decision fusion require further study, such as the optimal level of integrating the audio and visual log-likelihoods and the optimal function for this integration.

Further investigation of these issues is clearly warranted, and it is expected to lead to improved robustness and performance of audio-visual ASR. Progress in addressing some or all of these questions can also benefit other areas where joint audio and visual speech processing is suitable [120], such as speaker identification and verification [20], [45], [60], [94], [121–123], visual text-to-speech [124–128], speech event detection [129], video indexing and retrieval [130], speech enhancement [88], [90], coding [131], signal separation [132], and speaker localization [133], [134]. Improvements in these areas will clearly result in more robust and natural human-computer interaction.

REFERENCES

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, 1997.

[2] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech Lang.*, vol. 5, pp. 275–294, 1991.

[3] R. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition. Advanced Topics*, C.-H. Lee, F. K. Soong, and Y. Ohshima, Eds. Norwell, MA: Kluwer Academic Pub., 1997, ch. 15, pp. 357–384.

[4] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*. Berlin, Germany: Springer, 1996.

[5] R. Campbell, B. Dodd, and D. Burnham, Eds., *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998.

[6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoustical Society America*, vol. 26, pp. 212–215, 1954.

[7] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[8] M. Marschark, D. LePoutre, and L. Bement, "Mouth movement and signed communication," in *Hearing by Eye II*, R. Campbell, B. Dodd, and D. Burnham, Eds. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998, ch. 13, pp. 245–266.

[9] L. E. Bernstein, M. E. Demorest, and P. E. Tucker, "What makes a good speechreader? first you have to find one," in *Hearing by Eye II*, R. Campbell, B. Dodd, and D. Burnham, Eds. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998, ch. 11, pp. 211–227.

[10] A. Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, R. Campbell and B. Dodd, Eds. London, United Kingdom: Lawrence Erlbaum Associates, 1987, pp. 3–51.

[11] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, pp. 236–244, 1998.

[12] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, pp. 23–43, 1998.

[13] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," in *Proc. Conf. Audio-Visual Speech Processing*, Santa Cruz, CA, Aug. 7–9, 1999, pp. 112–117.

[14] J. Jiang, A. Alwan, P. A. Keating, B. Chaney, E. T. Auer, Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1174–1188, Nov. 2002.

[15] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke, "Lips, teeth, and the benefits of lipreading," in *Handbook of Research on Face Processing*, H. D. Ellis and A. W. Young, Eds. Amsterdam, The Netherlands: Elsevier Science Publishers, 1989, pp. 223–233.

[16] P. M. T. Smeele, "Psychology of human speechreading," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 3–15.

[17] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 331–349.

[18] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. Global Telecomm. Conf.*, Atlanta, GA, 1984, pp. 265–272.

[19] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[20] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23–37, Mar. 2002.

[21] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 461–471.

[22] Q. Su and P. L. Silsbee, "Robust audiovisual integration using semi-continuous hidden Markov models," in *Proc. Int. Conf. Spoken Lang. Processing*, Philadelphia, PA, Oct. 3–6, 1996, pp. 42–45.

[23] J. R. Movellan and G. Chadderdon, "Channel separability in the audio visual integration of speech: A Bayesian approach," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 473–487.

[24] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, Rhodes, Greece, Sept. 27–28, 1997, pp. 53–56.

[25] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. Works. Multimedia Signal Processing*, Redondo Beach, CA, Dec. 7–9, 1998, pp. 65–70.

[26] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, Sept. 2000.

[27] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, UT, May 7–11, 2001, pp. 177–180.

[28] F. J. Huang and T. Chen, "Consideration of Lombard effect for speechreading," in *Proc. Works. Multimedia Signal Processing*, Cannes, France, Oct. 3–5, 2001, pp. 613–618.

[29] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1274–1288, Nov. 2002.

[30] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 3733–3736.

[31] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition," in *Proc. Int. Conf. Multimedia Expo*, New York, NY, July 30–Aug. 2, 2000, pp. 1073–1076.

[32] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 505–515.

[33] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," in *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, Rhodes, Greece, Sept. 27–28, 1997, pp. 61–64.

[34] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Adelaide, Australia, Apr. 19–22, 1994, pp. 669–672.

[35] G. Krone, B. Talle, A. Wichert, and G. Palm, "Neural architectures for sensorfusion in speech recognition," in *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, Rhodes, Greece, Sept. 27–28, 1997, pp. 57–60.

[36] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. III, Beijing, China, Oct. 16–20, 2000, pp. 20–23.

