4

Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audio-Visual Speech Recognition

George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos

National Technical University of Athens, Athens, Greece

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 chapter we discuss the effects of feature measurement uncertainty on classification and learning rules. Such an approach can be 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, this framework leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We further show that more traditional multimodal fusion methods relying on stream weights fall under this scheme under certain assumptions; this provides novel insights into their applicability for various tasks and suggests new practical ways for estimating the stream weights adaptively. The potential of the approach is demonstrated in audiovisual speech recognition experiments using either synchronous or asynchronous models.

4.1 Multimodal Fusion: Benefits and Challenges

Motivated by the multimodal way humans perceive their environment, complementary information sources have been successfully utilized in many applications. Such a case is audiovisual speech recognition (AV-ASR) [413], where fusing visual and audio cues can lead to improved performance in comparison to audio-only recognition, especially in the presence of audio noise.

However, successfully integrating heterogeneous information streams is challenging, mainly because multimodal schemes need to adapt to dynamic environmental conditions, which can dissimilarly affect the reliability of the separate modalities by contaminating feature measurements with noise. For

P. Maragos et al. (eds.), *Multimodal Processing and Interaction*,
DOI: 10.1007/978-0-387-76316-3_4, © Springer Science+Business Media, LLC 2008

example, the visual stream in AV-ASR should be discounted when the visual front-end momentarily mistracks the speaker's face.

A common theme in many stream integration methods is the utilization of stream weights to equalize the different modalities. These weights operate as exponents to each stream's probability density and have been employed in fusion tasks of different audio streams [344] and audiovisual integration [147, 412]. Such stream weights have been applied not only in conventional Hidden Markov Models, but also in conjunction with Dynamic Bayesian Network architectures which better account for the asynchronicity of audiovisual speech [362]. Despite its favorable experimental properties, stream weighting requires setting the weights for the different streams; although various methods have been proposed for this purpose [184], a rigorous approach to adapt the stream weights is still missing.

In this chapter, building on the recent work of [248, 404, 389], we approach the problem of adaptive multimodal fusion by explicitly taking feature measurement uncertainty of the different modalities into account, both during model training and testing. In single modality scenarios, modeling feature noise has proven fruitful for noise-robust ASR [135, 442, 577, 130] and has been further pursued in applications such as speaker verification [578] and multi-band ASR [344]. We show in a probabilistic framework how multimodal learning and classification rules should be adjusted to account for feature measurement uncertainty. Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) are discussed in detail and modified algorithms for classification and EM maximum-likelihood estimation under uncertainty are derived. Uncertainty compensation leads to adaptive multimodal fusion rules which are widely applicable and easy to implement. We demonstrate that previous stream weight-based multimodal fusion formulations can be derived from the uncertainty-aware scheme under certain assumptions; this unveils their probabilistic underpinnings and provides novel insights into their applicability for various tasks. In this context, new practical ways for estimating stream weights adaptively are suggested. Regarding audiovisual speech, we describe techniques to extract uncertainty estimates for the visual and audio features and evaluate the method in AV-ASR experiments utilizing multistream HMM, demonstrating improved performance. Applying the proposed technique in conjunction with Product HMMs (P-HMM) [147, 312], which better account for cross-modal asynchrony, can yield further improvements.

4.2 Feature Uncertainty and Multimodal Fusion

Let us consider a pattern classification scenario. We measure a property (feature) of a pattern instance and try to decide to which of N classes $c_i, i = 1 \dots N$ it should be assigned. The measurement is a realization x of a random variable 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 functions $p(x|c_i), i = 1 \ldots N$. Our decision is then based on some proper rule, *e.g.*, the Maximum A Posteriori (MAP) criterion $\hat{c} = \text{argmax}\, p(c_i|x) = \text{argmax}\, p(x|c_i)p(c_i)$.

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/) [413]. Better choice of features and modeling schemes can reduce this uncertainty. Second, *parameter estimation uncertainty* that mainly originates from insufficient training. Using the Bayesian Predictive Classification rule can possibly alleviate it [220]. Third, *feature observation uncertainty* due to errors in the measurement process or noise contamination. This is the type of uncertainty we mainly address in this chapter.

4.2.1 Feature Observation Uncertainty and its Compensation in Classification

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

$$Y = X + E\,, \tag{4.1}$$

which is graphically depicted in Fig. 4.1 and assume that the noise density $p_E(e)$ is known. This scenario of contaminated measurements corresponds to the so-called *measurement error* models in statistics [172]. Under the observation model of Eq. (4.1), classification decisions must rely on $p(c_i|y) \propto p_Y(y|c_i)p(c_i)$, and thus $p_Y(y|c_i)$ needs to be computed.

