Adaptive Multimodal Fusion by Uncertainty Compensation With Application to Audiovisual Speech Recognition

George Papandreou, *Student Member, IEEE*, Athanassios Katsamanis, *Student Member, IEEE*, Vassilis Pitsikalis, *Member, IEEE*, and Petros Maragos, *Fellow, IEEE*

***Abstract*—While the accuracy of feature measurements heavily depends on changing environmental conditions, studying the consequences of this fact in pattern recognition tasks has received relatively little attention to date. In this paper, we explicitly take feature measurement uncertainty into account and show how multimodal classification and learning rules should be adjusted to compensate for its effects. Our approach is particularly fruitful in multimodal fusion scenarios, such as audiovisual speech recognition, where multiple streams of complementary time-evolving features are integrated. For such applications, provided that the measurement noise uncertainty for each feature stream can be estimated, the proposed framework leads to highly adaptive multimodal fusion rules which are easy and efficient to implement. Our technique is widely applicable and can be transparently integrated with either synchronous or asynchronous multimodal sequence integration architectures. We further show that multimodal fusion methods relying on stream weights can naturally emerge from our scheme under certain assumptions; this connection provides valuable insights into the adaptivity properties of our multimodal uncertainty compensation approach. We show how these ideas can be practically applied for audiovisual speech recognition. In this context, we propose improved techniques for person-independent visual feature extraction and uncertainty estimation with active appearance models, and also discuss how enhanced audio features along with their uncertainty estimates can be effectively computed. We demonstrate the efficacy of our approach in audiovisual speech recognition experiments on the CUAVE database using either synchronous or asynchronous multimodal integration models.**

***Index Terms*—Active appearance models (AAMs), audiovisual automatic speech recognition (AV-ASR), multimodal fusion, uncertainty compensation.**

I. Introduction

MOTIVATED by the multimodal way humans perceive their environment [1], complementary information sources have been successfully used in many applications. Such a case is audiovisual automatic speech recognition (AV-ASR) [2], [3], where fusing visual and audio cues can lead to substantially improved performance relatively to audio-only recognition, especially in the presence of audio noise.

However, successfully integrating heterogeneous information streams is a challenging task [4]–[7]. Devising robust combination mechanisms is highly nontrivial, mainly because multimodal schemes need to automatically adapt to dynamic environmental conditions which can dissimilarly affect the reliability of the separate modalities, essentially contaminating feature measurements with nonstationary noise. For example, the visual stream in AV-ASR should be discounted when the visual front-end momentarily mistracks the speaker's face. Other complicating factors, such as the lack of exact synchronization across different modalities, make traditional unimodal estimation/classification techniques less appropriate to handle multimodal data and further add to the complexity of the multimodal integration problem.

The technique presented in this work is exactly geared towards dynamic adaptation of multimodal fusion schemes to changing environmental conditions. We approach the problem of adaptive multimodal fusion by explicitly taking feature measurement uncertainty of the different modalities into account. A preliminary version of our work appeared in [8]–[10]. In single modality, audio-only scenarios, modeling audio feature noise has proven fruitful for noise-robust ASR [11]–[14] and also in applications such as speaker verification [15] and multiband ASR [16]; see [17] for further pointers to the related literature. We extend these ideas to the multimodal setting and show in Section II how multi-stream classification rules should be adjusted to compensate for feature measurement uncertainty. We discuss in detail and derive modified classification algorithms which take feature measurement uncertainty into account for Gaussian mixture models (GMMs) and hidden Markov models (HMMs), but the technique can also be seamlessly integrated with existing methods such as Product-HMMs that allow handling loosely synchronized multimodal data [18]–[21]. The proposed scheme leads to multimodal fusion rules which are adaptive at the frame level, widely applicable, and easy to implement. Multimodal model training under uncertain features is also covered, and modified expectation–maximization (EM) algorithms for GMMs and HMMs are presented in Section IV.

Of particular interest is the connection of our formulation with existing stream weight-based multimodal fusion techniques, which we discuss in Section III. In particular, we show that our scheme under certain assumptions effectively leads to

Manuscript received January 27, 2008; revised July 31, 2008. Current version published February 11, 2009. This work was supported in part by the European Network of Excellence MUSCLE (IST-FP6-507752), in part by the European FP6 FET research project ASPI (IST-FP6-021324), and in part by the projects ΠΕΝΕΔ-2003 ΕΔ-865 & 866, which are cofinanced by the E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerasimos (Makis) Potamianos.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece (e-mail: gpapan@cs.ntua.gr; nkatsam@cs.ntua.gr; vpitsik@cs.ntua.gr; maragos@cs.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2011515

adaptive stream weighting. This sheds new light onto the probabilistic underpinnings of stream weighting and also provides insights in the adaptivity properties of our scheme. Moreover, we suggest novel hybrid methods combining the stream weight approach and our adaptive compensation mechanism, in which stream weighting offers a discriminatively motivated bias towards the most informative modality, while uncertainty compensation offers a fine-grained adaptation mechanism which accounts for varying environmental conditions.

The applicability of the proposed multimodal fusion approach is illustrated in the context of audiovisual speech recognition, as described in Section V. Similarly to [22], our visual feature extraction front-end is based on active appearance models (AAMs) of the speaker's face [23]. An important novelty in our visual front-end is a speaker adaptation mechanism that discounts the inherent appearance variability of neutral-pose multiple person face images which is irrelevant to visual speech. The AAM can then concentrate its visual modeling power on the appearance variability caused by speech-related facial expressions; in the context of AV-ASR we term the resulting model a *visemic* AAM. We also demonstrate how AAM feature uncertainty can be estimated as part of the AAM face matching process. Regarding the audio front-end, we build on the recent technique of [14] which allows estimating both the enhanced speech feature vector and its corresponding uncertainty in a unified manner. We show that the same technique can be extended beyond the unimodal setting of [14] and be integrated in our adaptive multimodal fusion framework. We evaluate the proposed method in AV-ASR experiments using multi-stream HMMs, demonstrating improved performance. Applying our technique in conjunction with Product-HMMs, which better account for cross-modal asynchrony, we obtain further improvements.

II. Feature Uncertainty and Multimodal Fusion

Let us consider a pattern classification scenario, in which we measure a property (feature) of a pattern instance and try to decide to which class $c \in \{1, \ldots, N\}$ it should be assigned. The measurement is a realization $\mathbf{x}$ of a random vector X, whose statistics differ for the N classes. Typically, for each class we have trained a model that captures these statistics and represents the class-conditional probability densities $p(\mathbf{x}|c)$. Our decision is then based on some appropriate rule, e.g., the maximum *a posteriori* (MAP) criterion $\hat{c} = \arg\max p(c|\mathbf{x}) = \arg\max p(\mathbf{x}|c)p(c)$.

One may identify three major sources of uncertainty that could perplex classification. First, *class overlap* due to improper modeling or limited discriminability of the feature set for the classification task. For instance, visual cues cannot discriminate between members of the same viseme class (e.g., /p/, /b/) [3]. Better choice of features and modeling schemes can reduce this uncertainty. Second, *parameter estimation uncertainty* that mainly originates from insufficient training [24]. Third, *feature observation uncertainty* due to errors in the measurement process or noise contamination, which is the type of uncertainty we focus on in this paper. Note that feature measurement uncertainty is a central idea in classic estimation theory, playing a fundamental role, e.g., in the Kalman and Wiener filters [25]. In essence, our paper studies optimal fusion of noisy multimodal measurements for the task of classification, while estimation theory is about optimal fusion of multiple noisy information sources for the task of recovering an unknown continuous quantity.

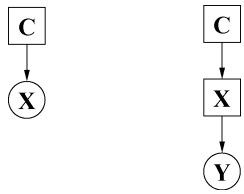

Fig. 1. Pictorial representation of feature measurement scenarios, with hidden and observed variables enclosed in squares and circles, respectively. Left: conventional case—we observe the features $\mathbf{x}$ directly. Right: noisy measurement case—we only observe the noisy features $\mathbf{y}$.

A. Feature Observation Uncertainty and Its Compensation in Classification

We can formulate feature observation uncertainty considering that the actual feature measurement $\mathbf{y}$ is just a noisy/corrupted version of the inaccessible clean feature $\mathbf{x}$. More specifically, we adopt the measurement model

$$Y = X + E \tag{1}$$

and assume that the noise probability density $p(\mathbf{e})$ is known; this scenario is graphically depicted in Fig. 1 and corresponds to *measurement error* models in statistics [26]. Under this observation model, classification decisions must rely on $p(c|\mathbf{y}) \propto p(\mathbf{y}|c)p(c)$, and thus $p(\mathbf{y}|c)$ needs to be computed.

To determine the desirable noisy feature density function $p(\mathbf{y}|c)$, we need to integrate-out the hidden clean feature variable $\mathbf{x}$

$$p(\mathbf{y}|c) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|c)d\mathbf{x}. \tag{2}$$

Although the integral in (2) is in general intractable, we can obtain a closed-form solution in the important special case of Gaussian data model, $p(\mathbf{x}|c) = N(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, with Gaussian observation noise, $p(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}; \mathbf{x}+\boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e)$, where $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the multivariate Gaussian probability density function on $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, one can show that $p(\mathbf{y}|c)$ is given by

$$p(\mathbf{y}|c) = N(\mathbf{y}; \boldsymbol{\mu}_c + \boldsymbol{\mu}_e, \boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_e) \tag{3}$$

implying that we can proceed by considering our features $\mathbf{y}$ clean, provided that we shift the model means $\boldsymbol{\mu}_c$ by $\boldsymbol{\mu}_e$ (*enhancement* step) and increase the model covariances $\boldsymbol{\Sigma}_c$ by $\boldsymbol{\Sigma}_e$ (*variance compensation* step). A similar approach has been previously followed in related audio-only applications [12], [14], [15].