[37] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report, Oct. 2000.

[38] G. Potamianos and C. Neti, "Automatic speechreading of impaired speech," in *Proc. Conf. Audio-Visual Speech Processing*, Aalborg, Denmark, Sept. 7–9, 2001, pp. 177–182.

[39] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Lang. Processing*, Yokohama, Japan, Sept. 18–22, 1994, pp. 547–550.

[40] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Processing*, vol. I, Chicago, IL, Oct. 4–7, 1998, pp. 173–177.

[41] G. Chiou and J.-N. Hwang, "Lipreading from color video," *IEEE Trans. Image Processing*, vol. 6, pp. 1192–1195, Aug. 1997.

[42] N. Li, S. Dettmer, and M. Shah, "Lipreading using eigensequences," in *Proc. Int. Works. Automatic Face Gesture Recognition*, Zurich, Switzerland, 1995, pp. 30–34.

[43] P. Scanlon and R. Reilly, "Feature analysis for automatic speechreading," in *Proc. Works. Multimedia Signal Processing*, Cannes, France, Oct. 3–5, 2001, pp. 625–630.

[44] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma, "A cascade visual front end for speaker independent automatic speechreading," *Int. J. Speech Technol.*, vol. 4, pp. 193–208, July/Oct. 2001.

[45] P. Jourlin, "Word dependent acoustic-labial weights in HMM-based speech recognition," in *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, Rhodes, Greece, Sept. 27–28, 1997, pp. 69–72.

[46] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 629–642, Nov. 1999.

[47] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1260–1273, Nov. 2002.

[48] L. Czap, "Lip representation by image ellipse," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. IV, Beijing, China, Oct. 16–20, 2000, pp. 93–96.

[49] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Computer Vision Image Understanding*, vol. 65, pp. 163–178, 1997.

[50] D. Chandramohan and P. L. Silsbee, "A multiple deformable template approach for visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing*, Philadelphia, PA, Oct. 3–6, 1996, pp. 50–53.

[51] B. Dalton, R. Kaucic, and A. Blake, "Automatic speechreading using dynamic contours," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 373–382.

[52] 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 Processing*, vol. 2002, pp. 1213–1227, Nov. 2002.

[53] M. T. Chan, "HMM based audio-visual speech recognition integrating geometric- and appearance-based visual features," in *Proc. Works. Multimedia Signal Processing*, Cannes, France, Oct. 3–5, 2001, pp. 9–14.

[54] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Europ. Conf. Computer Vision*, Freiburg, Germany, 1998, pp. 484–498.

[55] A. Adjoudani, T. Guiard-Marigny, B. L. Goff, L. Reveret, and C. Benoît, "A multimedia platform for audio-visual speech processing," in *Proc. Europ. Conf. Speech Commun. Technol.*, Rhodes, Greece, Sept. 22–25, 1997, pp. 1671–1674.

[56] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Computer Vision*, vol. 4, pp. 321–331, 1988.

[57] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Computer Vision*, vol. 8, pp. 99–111, 1992.

[58] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - Their training and application," *Computer Vision Image Understanding*, vol. 61, pp. 38–59, Jan. 1995.

[59] H. P. Graf, E. Cosatto, and G. Potamianos, "Robust recognition of faces and facial features with a multi-modal system," in *Proc. Int. Conf. Systems, Man, Cybernetics*, Orlando, FL, 1997, pp. 2034–2039.

[60] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1228–1247, Nov. 2002.

[61] P. Daubias and P. Deléglise, "Automatically building and evaluating statistical models for lipreading," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1202–1212, Nov. 2002.

[62] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.

[63] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 39–51, Jan. 1998.

[64] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. Int. Conf. Audio Video-based Biometric Person Authentication*, Washington, DC, Mar. 22–23, 1999, pp. 154–159.

[65] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York, NY: John Wiley and Sons, 1965.

[66] C. Chatfield and A. J. Collins, *Introduction to Multivariate Analysis*. London, United Kingdom: Chapman and Hall, 1991.

[67] G. Potamianos and C. Neti, "Improved ROI and within frame discriminant features for lipreading," in *Proc. Int. Conf. Image Processing*, vol. III, Thessaloniki, Greece, Oct. 7–10, 2001, pp. 250–253.

[68] R. C. Gonzalez and P. Wintz, *Digital Image Processing*. Reading, MA: Addison-Wesley Publishing Company, 1977.