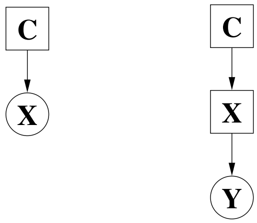

Fig. 4.1. Pictorial representation of feature measurement scenarios, with hidden variables denoted by squares and observed by circles. *Left*: Conventional case – we observe the features x directly. *Right*: Noisy measurement case – we only observe noisy features y.

To determine the desirable noisy feature probability density function $p_Y(y|c_i)$, we need to integrate out the clean feature variable x

$$p_Y(y|c_i) = \int p_X(x|c_i) p_E(y-x)\, dx. \tag{4.2}$$

Although the integral in Eq. (4.2) is in general intractable, we can obtain a closed-form solution in the important special case of Gaussian data model, $p_X(x|c_i) = N(x; \mu_i, \Sigma_i)$, with Gaussian observation noise, $p_E(e) = N(e; \mu_e, \Sigma_e)$. Then one can show that $p_Y(y|c_i)$ is given by

$$p_Y(y|c_i) = N(y; \mu_i + \mu_e, \Sigma_i + \Sigma_e), \tag{4.3}$$

implying that we can proceed by considering our features y clean, provided that we shift the model means by μ_e and increase the model covariances Σ_i by Σ_e. A similar approach has been previously followed in [442, 578, 130].

To illustrate Eq. (4.3), we discuss with reference to Fig. 4.2 how observation uncertainty influences decisions in a simple 2-class classification task. The two classes are modeled by 2D spherical Gaussian distributions, $N(\mu_1, \sigma_1^2 I)$, $N(\mu_2, \sigma_2^2 I)$ and they have equal prior probability. If our observation y contains zero mean spherical Gaussian noise with covariance matrix $\sigma_e^2 I$ then the modified decision boundary consists of those y for which $N(y; \mu_1, \sigma_1^2 I + \sigma_e^2 I) = N(y; \mu_2, \sigma_2^2 I + \sigma_e^2 I)$. When σ_e^2 is zero, the decision should be made as in the clean 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.

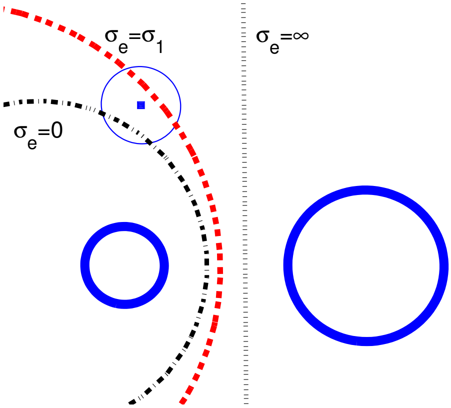

Fig. 4.2. Decision boundaries for classification of a noisy observation (square marker) in two classes, shown as circles, for various observation noise variances. Classes are modeled by spherical Gaussians of means μ_1, μ_2 and variances $\sigma_1^2 I$, $\sigma_2^2 I$ respectively. The decision boundary 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.

4.2.2 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 $x_s, s = 1, \ldots, S$. If the features are statistically independent given the class label c, the conditional probability of the full observation vector $x_{1:S} \equiv (x_1; \ldots ; x_S)$ is given by the product rule; application of Bayes' formula yields the class label probability given the features:

$$p(c|x_{1:S}) \propto p(c) \prod_{s=1}^{S} p(x_s|c) \ . \tag{4.4}$$

In an attempt to improve classification performance, several authors have introduced stream weights w_s as exponents in Eq. (4.4), resulting in the modified expression

$$b(c|x_{1:S}) = p(c) \prod_{s=1}^{S} p(x_s|c)^{w_s} \ , \tag{4.5}$$

which can be seen in a logarithmic scale as a weighted average of individual stream log-probabilities. Such schemes have been motivated by potential differences in reliability among different information streams, and larger weights are assigned to information streams with better classification performance. Using such weighting mechanisms has been experimentally proven to be beneficial for feature integration in both intra-modal (*e.g.*, multiband audio [344]) and inter-modal (*e.g.*, audiovisual speech recognition [147, 184, 362]) scenarios.

The stream weights formulation is however unsatisfactory in various respects. From a theoretical viewpoint, the weighted score b in Eq. (4.5) no longer has the probabilistic interpretation of Eq. (4.4) as class probability given the full observation vector $x_{1:S}$. Therefore it becomes unclear how to conceptually define, let alone implement, standard probabilistic operations, such as integrating-out a variable x_s (in the case of missing features), or conditioning the score on some other available information. From a more practical standpoint, it is not straightforward how 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. Although there have been some efforts towards dynamically adjustable stream weights [184], they are not rigorously justified and are difficult to generalize.