To illustrate (3), we discuss with reference to Fig. 2 how observation uncertainty influences decisions in a simple two-class classification task. The two classes are modeled by 2-D spherical Gaussian distributions, $N(\boldsymbol{\mu}_1, \sigma_1^2\mathbf{I})$, $N(\boldsymbol{\mu}_2, \sigma_2^2\mathbf{I})$ and they

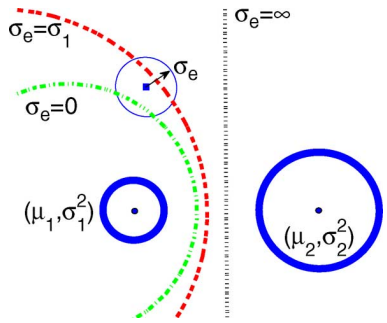

Fig. 2. Decision boundaries for classification of a noisy observation (square marker) in two classes, shown as circles, for various observation noise variances. Classes are modeled by spherical Gaussians of means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and variances $\sigma_1^2\mathbf{I}$, $\sigma_2^2\mathbf{I}$, respectively. The decision boundary is plotted for three values of noise variance (a) $\sigma_e = 0$ (i.e., no observation uncertainty), (b) $\sigma_e = \sigma_1$, and (c) $\sigma_e = \infty$. With increasing noise variance, the boundary moves away from its noise-free position.

have equal prior probability. If our observation $\mathbf{y}$ contains zero mean spherical Gaussian noise with covariance matrix $\sigma_e^2\mathbf{I}$, then the modified decision boundary consists of those $\mathbf{y}$ for which $N(\mathbf{y};\boldsymbol{\mu}_1,\sigma_1^2\mathbf{I}+\sigma_e^2\mathbf{I}) = N(\mathbf{y};\boldsymbol{\mu}_2,\sigma_2^2\mathbf{I}+\sigma_e^2\mathbf{I})$. When σ_e^2 is zero, the decision should be made as in the noise-free case. If σ_e^2 is comparable to the variances of the models, then the modified boundary significantly differs from the original one and neglecting observation uncertainty in the decision process increases misclassifications.

B. Observation Uncertainty and Multimodal Fusion

For many applications, one can get improved performance by exploiting complementary features, stemming from a single or multiple modalities. Let us assume that one wants to integrate S information streams which produce feature vectors $\mathbf{x}_s$, $s = 1,\ldots,S$. Application of Bayes' formula yields the posterior class label probability given the full observation vector $\mathbf{x}_{1:S} \equiv (\mathbf{x}_1;\ldots;\mathbf{x}_S)$

$$p(c|\mathbf{x}_{1:S}) \propto p(c)p(\mathbf{x}_{1:S}|c). \tag{4}$$

If the features are statistically independent given the class label c (see [27] for a discussion of this property in the context of audiovisual speech), the conditional probability of the aggregate observation vector $\mathbf{x}_{1:S}$ becomes separable and is given by the product rule, implying that (4) can be written as

$$p(c|\mathbf{x}_{1:S}) \propto p(c)\prod_{s=1}^{S} p(\mathbf{x}_s|c). \tag{5}$$

This case corresponds to what Clark and Yuille [4] call *weakly coupled* data fusion.

We will now show that accounting for feature uncertainty naturally leads to a novel adaptive mechanism for fusion of different information sources. Since in our stochastic measurement framework we do not have direct access to the features $\mathbf{x}_s$, our decision mechanism depends on their noisy counterparts $\mathbf{y}_s = \mathbf{x}_s + \mathbf{e}_s$. Assuming noise independence across the streams, the probability of interest is thus obtained by integrating out the hidden clean features $\mathbf{x}_s$, i.e.,

$$p(c|\mathbf{y}_{1:S}) \propto p(c)\prod_{s=1}^{S}\int p(\mathbf{y}_s|\mathbf{x}_s)p(\mathbf{x}_s|c)d\mathbf{x}_s. \tag{6}$$

In the common case that the clean feature emission probability is modeled as a GMM, i.e., $p(\mathbf{x}_s|c) = \sum_{m=1}^{M_{s,c}} \rho_{s,c,m}N(\mathbf{x}_s;\boldsymbol{\mu}_{s,c,m},\boldsymbol{\Sigma}_{s,c,m})$, and the observation noise at each stream is considered Gaussian, i.e.,

$$p(\mathbf{y}_s|\mathbf{x}_s) = N(\mathbf{y}_s;\mathbf{x}_s+\boldsymbol{\mu}_{e,s},\boldsymbol{\Sigma}_{e,s}) \tag{7}$$

it directly follows that

$$p(c|\mathbf{y}_{1:S}) \propto p(c)\prod_{s=1}^{S}\sum_{m=1}^{M_{s,c}} \rho_{s,c,m}N(\mathbf{y}_s;\boldsymbol{\mu}_{s,c,m}+\boldsymbol{\mu}_{e,s},\boldsymbol{\Sigma}_{s,c,m}+\boldsymbol{\Sigma}_{e,s}) \tag{8}$$

which, as in the single-stream case (3), involves considering our features $\mathbf{y}_s$ clean, while shifting the model means $\boldsymbol{\mu}_{s,c,m}$ by $\boldsymbol{\mu}_{e,s}$ and increasing the model covariances $\boldsymbol{\Sigma}_{s,c,m}$ by $\boldsymbol{\Sigma}_{e,s}$. Using mixtures of Gaussians for the measurement noise $p(\mathbf{y}_s|\mathbf{x}_s)$ is straightforward and could be useful in case of heavy-tailed noise distributions or for modeling observation outliers. Also note that, although the measurement noise covariance matrix $\boldsymbol{\Sigma}_{e,s}$ of each stream is the same for all classes c and all mixture components m, noise particularly affects the most peaked mixtures, for which $\boldsymbol{\Sigma}_{e,s}$ is substantial relative to the modeling uncertainty $\boldsymbol{\Sigma}_{s,c,m}$. The adaptive fusion effect of feature uncertainty compensation in a two-class classification task using two streams is illustrated in Fig. 3.

III. Stream Weights and Uncertainty Compensation

A. Stream Weights in Multimodal Fusion

A common theme in many stream integration methods is the use of stream weights to equalize the different modalities. Stream weights w_s act as exponents in the original product rule (5), resulting in the modified posterior-like score

$$b(c|\mathbf{x}_{1:S}) = p(c)\prod_{s=1}^{S} p(\mathbf{x}_s|c)^{w_s} \tag{9}$$

which can be seen in a logarithmic scale as a weighted average of the individual stream log-probabilities. Selection of stream weights is typically governed by two factors, namely 1) discrimination capacity of each modality for the given task and 2) amount of feature degradation caused by adverse environmental conditions. For example, in the context of AV-ASR, bigger weight is typically assigned to the more informative audio modality than to the visual modality in clean acoustic conditions, but the visual share is gradually increased as acoustic conditions deteriorate. The technique has been routinely employed in fusion tasks involving either different audio-only

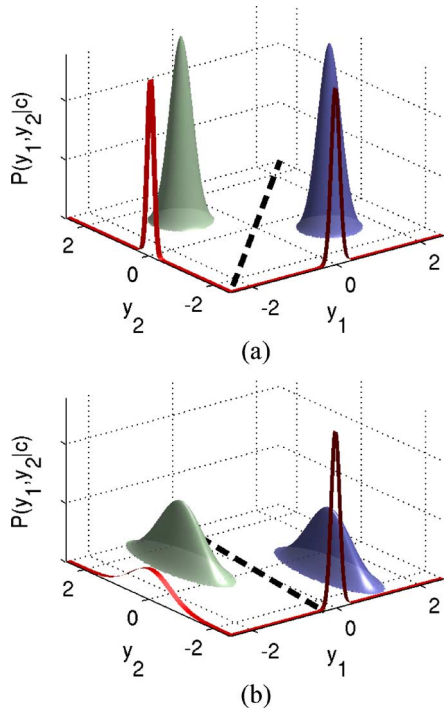

Fig. 3. Multimodal variance compensation leads to adaptive fusion. We illustrate a two-class classification scenario using two Gaussian feature streams, $\mathbf{y}_1$ and $\mathbf{y}_2$, with equal model covariances $\boldsymbol{\Sigma}_{s,c} = \sigma^2\mathbf{I}$. The measurement noise density of each stream is plotted on top of its corresponding axis, while the classification decision boundary is drawn with a dashed line. (a) Negligible measurement noise in either stream: the decision boundary lies on the axes' diagonal. (b) Substantial measurement noise in the $\mathbf{y}_2$ stream, $\sigma_{e,2} \gg \sigma_{e,1}$: the decision boundary moves and classification is mostly influenced by the feature value of the reliable $\mathbf{y}_1$ stream.

streams [16] or multimodal audio and visual streams [3]; early related AV-ASR references are [28] and [29]. Such stream weights have been applied not only in conventional HMMs, but also in conjunction with more flexible architectures which better account for the asynchronicity of audiovisual speech, such as Product-HMMs and more general dynamic Bayesian networks [18]–[21].