[69] I. Daubechies, *Wavelets*. Philadelphia, PA: S.I.A.M., 1992.

[70] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, May 7–10, 1996, pp. 821–824.

[71] M. S. Gray, J. R. Movellan, and T. J. Sejnowski, "Dynamic features for visual speech-reading: A systematic comparison," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 751–757.

[72] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. United Kingdom: Entropic Ltd., 1999.

[73] L. D. Rosenblum and H. M. Saldana, "Time-varying information for visual speech perception," in *Hearing by Eye II*, R. Campbell, B. Dodd, and D. Burnham, Eds. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998, ch. 3, pp. 61–81.

[74] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 661–664.

[75] L. Polymenakos, P. Olsen, D. Kanevsky, R. A. Gopinath, P. Gopalakrishnan, and S. Chen, "Transcription of broadcast news - some recent improvements to IBM's LVCSR system," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 901–904.

[76] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading." in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, May 7–10, 1996, pp. 833–836.

[77] O. Vanegas, A. Tanaka, K. Tokuda, and T. Kitamura, "HMM-based visual speech recognition using intensity and location normalization," in *Proc. Int. Conf. Spoken Lang. Processing*, Sydney, Australia, Nov. 30–Dec. 4, 1998, pp. 289–292.

[78] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, Jan. 2001.

[79] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1248–1259, Nov. 2002.

[80] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 2009–2012.

[81] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1189–1201, Nov. 2002.

[82] A. Rogozan, "Discriminative learning of visual data for audiovisual speech recognition," *Int. J. Artificial Intell. Tools*, vol. 8, pp. 43–52, 1999.

[83] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 337–351, Sept. 1996.

[84] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[85] L. R. Bahl, P. F. Brown, P. V. DeSouza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, Apr. 8–11, 1986, pp. 49–52.

[86] W. Chou, B.-H. Juang, C.-H. Lee, and F. Soong, "A minimum error rate pattern recognition approach to speech recognition," *Int. J. Pattern Recognition Artificial Intell.*, vol. 8, pp. 5–31, Jan. 1994.

[87] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips," in *Proc. Europ. Conf. Speech Commun. Technol.*, Madrid, Spain, Sept. 18–21, 1995, pp. 1559–1562.

[88] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoustical Society America*, vol. 109, pp. 3007–3020, 2001.

[89] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 2025–2028.

[90] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization)," in *Proc. Int. Conf. Spoken Lang. Processing*, Denver, CO, Sept. 16–20, 2002, pp. 1449–1452.

[91] L. Xu, A. Krzyżak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications in handwritten recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418–435, May/June 1992.

[92] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, Mar. 1998.

[93] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4–37, Jan. 2000.

[94] A. Jain, R. Bolle, and S. Pankanti, "Introduction to biometrics," in *Biometrics. Personal Identification in Networked Society*, A. Jain, R. Bolle, and S. Pankanti, Eds. Norwell, MA: Kluwer Academic Publishers, 1999, ch. 1, pp. 1–41.

[95] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. Int. Conf. Spoken Lang. Processing*, Philadelphia, PA, Oct. 3–6, 1996, pp. 426–429.

[96] H. Glotin and F. Berthommier, "Test of several external posterior weighting functions for multiband full combination ASR," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. I, Beijing, China, Oct. 16–20, 2000, pp. 333–336.

[97] C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. II, Beijing, China, Oct. 16–20, 2000, pp. 1023–1026.

[98] P. Beyerlein, "Discriminative model combination," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 481–484.

[99] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, UT, May 7–11, 2001, pp. 173–176.

[100] P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, NM, Apr. 3–6, 1990, pp. 845–848.

[101] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. Conf. Computer Vision Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 994–999.

[102] K. W. Grant and S. Greenberg, "Speech intelligibility derived from asynchronous processing of auditory-visual information," in *Proc. Conf. Audio-Visual Speech Processing*, Aalborg, Denmark, Sept. 7–9, 2001, pp. 132–137.

[103] X. Liu, Y. Zhao, X. Pi, L. Liang, and A. V. Nefian, "Audio-visual continuous speech recognition using a coupled hidden Markov model," in *Proc. Int. Conf. Spoken Lang. Processing*, Denver, CO, Sept. 16–20, 2002, pp. 213–216.

[104] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Human Lang. Technol. Conf.*, San Diego, CA, Mar. 24–27, 2002, pp. 1–6.

[105] S. Nakamura, "Fusion of audio-visual information for integrated speech processing," in *Audio-and Video-Based Biometric Person Authentication*, J. Bigun and F. Smeraldi, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 127–143.