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 x_s, our decision mechanism depends on the noisy version $y_s = x_s + e_s$ of the underlying quantity. The probability of interest is thus obtained by integrating out the hidden clean features x_s, *i.e.*,

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \int p(x_s|c) p(y_s|x_s) dx_s \ . \tag{4.6}$$

In the common case that the clean feature emission probability is modeled as a Gaussian mixture model (GMM), *i.e.*,

$$p(x_s|c) = \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(x_s; \mu_{s,c,m}, \Sigma_{s,c,m}), \tag{4.7}$$

and the observation noise at each stream is considered independent across streams and Gaussian, $p(y_s|x_s) = N(y_s; x_s + \mu_{e,s}, \Sigma_{e,s})$, it directly follows that

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(y_s; \mu_{s,c,m} + \mu_{e,s}, \Sigma_{s,c,m} + \Sigma_{e,s}) \ , \tag{4.8}$$

which, as in the single-stream case (4.3), involves considering our features y_s clean, while shifting the model means by $\mu_{e,s}$, and increasing the model covariances $\Sigma_{s,c,m}$ by $\Sigma_{e,s}$. Using mixtures of Gaussians for the measurement noise $p(y_s|x_s)$ is straightforward and could be useful in case of heavy-tailed noise distribution or for modeling observation outliers. Also note that, although the measurement noise covariance matrix $\Sigma_{e,s}$ of each stream is the same for all classes c and all mixture components m, noise particularly affects the most peaked mixtures, for which $\Sigma_{e,s}$ is substantial relative to the modeling uncertainty due to $\Sigma_{s,c,m}$. The adaptive fusion effect of feature uncertainty compensation in a simple 2-class classification task using two streams is illustrated in Fig. 4.3.

Although Eq. (4.8) is conceptually simple and easy to implement, given an estimate of the measurement noise variance $\Sigma_{e,s}$ of each stream, it actually constitutes a highly adaptive rule for multisensor fusion. To appreciate this, and also to show how our scheme is related to the stream weights formulation of Eq. (4.5), we examine a particularly illuminating special case of our result. We make two simplifying assumptions:

1. The measurement noise covariance is a scaled version of the model covariance, *i.e.*, $\Sigma_{es} = r_{s,c,m} \Sigma_{s,c,m}$ for some positive constant $r_{s,c,m}$ interpreted as the relative measurement error. Intuitively, as the SNR for the s-stream drops, the corresponding relative measurement error $r_{s,c,m}$ increases.
2. For every stream observation 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.

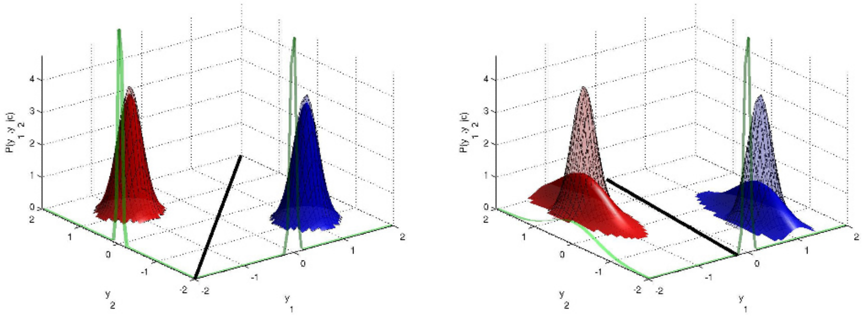

Fig. 4.3. Multimodal variance compensation leads to adaptive fusion. Figures describe a 2-class classification scenario, using two Gaussian feature streams, y_1 and y_2, with equal model covariances $\Sigma_{s,c} = \sigma^2$. The 1-D plots on the y_1 and y_2 axes represent the measurement uncertainty in the corresponding stream. *Left*: Conventional negligible measurement uncertainty scenario; the decision boundary lies on the axes' diagonal. *Right*: Significant measurement noise at the y_2 stream, $\Sigma_{e,2} \gg \Sigma_{e,1}$, in which case $p(y_S|c)$ (solid surfaces) differ significantly from $p(x_S|c)$ (transparent surfaces); the decision boundary moves and classification is mostly influenced by the reliable y_1 stream.