The stream weights formulation has however some important shortcomings. From a theoretical viewpoint, the weighted score $b(c|\mathbf{x}_{1:S})$ in (9) no longer has the probabilistic interpretation of (5) as class probability given the full observation vector $\mathbf{x}_{1:S}$. From a more practical standpoint, it is not straightforward to optimally select stream weights. Most authors set them discriminatively for a given set of environment conditions (e.g., audio noise level in the case of audiovisual speech recognition) by minimizing the classification error on a held-out set, and then keep them constant throughout the recognition phase. However, this is insufficient, since attaining optimal performance requires that we dynamically adjust the share of each stream in the decision process, e.g., to account for visual tracking failures in the AV-ASR case. There have been some efforts towards dynamically adjustable stream weights, as well as stream weights adapted to the phonemic content of audiovisual speech (in the form of unit- or even class-dependent stream weights) [30]–[32]; however, stream weight tuning in this context is challenging, typically requiring extensive training sets.

B. Effective Stream Weights in Uncertainty Compensation

Although our multimodal fusion scheme for uncertainty compensation given by (8) seemingly bears little resemblance to the stream weights formulation of (9), there are interesting connections between the two approaches which become apparent if we examine a particularly illuminating special case of our uncertainty compensation result. Specifically, with reference to (8), we consider a scenario in which the following two assumptions hold.

1) The measurement noise covariance is a scaled version of the model covariance, i.e., $\boldsymbol{\Sigma}_{e,s} = r_{s,c,m}\boldsymbol{\Sigma}_{s,c,m}$. Note that the $r_{s,c,m}$ are not parameters to be tuned but just the relative measurement errors $\boldsymbol{\Sigma}_{e,s}/\boldsymbol{\Sigma}_{s,c,m}$. Intuitively, as the signal-to-noise ratio (SNR) for stream s drops, the corresponding relative measurement error $r_{s,c,m}$ increases.
2) For every stream observation $\mathbf{y}_s$, the Gaussian mixture response of that stream is dominated by a single component m_0 or, equivalently, there is little overlap among different Gaussian mixtures.

Under these conditions, the Gaussian densities in (8) can be approximated by $N(\mathbf{y}_s; \boldsymbol{\mu}_{s,c,m_0} + \boldsymbol{\mu}_{e,s}, (1 + r_{s,c,m_0})\boldsymbol{\Sigma}_{s,c,m_0})$; using the power-of-Gaussian identity $N(\mathbf{x}; \boldsymbol{\mu}, w^{-1}\boldsymbol{\Sigma}) = (\det(w(2\pi\boldsymbol{\Sigma})^{w-1}))^{1/2} N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})^w \propto N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})^w$ yields

$$p(c|\mathbf{y}_{1:S}) \propto p(c) \times \prod_{s=1}^{S} \left[\tilde{\rho}_{s,c,m_0} N\left(\mathbf{y}_s; \boldsymbol{\mu}_{s,c,m_0} + \boldsymbol{\mu}_{e,s}, \boldsymbol{\Sigma}_{s,c,m_0}\right)\right]^{w_{s,c,m_0}} \quad (10)$$

where

$$w_{s,c,m_0} = 1/(1 + r_{s,c,m_0}) \quad (11)$$

is the *effective stream weight* and $\tilde{\rho}_{s,c,m_0} = (\det(w_{s,c,m_0}(2\pi\boldsymbol{\Sigma}_{s,c,m_0})^{w_{s,c,m_0}-1}))^{1/2}\rho_{s,c,m_0}$ is a modified mixture weight which is independent of the observation $\mathbf{y}_s$. Note that the effective stream weights are between 0 (for $r_{s,c,m_0} \gg 1$) and 1 (for $r_{s,c,m_0} \approx 0$) and discount the contribution of each stream to the final result by properly taking its relative measurement error into account. The most important aspect of our effective stream weights w_{s,c,m_0} in (11) is that they are adaptive at the finest possible granularity: 1) environmental noise compensation is tailored to the error characteristics $(\boldsymbol{\mu}_{e,s}, \boldsymbol{\Sigma}_{e,s})$ of each new measurement $\mathbf{y}_s$, implying frame-level adaptation in applications such as AV-ASR; 2) content-based effective weight adjustment goes down to the class label and Gaussian mixture component. This level of adaptivity is beyond the reach of conventional stream weight adaptation techniques and is achieved without the need to tune numerous parameters on large validation datasets.

The simplifying assumptions behind the effective stream weights formula (11) will typically not hold in practice. In our implementation, we never use (10) or compute w_{s,c,m_0}, but rather always use the general variance compensation formula (8). Nevertheless, the arguments above qualitatively suggest that our uncertainty compensation scheme of (8) is actually a highly adaptive method for multimodal fusion.

C. Stream Weights and Uncertainty Compensation Hybrids

The preceding analysis in Section III-B has unveiled some interesting ties between the traditional stream weights approach and our uncertainty compensation scheme. We will build on these ties to propose hybrid schemes which combine the advantages of both formulations.

While our uncertainty compensation scheme has been derived from a model-based probabilistic perspective and the underlying model training principle is maximum likelihood, the stream weights formulation could be justified under discriminative arguments and discriminative training criteria are appropriate for it [33], [34]. The importance of discriminative approaches to audio-only ASR has been highlighted by the success of discriminative model training techniques using the maximum mutual information [35] or the minimum classification error rate [36] criteria, which often produce models with improved recognition performance relative to maximum likelihood. The success of discriminative criteria stems from the fact that, in contrast to model-based approaches, they take account of competing classification hypotheses and try to reduce the probability of incorrect assignments, or even directly minimize recognition errors. This pragmatic viewpoint makes discriminative approaches more robust to model mis-specification, e.g., when the actual data statistics are poorly described by the GMM/HMM assumptions.

In this context, it is reasonable to propose combining our model-based uncertainty compensation scheme with stream weighting, resulting to the following multimodal fusion scheme which is a hybrid of (8) and (9)

$$\begin{aligned} b(c|\mathbf{y}_{1:S}) &= p(c)\prod_{s=1}^{S} p(\mathbf{y}_s|c)^{w_s} \\ &\propto p(c)\prod_{s=1}^{S}\left(\sum_{m=1}^{M_{s,c}} \rho_{s,c,m} \times N(\mathbf{y}_s;\boldsymbol{\mu}_{s,c,m}+\boldsymbol{\mu}_{e,s},\boldsymbol{\Sigma}_{s,c,m}+\boldsymbol{\Sigma}_{e,s})\right)^{w_s}. \end{aligned} \tag{12}$$

This hybrid scheme combines the improved discriminative characteristics of stream weights with the advantageous adaptivity properties of our uncertainty compensation scheme into a powerful blend. Such a scheme also makes sense intuitively, since, for example, in AV-ASR experiments performed under controlled conditions with very little acoustic noise it is beneficial to place bigger weight to the more informative audio stream. The experiments reported in Section VI demonstrate the effectiveness of the hybrid scheme.

IV. EM Training Under Uncertainty

In many real-world applications requiring large amounts of training data, very accurate training sets collected under strictly controlled conditions are very difficult to gather. For example, in audiovisual speech recognition it is unrealistic to assume that a human expert annotates each frame in the training videos. A usual compromise is to adopt a semi-automatic annotation technique which yields a sufficiently diverse training set; since such a technique can introduce non-negligible feature errors in the training set, it is desirable to take training set feature uncertainty into account in learning procedures.

A. EM Training for GMMs

Under our feature uncertainty viewpoint, only a noisy version $\mathbf{y}$ of the underlying true property $\mathbf{x}$ can be observed. Maximum-likelihood estimation of the GMM parameters $\boldsymbol{\theta}$ from a training set $\mathcal{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_T\}$ under the EM algorithm [37] should thus consider as hidden variables not only the class memberships $\mathcal{M}$, but also the corresponding clean features $\mathcal{X}$. The expected complete-data log-likelihood $Q(\boldsymbol{\theta},\boldsymbol{\theta}') = E[\log p(\mathcal{Y},\{\mathcal{X},\mathcal{M}\}|\boldsymbol{\theta})|\mathcal{Y},\boldsymbol{\theta}']$ of the parameters $\boldsymbol{\theta}$ in the EM algorithm's current iteration given the previous guess $\boldsymbol{\theta}'$ in the **E-step** should thus be obtained by summing over discrete and integrating over continuous hidden variables. In the single stream case this translates to

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\theta}') = &\sum_{t=1}^{T}\sum_{m=1}^{M}\log \pi_m p(m|\mathbf{y}_t,\boldsymbol{\theta}') \\ &+\sum_{t=1}^{T}\sum_{m=1}^{M}\int \log p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t,m|\mathbf{y}_t,\boldsymbol{\theta}')d\mathbf{x}_t \\ &+\sum_{t=1}^{T}\sum_{m=1}^{M}\int \log p(\mathbf{x}_t|m,\boldsymbol{\theta})p(\mathbf{x}_t,m|\mathbf{y}_t,\boldsymbol{\theta}')d\mathbf{x}_t. \end{aligned} \tag{13}$$