[106] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-

visual ASR," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 853–856.
[107] J. A. Nelder and R. Mead, "A simplex method for function minimisation," *Computing J.*, vol. 7, pp. 308–313, 1965.
[108] F. Berthommier and H. Glotin, "A new SNR-feature mapping for robust multistream speech recognition," in *Proc. Int. Congress Phonetic Sciences*, San Francisco, CA, 1999, pp. 711–715.
[109] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. III, Beijing, China, Oct. 16–20, 2000, pp. 746–749.
[110] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
[111] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Lang.*, vol. 9, pp. 171–185, 1995.
[112] L. Neumeyer, A. Sankar, and V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proc. Europ. Conf. Speech Commun. Technol.*, Madrid, Spain, Sept. 18–21, 1995, pp. 1127–1130.
[113] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Munich, Germany, 1997, pp. 1043–1046.
[114] M. J. F. Gales, "Maximum likelihood multiple projection schemes for hidden Markov models," Cambridge University, Cambridge, United Kingdom, Tech. Rep., 1999.
[115] G. Potamianos and A. Potamianos, "Speaker adaptation for audio-visual speech recognition," in *Proc. Europ. Conf. Speech Commun. Technol.*, Budapest, Hungary, Sept. 5–9, 1999, pp. 1291–1294.
[116] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Audio-and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds. Berlin, Germany: Springer, 1997, pp. 403–409.
[117] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTS: The extended M2VTS database," in *Proc. Int. Conf. Audio Video-based Biometric Person Authentication*, Washington, DC, Mar. 22–23, 1999, pp. 72–76.
[118] G. Iyengar and C. Neti, "Detection of faces under shadows and lighting variations," in *Proc. Works. Multimedia Signal Processing*, Cannes, France, Oct. 3–5, 2001, pp. 15–20.
[119] L. R. Bahl, S. V. De Gennaro, P. S. Gopalakrishnan, and R. L. Mercer, "A fast approximate acoustic match for large vocabulary speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 59–67, Jan. 1993.
[120] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, pp. 837–852, May 1998.
[121] T. Wark and S. Sridharan, "A syntactic approach to automatic lip feature extraction for speaker identification," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 3693–3696.
[122] B. Fröba, C. Küblbeck, C. Rothe, and P. Plankensteiner, "Multi-sensor biometric person recognition in an access control system," in *Proc. Int. Conf. Audio Video-based Biometric Person Authentication*, Washington, DC, Mar. 22–23, 1999, pp. 55–59.
[123] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for broadcast news: some fusion techniques," in *Proc. Works. Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 13–15, 1999, pp. 161–167.
[124] M. M. Cohen and D. W. Massaro, "What can visual speech synthesis tell visual speech recognition?" in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1994.
[125] T. Chen, H. P. Graf, and K. Wang, "Lip synchronization using speech-assisted video processing," *IEEE Signal Processing Lett.*, vol. 2, pp. 57–59, Apr. 1995.
[126] E. Cosatto, G. Potamianos, and H. P. Graf, "Audio-visual unit selection for the synthesis of photo-realistic talking-heads," in *Proc. Int. Conf. Multimedia Expo*, New York, NY, July 30–Aug. 2, 2000, pp. 1097–1100.
[127] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, vol. 2, pp. 152–163, Sept. 2000.
[128] J. J. Williams and A. K. Katsaggelos, "An HMM-based speech-to-video synthesizer," *IEEE Trans. Neural Networks*, vol. 13, pp. 900–915, July 2002.
[129] P. De Cuetos, C. Neti, and A. Senior, "Audio-visual intent to speak detection for human computer interaction," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, Turkey, June 5–9, 2000, pp. 1325–1328.
[130] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, "Integration of multimodal features for video scene classification based on HMM," in *Proc. Works. Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 13–15, 1999, pp. 53–58.
[131] E. Foucher, L. Girin, and G. Feng, "Audiovisual speech coder: Using vector quantization to exploit the audio/video correlation," in *Proc. Conf. Audio-Visual Speech Processing*, Terrigal, Australia, Dec. 4–6, 1998, pp. 67–71.
[132] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1165–1173, Nov. 2002.
[133] U. Bub, M. Hunke, and A. Waibel, "Knowing who to listen to in speech recognition: Visually guided beamforming," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, May 9–12, 1995, pp. 848–851.
[134] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1154–1164, Nov. 2002.