We get the updated parameters $\boldsymbol{\theta}$ in the **M-step** by maximizing $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ over $\boldsymbol{\theta}$, yielding

$$\pi_m = \frac{r_m}{T}, \quad \text{with} \quad r_m = \sum_{t=1}^{T} r_{t,m}, \tag{14}$$

$$\boldsymbol{\mu}_m = \frac{1}{r_m}\sum_{t=1}^{T} r_{t,m}\hat{\mathbf{x}}_{t,m} \tag{15}$$

$$\boldsymbol{\Sigma}_m = \frac{1}{r_m}\sum_{t=1}^{T} r_{t,m} \times \left(\boldsymbol{\Sigma}_{x_{t,m}} + (\hat{\mathbf{x}}_{t,m}-\boldsymbol{\mu}_m)(\hat{\mathbf{x}}_{t,m}-\boldsymbol{\mu}_m)^T\right) \tag{16}$$

where (the prime denotes previous-step parameter estimates)

$$\begin{aligned} r_{t,m} &= p(m|\mathbf{y}_t,\boldsymbol{\theta}') \\ &\propto \pi'_m N\left(\mathbf{y}_t;\boldsymbol{\mu}'_m+\boldsymbol{\mu}_{e,t},\boldsymbol{\Sigma}'_m+\boldsymbol{\Sigma}_{e,t}\right) \end{aligned} \tag{17}$$

$$\hat{\mathbf{x}}_{t,m} = \boldsymbol{\Sigma}_{x_{t,m}}\left(\left(\boldsymbol{\Sigma}'_m\right)^{-1}\boldsymbol{\mu}'_m + (\boldsymbol{\Sigma}_{e,t})^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_{e,t})\right) \tag{18}$$

$$\boldsymbol{\Sigma}_{x_{t,m}} = \left(\left(\boldsymbol{\Sigma}'_m\right)^{-1} + (\boldsymbol{\Sigma}_{e,t})^{-1}\right)^{-1}. \tag{19}$$

The resulting EM algorithm has some notable differences with respect to the noise-free case. Specifically, in computing the responsibilities $r_{t,m}$ in (17) during the *E-step*, error-compensated scores are used. Also, in updating the model's means and variances during the *M-step* in (15) and (16), one should replace each noisy measurement $\mathbf{y}_t$ used in conventional GMM training with its model-enhanced counterparts, described by the expected values $\hat{\mathbf{x}}_{t,m}$ and the uncertainties $\boldsymbol{\Sigma}_{x_{t,m}}$. In

particular, the enhancement uncertainty $\boldsymbol{\Sigma}_{x_{t,m}}$ enters in (16) and regularizes the computation of the model variance $\boldsymbol{\Sigma}_m$. Furthermore, in the multimodal case with multiple streams $s = 1, \ldots, S$, one should compute the responsibilities by $r_{t,m} \propto \pi'_m \prod_{s=1}^{S} N(\mathbf{y}_{s,t}; \boldsymbol{\mu}'_{s,m} + \boldsymbol{\mu}_{s,e,i}, \boldsymbol{\Sigma}'_{s,m} + \boldsymbol{\Sigma}_{s,e,t})$, which generalizes (17) and introduces interactions among the modalities. Analogous EM formulas for HMM parameter estimation are given in the Appendix.

Similarly to the analysis in Section III-B, we can gain further insight into the previous EM formulas by considering the special case of zero-mean errors with constant and model-aligned covariance matrices, i.e., $\boldsymbol{\mu}_{e,t} = 0$ and $\boldsymbol{\Sigma}_{e,t} = \boldsymbol{\Sigma}_e = \lambda_m \boldsymbol{\Sigma}_m$. Then, one can easily show that, after convergence, the covariance formula in (16) can be written as

$$\boldsymbol{\Sigma}_m = \frac{1}{1+\lambda_m} \tilde{\boldsymbol{\Sigma}}_m, \text{ or, equivalently, } \boldsymbol{\Sigma}_m = \tilde{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_e \quad (20)$$

i.e., we simply subtract from the conventional (uncompensated) covariance estimate $\tilde{\boldsymbol{\Sigma}}_m = (1/r_m) \sum_{t=1}^{T} r_{t,m} (\mathbf{y}_t - \boldsymbol{\mu}_m)(\mathbf{y}_t - \boldsymbol{\mu}_m)^T$ the noise covariance $\boldsymbol{\Sigma}_e$. The rule in (20) has been used before as a heuristic for fixing the model covariance estimate after conventional EM training with noisy data (e.g., [38]). We have shown that it is partly justified in the constant and model-aligned errors case; otherwise, one should use the more general rules in (16).

V. Audiovisual Speech Recognition

A challenging application domain for multimodal fusion schemes is audiovisual automatic speech recognition (AV-ASR), since it requires modeling both the relative reliability and the synchronicity of the audio and visual modalities. We demonstrate that the proposed fusion scheme can be readily integrated with multistream HMMs or other multimodal sequence processing techniques and improve their performance in AV-ASR.

A. Visual Feature Extraction and Uncertainty Estimation

Salient visual speech information can be obtained from the speaker's visible articulators, mainly the lips and the jaw, which constitute the *region of interest* (ROI) around the mouth [3]. Visual information typically comprises geometrical shape characteristics, as well as texture information which corresponds to the greyscale intensity or the color values of facial images.

We use AAMs [23] to accurately track the speaker's face and extract visual speech features from it. Active appearance models, which were first used for AV-ASR in [22], are generative models of object appearance and have proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. Their distinctive difference relative to image transform-based methods based on DCT/PCA/DWT/ICA of the raw face image pixels, is that AAMs explicitly capture separately the shape and texture variation of the face [3]. In particular, in the AAM scheme an object's shape is modeled as a wireframe mask defined by a set of landmark points $\{\mathbf{z}_i, i = 1, \ldots, L\}$, whose coordinates constitute a shape vector

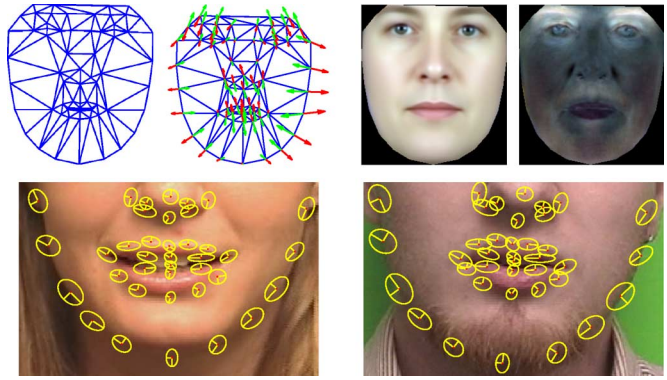

Fig. 4. Visual front-end. Top-left: mean shape s_0 and the first eigenshape s_1, which is illustrated with arrows denoting departure from the mean shape. Top-right: mean texture A_0 and the first eigentexture A_1. Bottom: tracked face shape and feature point uncertainty.

$\mathbf{s}$ of length $2L$. We allow for deviations from the mean shape $\mathbf{s}_0$ by letting $\mathbf{s}$ lie in a linear n-dimensional subspace, yielding

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i. \quad (21)$$

The deformation of the shape $\mathbf{s}$ from the mean shape $\mathbf{s}_0$ defines a mapping $\mathbf{z} \mapsto \mathbf{W}(\mathbf{z}; \mathbf{p})$, $\mathbf{z}$ standing for any point in the interior of the mean shape, which brings the face exemplar on the current image frame I into registration with the mean face template. After canceling out shape deformation, the face texture registered with the mean face can be modeled as a weighted sum of m "eigentextures" $\{A_i\}$, i.e.,

$$I(\mathbf{W}(\mathbf{z}; \mathbf{p})) = A_0(\mathbf{z}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{z}) \quad (22)$$

where A_0 is the mean face texture. Both eigenshape and eigentexture bases are learned during a training phase, using a representative set of hand-labeled face images [23]. The training set shapes are first aligned and then a principal component analysis (PCA) of these aligned shapes yields the main modes of shape variation $\{s_i\}$. Similarly, the leading principal components of the training set texture vectors constitute the eigentexture set $\{A_i\}$. The mean shape/texture and the first shape/texture eigenvector extracted by such a procedure are visualized in the upper part of Fig. 4.

Given a trained AAM, model fitting amounts to finding for each video frame I the shape and texture parameters $\tilde{\mathbf{p}} \equiv \{\mathbf{p}, \boldsymbol{\lambda}\}$ which minimize the penalized error functional

$$f(\tilde{\mathbf{p}}) = \frac{\kappa}{2\sigma^2} \|E(\tilde{\mathbf{p}})\|^2 + Q(\tilde{\mathbf{p}}) \quad (23)$$

where $E(\tilde{\mathbf{p}}) \equiv I(\mathbf{W}(\mathbf{p})) - A_0 - \sum_{i=1}^{m} \lambda_i A_i$ is the model's texture reconstruction error image, σ^2 is the variance of the reconstruction error, $Q(\tilde{\mathbf{p}}) = (1/2)(\tilde{\mathbf{p}} - \tilde{\mathbf{p}}_0)^T \boldsymbol{\Sigma}_{\tilde{\mathbf{p}},0}^{-1} (\tilde{\mathbf{p}} - \tilde{\mathbf{p}}_0)$ is a quadratic penalty corresponding to a Gaussian coefficient prior with mean $\tilde{\mathbf{p}}_0$ and covariance matrix $\boldsymbol{\Sigma}_{\tilde{\mathbf{p}},0}$, and κ is a positive parameter which adjusts the share of the prior and reconstruction error terms in the AAM fitting criterion. Efficient, real-time, iterative algorithms for solving this nonlinear least

squares problem and obtaining the best estimate for $\tilde{\mathbf{p}}$ can be found in [23], [39], [40]. The covariance matrix $\mathbf{\Sigma}_{\tilde{\mathbf{p}}}$ in the least-squares estimate for $\tilde{\mathbf{p}}$ is related to the Hessian matrix of the error functional $f(\tilde{\mathbf{p}})$, evaluated at its minimum [41, ch. 15] and can be efficiently obtained as a by-product of the fitting process [40]. In our audiovisual fusion experiments, we consider the least-squares AAM solution $\tilde{\mathbf{p}}$ as an unbiased measurement for the visual features. We also consider the measurement noise uncertainty Gaussian and use $\mathbf{\Sigma}_{\tilde{\mathbf{p}}}$ as its covariance matrix. In the notation of Section II-B, we thus have for the visual stream $(s \equiv v)$ $\mathbf{y}_v = \tilde{\mathbf{p}}$, $\boldsymbol{\mu}_{e,v} = 0$, and $\mathbf{\Sigma}_{e,v} = \mathbf{\Sigma}_{\tilde{\mathbf{p}}}$. We employ a face detector [42] to initialize face tracking or help recover it in case of failure, rendering the visual feature extraction process fully automatic.

A novel aspect of our visual front-end which differentiates it from previous AAM-based implementations for AV-ASR [22], [43] is that we use a cascade of two AAMs. The first, full-face AAM spans the whole face area, as shown in the upper part of Fig. 4, and can reliably *track* the speaker in long video sequences. However, this is not particularly appropriate for visual speech feature extraction, since visual speech-related information is mostly confined in the lower-half part of the face. Therefore, we also use a second ROI-AAM which covers the face area around the mouth, as depicted in the lower part of Fig. 4, and is used to *analyze* the ROI's shape and texture. Since the ROI-AAM covers too small an area to allow for stable tracking, we pinpoint it with the full-face AAM. As visual feature vector for speech recognition we use at each new video frame I_t the analysis parameters $\tilde{\mathbf{p}}_t$ of the ROI-AAM along with their uncertainty estimates computed as described above. Plots of the corresponding landmark positions and their localization uncertainty ellipses for two example video frames are illustrated in Fig. 4.

Since we are interested in speaker-independent AV-ASR, deriving visual speech features with good speaker invariance properties has been a particular concern in our visual front-end design. Active appearance models trained with the conventional procedure described above on annotated datasets depicting multiple persons, as has been done in [22], are deficient in this respect, because AAM modeling is expended on representing the extensive appearance variability across different speakers instead of concentrating on the speech-induced intra-person variability. Using feature mean subtraction [3] can only partly alleviate this deficiency because it cannot cancel the fact that the leading PCA modes selected during training mostly account for speaker identity rather than visual speech variability. To address this issue, we allow speaker-dependent mean shape $\mathbf{s}_0$ and texture A_0 vectors in our AAM-based facial analysis front-end. In practice, in the ROI-AAM training phase we subtract person-specific (as distinct from global) shape and texture means from the annotated dataset. We also modify the AAM feature extraction by subtracting an estimate of the speaker's mean shape and texture before analyzing with the mouth ROI-AAM. In the experiments reported in Section VI, we have found it adequate to use as such estimates just the average of the speaker's shape and texture over ten video frames at the beginning of each subject's recording, with 1-s delay between the considered frames. In the context of AV-ASR, we term this modified AAM model a *visemic AAM*, since its leading modes of shape and texture variation are directly related to visual speech and are thus more immune to variability across speakers. A similar approach has been applied in conjunction with image transform-based visual analysis techniques [44], but the lack of explicit control on facial shape deformation can make it less effective than with AAMs. A more thorough study of person-independent visual feature extraction for facial analysis, which will include a more detailed analysis of our visemic AAM technique, as well as an extensive comparison with other methods will be included in another paper under preparation.

B. Audio Feature Extraction and Uncertainty Estimation

With some notable exceptions, e.g., [18], most AV-ASR research to date has studied the performance gain of audiovisual fusion in comparison to relatively simple audio-only systems. Since AV-ASR is mostly motivated for speech recognition applications under noisy acoustic conditions, it is important to examine the effectiveness of AV-ASR systems in conjunction with more advanced noise-robust audio front-ends.

From the extensive recent literature on noise-robust audio-only ASR, we have integrated into our AV-ASR system the technique of [14]. Their approach fits especially well in our framework since it addresses both speech enhancement and computation of uncertainty estimates of the enhanced audio-only features in a unified manner. Following [14], our audio features correspond to the log-filter energies of a Mel-scale filterbank applied on the audio signal, which we subsequently refer to as FBANK representation. Assuming an additive time-domain noise model, the noise degradation process in the FBANK audio feature domain can be effectively modeled by

$$\mathbf{y}_a = \mathbf{x}_a + \mathbf{g}(\mathbf{n}_a - \mathbf{x}_a) + \mathbf{e}_a \tag{24}$$

where $\mathbf{y}_a$, $\mathbf{x}_a$, and $\mathbf{n}_a$ are the FBANK features corresponding to the degraded audio signal, the clean audio signal, and the noise, respectively. The modeling error of the approximation $\mathbf{e}_a$ is assumed zero-mean Gaussian with variance Ψ, while $\mathbf{g}(\mathbf{n}_a - \mathbf{x}_a) = \log(1 + \exp(\mathbf{n}_a - \mathbf{x}_a))$. Since the $\mathbf{g}(\mathbf{n}_a - \mathbf{x}_a)$ term in (24) is nonlinear with respect to $\mathbf{x}_a$, as in [14], we iteratively take a zero-order Taylor approximation of it around a current estimate of $\hat{\mathbf{x}}_a^{(j)}$, to obtain $\mathbf{y}_a = \mathbf{x}_a + \mathbf{g}^{(j)} + \mathbf{e}_a$, where $\mathbf{g}^{(j)} = \mathbf{g}(\mathbf{n}_a - \hat{\mathbf{x}}_a^{(j)})$. We also assume that an M-component GMM trained on clean speech is available. This GMM is described by the mean vectors and covariance matrices $\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m$ and the prior probabilities π_m. Combining the linearized feature degradation model of (24) with the clean speech GMM yields the improved enhanced audio feature estimate

$$\hat{\mathbf{x}}_a^{(j+1)} = \sum_{m=1}^{M} r_m^{(j)} \left[\mathbf{\Sigma}_m^{-1} \boldsymbol{\mu}_m + \Psi^{-1} \left(\mathbf{y}_a - \mathbf{g}^{(j)} \right) \right] \tag{25}$$

where $r_m^{(j)} \propto \pi_m N(\mathbf{y}_a; \boldsymbol{\mu}_m + \mathbf{g}^{(j)}, \mathbf{\Sigma}_m + \Psi)$ is the assignment probability of the audio feature to the mth clean speech GMM mixture component after the jth iteration of the enhancement process. Upon convergence, we obtain the final enhanced audio estimate $\hat{\mathbf{x}}_a$ along with its accompanying uncertainty $\mathbf{\Sigma}_{e,a}$, given in [14, Eq. (25)]. We refer to [14] for further details and extensions of the method. The obtained noisy-clean difference

vector $\boldsymbol{\mu}_{e,a} \equiv \mathbf{y}_a - \hat{\mathbf{x}}_a$ and the measurement uncertainty $\boldsymbol{\Sigma}_{e,a}$ correspond to the audio stream quantities $\boldsymbol{\mu}_{e,s}$ and $\boldsymbol{\Sigma}_{e,s}$ in (7) and describe the audio feature degradation process, which we consider Gaussian. We can then straightforwardly integrate the audio enhancement vector $\boldsymbol{\mu}_{e,a}$ and its uncertainty $\boldsymbol{\Sigma}_{e,a}$ into our audiovisual fusion scheme.

C. Synchronous and Asynchronous Integration Models

Although our discussion so far has focused on multimodal fusion using simple GMMs for static data and state-synchronous Multistream-HMMs for dynamic data, our uncertainty compensation scheme has a much wider applicability and is compatible with more general sequence modeling architectures for asynchronous audiovisual modeling. The audio and visual speech streams are often naturally sampled at different frame rates or are only loosely synchronized [3]. Human speech perception has adapted to these challenges; for example, human speech-reading performance is robust to large artificial delays between the audiovisual streams [45]. Moreover, traditional unimodal HMMs cannot naturally handle the inherently different categorization of audio and visual primitive units into phonemes and visemes, respectively.

A number of multimodal integration techniques have been recently developed to address these issues. Depending on the stage at which the audio and visual streams are fused, one can generally classify these approaches into three main categories, namely early, intermediate, and late integration techniques [46], ranging from methods that enforce strict stream alignment to methods that process each stream independently. Intermediate integration techniques, which allow moderate asynchrony between the modalities, are perhaps best suited for modeling audiovisual speech. Successful representative intermediate integration approaches are the state-asynchronous Multistream-HMMs [18], Product-HMMs [19], [20], Asynchronous-HMMs [47], and various dynamic Bayesian network alternatives which have been investigated in the context of audiovisual speech recognition in [21]. Our adaptive fusion by uncertainty compensation scheme can be seamlessly integrated with these multimodal fusion architectures; in particular, in Section VI, we also present AV-ASR experiments employing our scheme in conjunction with Product-HMMs.

VI. Experiments

The proposed scheme for fusion by uncertainty compensation has been evaluated with audiovisual speech recognition experiments.

A. Dataset and Evaluation Methodology

We have used the Clemson University audiovisual experiments (CUAVE) database [48] on which we have performed digit classification experiments. The experiments are performed on the "Normal" part of the database comprising audiovisual recordings of 36 (17 female and 19 male) speakers uttering 50 isolated English digits each. The speakers in this part of the database are facing the camera and are standing relatively still. Video recordings have been performed under good illumination conditions at 720 × 480 pixels resolution and at 29.97-Hz frame rate; one representative image frame from each of the speakers is shown in Fig. 5. For the tests in noise, the audio recordings in the testing subset have been contaminated with additive babble noise from the NOISEX-92 database at various SNR levels.

Fig. 5. Sample frames from all 36 CUAVE database subjects.

Recognition performance is tested on data from six speakers, while the recordings of the remaining 30 speakers have been used for training the digit models. Since the CUAVE dataset is relatively small compared to audio-only corpora, we have performed all our experiments multiple times using different splits of the database into test/training sets in order to increase the statistical significance of our results. More specifically, we have partitioned our dataset into six nonoverlapping subsets, each corresponding to the six speakers of a single row in Fig. 5. Then, we have used each of the subsets in rotation as test set, training the models on the remaining five subsets. This yields a total of six repetitions of our experiments on independent test sets. The audio- and visual-only recognition results we report in Section VI-B have been averaged over these six repetitions, while for the audiovisual recognition experiments of Section VI-C we have retained the first subset for determining the best stream weights and thus the reported results have been averaged over the remaining five repetitions.

As audio features, we use log-filter energies of a Mel-scale filterbank applied on the audio signal (FBANK representation). Specifically, we extract 26 FBANK coefficients from 25-ms Hamming windowed frames of the preemphasized (factor: 0.97) audio signal at a rate of 100 Hz. As visual features we use the AAM coefficients of mouth-ROI visemic AAMs, computed as described in Section V-A. To match the audio frame rate, the visual features have been upsampled from the video frame rate of 29.97 to 100 Hz by simple linear interpolation. In all our experiments, derivative and acceleration parameters accompany both audio and visual features. Also, in all cases we use whole-digit left-to-right hidden Markov models, each with eight states and with a single diagonal-covariance Gaussian observation probability distribution per stream and per state. All models have been trained once on clean speech before testing under different noise conditions. Our experiments have been carried out using the HMM Toolkit (HTK) [49], which we

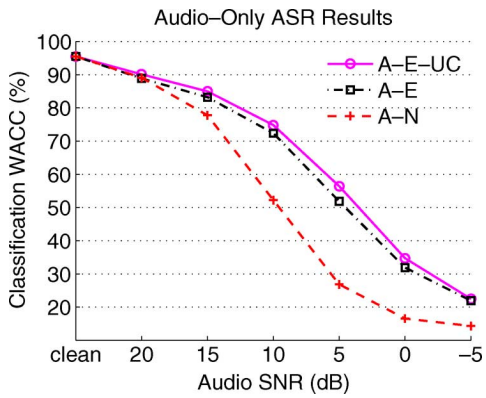

Fig. 6. Audio-only digit classification results for various babble noise SNR levels. We compare between using the raw noisy audio features (A-N), the enhanced audio features (A-E), and the enhanced audio features decoded with uncertainty compensation (A-E-UC).

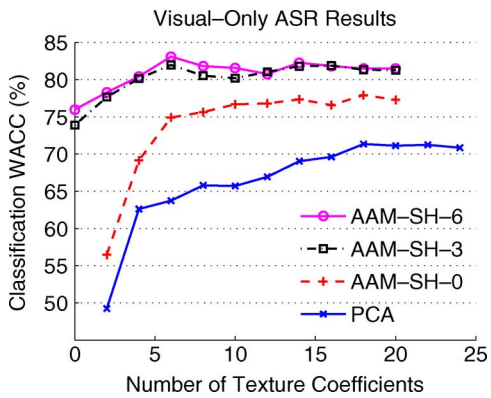

Fig. 7. Visual-only digit classification results for AAM and PCA visual features for varying number of texture coefficients. For the AAM features we also show how classification performance depends on the number of shape coefficients.

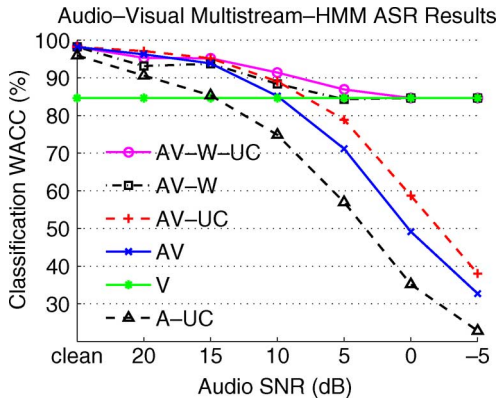

Fig. 8. Multistream-HMM audiovisual digit classification results at various babble noise SNR levels. We depict word accuracy results for the following methods: enhanced audio with uncertainty compensation (A-UC); visual-only (V); audiovisual (AV); audiovisual with uncertainty compensation (AV-UC); audiovisual with weights (AV-W); and audiovisual with weights and uncertainty compensation (AV-W-UC). In all experiments involving audio we have used the enhanced audio features. Active appearance model features have been used for the visual modality.

have modified so as to implement the uncertainty compensation fusion scheme.

B. Single-Modality Speech Recognition Experiments

We first present audio-only and visual-only digit recognition experiments examining the relative performance of different audio and visual front-end configurations.

We start with an audio-only classification experiment which examines the performance in our recognition task of the speech enhancement and uncertainty compensation technique described in Section V-B. In applying the method of [14], we used a clean-speech, 50-mixture GMM of the static FBANK features, trained on all CUAVE database clean recordings. We compare between using the raw noisy audio features (A-N) and the enhanced audio features (A-E and A-E-UC). The uncertainty estimates provided by the enhancement process are ignored in conventional decoding (A-E), while they are incorporated into the decision process in uncertainty compensated decoding (A-E-UC). The results summarized in Fig. 6 demonstrate that using the unprocessed noisy features leads to very poor recognition performance in low SNR levels. Using enhanced features is thus crucial in sustaining good performance, while uncertainty compensation provides a significant additional improvement. For example, at 5-dB SNR, word accuracy (WACC) after enhancement increases by roughly 25% absolute, while uncertainty compensation gives an additional 5% gain. In all our audiovisual ASR experiments reported next we will thus use the enhanced audio feature set.

We subsequently examine the relative performance of different visual front-end variants in a visual-only experiment. To compare our visemic AAM-based technique with alternative image-transform-based visual feature extraction methods, we have also extracted PCA visual features from the same mouth ROI area. Localization for both the AAM and PCA masks has been supplied by the full-face AAM. The mean shape and texture of the AAM, as well as the mean texture of the PCA feature extraction technique have both been updated for each speaker, as described in Section V-A, to increase the speaker-independence of the extracted features. In Fig. 7, we summarize the results obtained by the two alternative methods for varying number of retained texture coefficients. For the AAM case we give three plots, corresponding to retaining 0, 3, and 6 shape coefficients. Our visemic AAM with six shape and six texture coefficients performs overall the best (83% WACC), while the maximum performance of the PCA-based technique is 71% and achieved for 18 texture coefficients. What is particularly remarkable is the recognition capacity of visemic AAM models using very few AAM parameters. For example, using just three shape and no texture AAM coefficients yields 74% WACC, which surpasses the performance of the 18-coefficient PCA model; this should be attributed to the increased specificity of the proposed visemic AAM speaker adaptation algorithm. Our work is the first to demonstrate superior AV-ASR performance for the AAM features. In the previous study of [43], in which the AAM features were outperformed by simpler PCA-like image transform features, full-face AAMs where used for both facial analysis and tracking, while no mechanism for speaker invariance was applied. Our cascaded pair of AAMs (one for robust tracking and one for mouth-ROI analysis) and the proposed visemic AAM mechanism for speaker invariance seem to

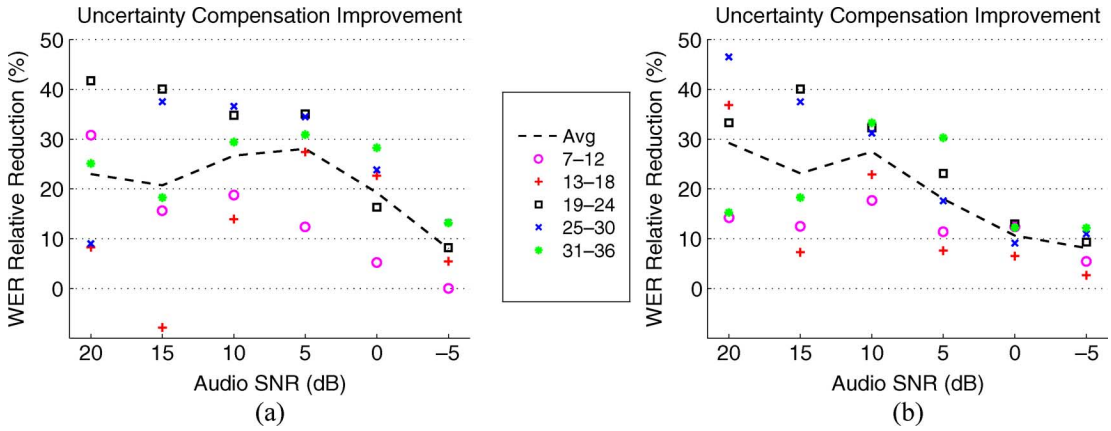

Fig. 9. Performance gain due to uncertainty compensation in Multistream-HMM audiovisual digit classification for various babble noise SNR levels and all five repetitions of the experiment over different test sets. We show the relative word error rate reduction when using uncertainty compensation. (a) Without using stream weights, i.e., AV-UC over AV. (b) With stream weights, i.e., AV-W-UC over AV-W. In all cases the enhanced audio features have been used.

effectively address both shortcomings of previous AAM-based techniques for AV-ASR and suggest that model-based computer-vision approaches can be particularly effective for visual speech facial feature extraction.

The audiovisual experiments reported next use the best-performing six-shape/six-texture visemic AAM visual feature set.

C. Audiovisual Speech Recognition Experiments

Having studied the performance of each modality separately, we present next our main set of audiovisual speech recognition experiments examining the performance of the uncertainty compensation fusion scheme, both with and without stream weighting. In all experiments the enhanced audio features are used. In Fig. 8, we plot the performance of the best audio-only result using uncertainty compensation (A-UC) (corresponding to the A-E-UC label in Fig. 6) and the best visual-only result (V) and compare them with the performance of four audiovisual state-synchronous multistream-HMM fusion variants: audiovisual with equal weights for the two streams and conventional decoding (AV); equal-weight audiovisual with uncertainty compensation decoding (AV-UC); audiovisual with optimized weights (AV-W); and audiovisual with optimized weights and uncertainty compensation decoding (AV-W-UC).

To illustrate the performance improvement due to uncertainty compensation alone, we show in Fig. 9 the relative reduction in word error rate (WER) when comparing AV-UC to AV (no stream weights) and AV-W-UC to AV-W (with stream weights); the relative WER reduction is given by $(\mathrm{WER}_{\mathrm{AV}} - \mathrm{WER}_{\mathrm{AV-UC}})/\mathrm{WER}_{\mathrm{AV}}$. As described in Section VI-A, all results are fivefold averages over different repetitions of the experiments with independent test subsets. For the experiments including weights we have used stream exponents summing to 1 and exhaustively searched at each noise level for the audio weight between 0.0 and 1.0 (in steps of 0.1) which yielded the best results on a reserved experiment repetition (comprising the first six speakers as test set). The best audio stream weight for the 0 and −5-dB noise levels turned out to be 0.0, meaning that the corresponding AV-W and AV-W-UC values in Fig. 8 coincide with the visual-only result. Since focusing on the improvement due to uncertainty compensation fusion makes sense only when both streams are active, the 0- and −5-dB noise level values in Fig. 9(b) have been obtained after setting the audio stream weight to 0.1, i.e., its minimum positive value.

Comparing between the AV and AV-UC results, we see that fusion by uncertainty compensation gives a consistent improvement for all acoustic conditions (4.8% mean absolute WACC improvement or 20.9% relative WER reduction averaged over all noisy conditions) over conventional decoding. Similarly consistent improvement is obtained when we combine uncertainty decoding with stream weighting (2.3% mean absolute WACC improvement or 19.4% relative WER reduction averaged over all noisy conditions), as can be seen by comparing AV-W with AV-W-UC. Stream weights are necessary for keeping audiovisual recognition performance above visual-only performace at very low SNRs; this can be attributed to an overestimation of the confidence in the feature estimate by the audio enhancement method. The best multistream-HMM audiovisual results in Fig. 8 are obtained with the AV-W-UC scheme which improves the WACC over the best audio-only recognition (A-UC) by an absolute 28.7% on average over all six noise levels.

To increase our confidence in the statistical significance of the improved audiovisual fusion results due to uncertainty compensated decoding, we show in Fig. 9 not only the average relative WER reduction, but also all the individual results for each of the five repetitions of the experiment on the disjoint test sets. Such comparisons across many experiment repetitions allow one to draw statistically safer conclusions about the relative performance of two competing techniques, since the variability in the results due to inter-speaker differences is reduced [50]–[52]. We see that the improvement in multimodal fusion due to uncertainty decoding is consistent over the repetitions of the experiments on independent test sets, both when we use stream weights or not. This fact further strengthens the statistical validity of our arguments.

Our last experiment investigates the performance of uncertainty decoding in conjunction with Product-HMMs, which, as discussed in Section V-C, better account for audio and visual

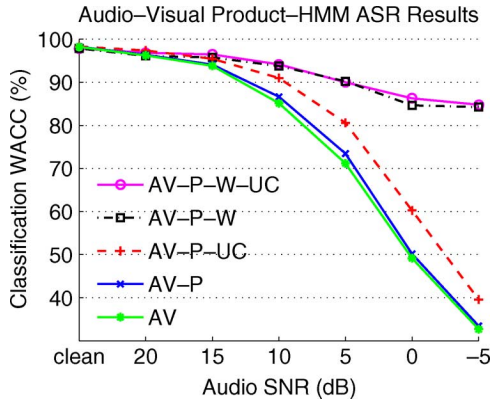

Fig. 10. Product-HMM-based audiovisual digit classification results for various babble noise SNR levels. We show recognition word accuracy results for four (weighted) Product-HMM variants, two with conventional decoding (AV-P and AV-P-W) and two with uncertainty decoding (AV-P-UC and AV-P-W-UC). The Multistream-HMM/conventional decoding (AV) results are also given for comparison. In all cases enhanced audio features and AAM visual features have been used.

speech asynchrony effects. In Fig. 10, we show Product-HMM results with conventional decoding (AV-P) and uncertainty decoding (AV-P-UC), their variants using stream weights (AV-P-W and AV-P-W-UC), as well as the state-synchronous Multistream-HMM with conventional decoding (AV) result as baseline. Using uncertainty decoding gives an (average over all noise levels) absolute WACC gain of 5.0% in the case of equal weight models (AV-P-UC vs. AV-P) and 0.6% when using stream weights (AV-P-W-UC versus AV-P-W). The average absolute WACC improvement of Product-HMMs over Multistream-HMMs is 1.0% when using conventional decoding and 1.2% with uncertainty compensated decoding. In total, our best audiovisual recognition results are obtained with the AV-P-W-UC model.

All reported experiments show a consistent improvement in recognition rates when using uncertainty compensation during decoding. Particularly noteworthy is the fact that adaptive fusion with uncertainty compensation integrates transparently with proven multimodal analysis techniques, such as stream weighting or Product-HMMs. In previous work [10] we have also demonstrated a further small improvement when considering visual feature uncertainty estimates also during model training. Uncertainty compensation thus proves to be a flexible and reliable tool in a wide range of multimodal fusion contexts.

VII. Conclusion

We have presented a novel framework for multimodal fusion by uncertainty compensation and demonstrated its effectiveness in audiovisual ASR. Given an estimate of each stream's feature uncertainty, the proposed framework naturally leads to highly adaptive multimodal fusion rules which are easy and efficient to implement. Our technique is widely applicable and can be transparently integrated with either synchronous or asynchronous multimodal sequence integration architectures typically encountered in multimodal applications. We have further shown that our scheme is compatible with the widely used stream-weights formulation; combination of both techniques consistently yields the best results in our AV-ASR experiments.

Appendix
EM Training for HMMs Under Uncertainty

For continuous-density HMMs modeling emission probabilities with M mixtures of Gaussians, similarly to the GMM case covered in Section IV, the expected complete-data log-likelihood $Q(\boldsymbol{\theta},\boldsymbol{\theta}') = E[\log p(\mathcal{Y},\{\mathcal{Q},\mathcal{X},\mathcal{M}\}|\boldsymbol{\theta})|\mathcal{Y},\boldsymbol{\theta}']$ of the parameters $\boldsymbol{\theta}$ in the EM algorithm's current iteration given the previous guess $\boldsymbol{\theta}'$ is obtained in the E-step as

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\theta}') = &\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\log a_{q_{t-1}q_t}P(\mathcal{Y},q|\boldsymbol{\theta}') \\ &+\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\int\log p(\mathbf{y}_t|\mathbf{x}_t,q_t,\boldsymbol{\theta}')P(\mathcal{Y},q,\mathbf{x}_t|\boldsymbol{\theta}')d\mathbf{x}_t \\ &+\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\sum_{m=1}^{M}\int\log p(\mathbf{x}_t|m_t,q_t,\boldsymbol{\theta}') \\ &\qquad\times P(\mathcal{Y},q,m,\mathbf{x}_t|\boldsymbol{\theta}')d\mathbf{x}_t \\ &+\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\sum_{m=1}^{M}p(m|q_t,\boldsymbol{\theta}')P(\mathcal{Y},q,m|\boldsymbol{\theta}') \\ &+\sum_{\mathbf{q}\in\mathcal{Q}}\log\pi_{q_0}P(\mathcal{Y},q|\boldsymbol{\theta}'). \end{aligned} \tag{26}$$

The responsibilities $\gamma_t(i,k) = p(q_t = i, m = k|\mathbf{y}_{1:t}) \propto a_t(i)\beta_t(i)$ are estimated via a forward–backward procedure [53] modified so that uncertainty compensated scores are used

$$\begin{aligned} a_{t+1}(j) &= P(\mathbf{y}_{1:t}, q_t = j|\boldsymbol{\theta}') \\ &= \left[\sum_{i=1}^{N}\alpha_{ij}a_t(i)\right]b'_j(\mathbf{y}_{t+1}) \end{aligned} \tag{27}$$

$$\begin{aligned} \beta_t(i) &= P(\mathbf{y}_{t+1:T}|q_t = i,\boldsymbol{\theta}') \\ &= \sum_{j=1}^{N}\alpha_{ij}b'_j(\mathbf{y}_{t+1})\beta_{t+1}(j) \end{aligned} \tag{28}$$

where $b'_j(\mathbf{y}_t) = \sum_{m=1}^{M}\rho_m N(\mathbf{y}_t;\boldsymbol{\mu}'_{j,m}+\boldsymbol{\mu}_{e_t},\boldsymbol{\Sigma}'_{j,m}+\boldsymbol{\Sigma}_{e_t})$. Scoring is done similarly to the conventional case by the forward algorithm, i.e., $P(\mathbf{y}_{1:T}|\boldsymbol{\theta}) = \sum_{i=1}^{N}a_T(i)$. The updated parameters $\boldsymbol{\theta}$ are estimated using formulas similar to the GMM case in Section IV. In particular, for updating $\boldsymbol{\mu}_{q,m},\boldsymbol{\Sigma}_{q,m}$ in the M-step, the filtered estimate for the observation is used as in (15) and (16).

Acknowledgment

The authors would like to thank A. Potamianos for providing an initial experimental setup for AV-ASR, G. Gravier for his extensive feedback on an early manuscript, particularly regarding Section III-C, I. Kokkinos for visual front-end discussions, K. Murphy for making his HMM toolkit publicly available, and J. N. Gowdy for providing the CUAVE database. They would also like to thank the associate editor and the anonymous reviewers for their comments and suggestions which have considerably improved the paper.

REFERENCES

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

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

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[4] J. Clark and A. Yuille, *Data Fusion for Sensory Information Processing*. Norwell, MA: Kluwer, 1990.

[5] J. R. Movellan and P. Mineiro, "Robust sensor fusion: Analysis and application to audio visual speech recognition," *Mach. Learn.*, vol. 32, pp. 85–100, 1998.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[7] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[8] A. Katsamanis, G. Papandreou, V. Pitsikalis, and P. Maragos, "Multimodal fusion by adaptive compensation for feature uncertainty with application to audiovisual speech recognition," in *Proc. Eur. Signal Process. Conf.*, 2006.

[9] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 2458–2461.

[10] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition," in *Proc. IEEE Int. Workshop Multimedia Signal Proc.*, 2007, pp. 264–267.

[11] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[12] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[13] N. B. Yoma, F. McInnes, and M. Jack, "Weighted matching algorithms and reliability in noise canceling by spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, vol. 2, pp. 1171–1174.

[14] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[15] N. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 158–166, Mar. 2002.

[16] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Commun.*, vol. 34, pp. 25–40, 2001.

[17] N. B. Yoma, C. Molina, C. Garreton, and F. Huenupan, "Uncertainty in signal estimation and stochastic weighted Viterbi algorithm: A unified framework to address robustness in speech recognition and speaker verification," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech Education and Publishing, 2007, ch. 12, pp. 187–218.

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

[19] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 1, pp. 169–172.

[20] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Int. Conf. Human Lang. Technol. Res.*, 2002, pp. 1–6.

[21] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1274–1288, 2002.

[22] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[24] Q. Huo and C. Lee, "A Bayesian predictive approach to robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 200–204, Nov. 2000.

[25] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[26] W. A. Fuller, *Measurement Error Models*. New York: Wiley, 1987.

[27] D. Massaro and D. Stork, "Speech recognition and sensory integration," *Amer. Sci.*, vol. 86, no. 3, pp. 236–244, 1998.

[28] 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. Stork and M. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 461–471.

[29] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 337–351, Sep. 1996.

[30] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Commun.*, vol. 26, pp. 149–161, 1998.

[31] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 1, pp. 173–176.

[32] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, no. 11, pp. 1260–1273, 2002.

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

[34] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 6, pp. 3733–3736.

[35] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, vol. 11, pp. 49–52.

[36] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. (B)*, vol. 39, no. 1, pp. 1–38, 1977.

[38] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[39] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.

[40] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognition*, 2008.

[41] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[42] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition*, 2001, vol. 1, pp. 511–518.

[43] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. Int. Conf. Multimedia and Expo.*, 2001, pp. 825–828.

[44] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2003, pp. 260–267.

[45] Q. Summerfield, "Lipreading and audio-visual speech perception," *Phils. Trans.: Biol. Sci.*, vol. 335, no. 1273, pp. 71–78, 1992.

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

[47] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2003, vol. 15, pp. 1237–1244.

[48] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. 2017–2020.

[49] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.2)," Cambridge Univ. Eng. Dept., 2002, Tech. Rep..

[50] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1989, vol. 1, pp. 532–535.

[51] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. 409–412.

[52] M. Keller, S. Bengio, and S. Wong, "Benchmarking non-parametric statistical tests," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2005, vol. 18, pp. 651–658.
[53] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

George Papandreou (S'03) received the Diploma in electrical and computer engineering (with highest honors) from the National Technical University of Athens (NTUA), Athens, Greece, in 2003, where he is currently working towards the Ph.D. degree.

Since 2003, he has been a Graduate Research Assistant at the NTUA, participating in national and European research projects in the areas of computer vision and audiovisual speech analysis. During the summer of 2006, he visited Trinity College Dublin, Dublin, Ireland, working on image restoration. From 2001 to 2003, he was an Undergraduate Research Associate with the Institute of Informatics and Telecommunication, Greek National Center for Scientific Research "Demokritos," participating in projects on wireless Internet technologies. His research interests are in image analysis, computer vision, and multimodal processing. His published research in these areas includes work on image segmentation with multigrid geometric active contours (accompanied with an open-source software toolbox), image restoration for cultural heritage applications, human face image analysis, and multimodal fusion for audiovisual speech processing.

Athanassios Katsamanis (S'03) received the Diploma in electrical and computer engineering (with highest honors) in 2003 from the National Technical University of Athens (NTUA), Athens, Greece, where he is currently pursuing the Ph.D. degree.

He is currently a Graduate Research Assistant at the NTUA. From 2000 to 2002, he was an Undergraduate Research Associate with the Greek Institute for Language and Speech Processing (ILSP), participating in projects in speech synthesis, signal processing education, and machine translation. During the summer of 2002, he worked on Cantonese speech recognition at the Hong Kong Polytechnic University, while in the summer of 2007 he visited Télécom Paris (ENST) working on speech production modeling. His research interests lie in the area of speech analysis and include speech production, synthesis, recognition, and multimodal processing. In these domains and in the frame of his Ph.D. degree and European research projects, since 2003 he has worked on multimodal speech inversion, aeroacoustics for articulatory speech synthesis, speaker adaptation for non-native speech recognition, and multimodal fusion for audiovisual speech recognition.

Vassilis Pitsikalis (S'02-M'08) received the Diploma in electrical and computer engineering and the Ph.D degree, both from the National Technical University of Athens (NTUA), Athens, Greece, in 2001 and 2007, respectively.

Since 2008, he has been a Postdoctoral Research Associate at the NTUA. During his studies, he has participated as a Graduate Research Assistant in several National and European research projects in the areas of nonlinear speech processing and automatic speech recognition (ASR). During the spring semester of 2002, he visited as a Research Assistant Lucent Technologies, Murray Hill, NJ. His research interests are in the areas of speech analysis and recognition and include fractal speech processing and analysis, robust speech recognition, and multistream and multimodal fusion and recognition.

Petros Maragos (S'81–M'85–SM'91–F'95) received the Diploma in electrical engineering from the National Technical University of Athens (NTUA) in 1980 and the M.Sc.E.E. and Ph.D. degrees from Georgia Institute of Technology (Georgia Tech), Atlanta, in 1982 and 1985, respectively.

In 1985, he joined the faculty of the Division of Applied Sciences, Harvard University, Cambridge, MA, where he worked for eight years as a Professor of electrical engineering. In 1993, he joined the faculty of the Department of Electrical and Computer Engineering, Georgia Tech. During parts of 1996 to 1998, he was on sabbatical and academic leave working as a Director of Research at the Institute for Language and Speech Processing in Athens. Since 1998, he has been working as a Professor at the NTUA School of Electrical and Computer Engineering. His research and teaching interests include signal processing, systems theory, pattern recognition, informatics, communications, and their applications to image processing and computer vision, speech and language processing, and multimedia. He has served as editorial board member for the journals *Signal Processing* and *Visual Communications and Image Representation*, as General Chairman or Co-Chair of conferences or workshops (VCIP'92, ISMM'96, VLBV'01, MMSP'07), and as member of IEEE Signal Processing Society committees. He recently coedited a book on multimodal processing and interaction.

Prof. Maragos received the 1987 NSF Presidential Young Investigator Award, the 1988 IEEE Signal Processing Society Young Author Paper Award for the paper "Morphological Filters," the 1994 IEEE Signal Processing Society Senior Award and the 1995 IEEE W. R. G. Baker Prize Award for the paper "Energy Separation in Signal Modulations with Application to Speech Analysis," the 1996 Pattern Recognition Society's Honorable Mention Award for the paper "Min-Max Classifiers," and the 2007 EURASIP Technical Achievements Award for contributions to nonlinear signal processing and systems theory, image processing, and speech processing. He has served as an Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.