Under these conditions the Gaussian densities in Eq. (4.8) can be approximated by $N(y_s; \mu_{s,c,m_0} + \mu_{es}, (1 + r_{s,c,m_0})\Sigma_{s,c,m_0})$; using the power-of-Gaussian identity $N(x; \mu, w^{-1}\Sigma) = (\det(w(2\pi\Sigma)^{w-1}))^{1/2} N(x; \mu, \Sigma)^w \propto N(x; \mu, \Sigma)^w$ yields

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \left[\tilde{\rho}_{s,c,m_0} N(y_s; \mu_{s,c,m_0} + \mu_{e,s}, \Sigma_{s,c,m_0}) \right]^{w_{s,c,m_0}} , \qquad (4.9)$$

where

$$w_{s,c,m_0} = 1/(1 + r_{s,c,m_0}) \qquad (4.10)$$

is the *effective stream weight* and $\tilde{\rho}_{s,c,m_0}$ is a properly modified mixture weight which is independent of the observation 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; however they do not need to satisfy a sum-to-one constraint $\sum_{s=1}^{S} w_{s,c,m_0} = 1$, as is conventionally considered by other authors.

This is an appealing result. Our framework unveils the probabilistic assumptions under stream weight-based formulations; furthermore, Eq. (4.10) provides a rigorous mechanism to select for each new measurement y_s and uncertainty estimate $(\mu_{e,s}, \Sigma_{e,s})$ all involved stream weights *fully adaptively*, *i.e.*, with respect to both class label c and mixture component m.

4.3 Uncertainty in Expectation-Maximization Training

In many real-world applications requiring big volumes 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 important to take training set feature uncertainty into account in learning procedures.

4.3.1 GMM Training Under Uncertainty

Under our feature uncertainty viewpoint, only a noisy version y of the underlying true property x can be observed. Maximum-likelihood estimation of the GMM parameters θ from a training set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ under the EM algorithm [129] should thus consider the corresponding clean features $\mathcal{X}$, besides the class memberships $\mathcal{M}$, as hidden variables. The expected complete-data log-likelihood $Q(\theta, \theta') = E[\log p(\mathcal{Y}, \{\mathcal{X}, \mathcal{M}\}|\theta)|\mathcal{Y}, \theta']$ of the parameters θ in the EM algorithm's current iteration given the previous guess θ' 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

$$\begin{aligned} Q(\theta, \theta') = &\sum_{i=1}^{N} \sum_{m=1}^{M} \log \pi_m p(m|y_i, \theta') + \\ &\sum_{i=1}^{N} \sum_{m=1}^{M} \int \log p(y_i|x_i) p(x_i, m|y_i, \theta') dx_i + \\ &\sum_{i=1}^{N} \sum_{m=1}^{M} \int \log p(x_i|m, \theta) p(x_i, m|y_i, \theta') dx_i \; . \end{aligned} \tag{4.11}$$

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

$$r_m = \sum_{i=1}^{N} r_{i,m}, \quad \pi_m = \frac{r_m}{N}, \quad \mu_m = \frac{1}{r_m} \sum_{i=1}^{N} r_{i,m} \hat{x}_{i,m},$$

$$\Sigma_m = \frac{1}{r_m} \sum_{i=1}^{N} r_{i,m} \left(\Sigma_{x_{i,m}} + (\hat{x}_{i,m} - \mu_m)(\hat{x}_{i,m} - \mu_m)^T \right) \; , \tag{4.12}$$

where (the prime denotes previous-step parameter estimates)

$$r_{i,m} = p(m|y_i, \theta') \propto \pi'_m N(y_i; \mu'_m + \mu_{e,i}, \Sigma'_m + \Sigma_{e,i}) \tag{4.13}$$

$$\hat{x}_{i,m} = \Sigma_{x_{i,m}} \left((\Sigma'_m)^{-1} \mu'_m + (\Sigma_{e,i})^{-1} (y_i - \mu_{e,i}) \right), \tag{4.14}$$

$$\Sigma_{x_{i,m}} = \left((\Sigma'_m)^{-1} + (\Sigma_{e,i})^{-1} \right)^{-1}. \tag{4.15}$$

Two important differences w.r.t. the noise-free case are notable: *first*, error-compensated scores are utilized in computing the responsibilities $r_{i,m}$ in Eq. (4.13); *second*, in updating the model's means and variances, one should replace the noisy measurements y_i used in conventional GMM training with their model-enhanced counterparts, described by the expected value $\hat{x}_{i,m}$ and variance $\Sigma_{x_{i,m}}$. Furthermore, in the multimodal case with multiple streams $s = 1, \ldots, S$, one should compute the responsibilities by $r_{i,m} \propto \pi'_m \prod_{s=1}^{S} N(y_{s,i}; \mu'_{s,m} + \mu_{s,e,i}, \Sigma'_{s,m} + \Sigma_{s,e,i})$, which generalizes Eq. (4.13) and introduces interactions among modalities.

4.3.2 HMM Training Under Uncertainty

For the HMM, similarly to the GMM case just covered, the expected complete-data log-likelihood $Q(\theta, \theta') = E[\log p(O, \{\mathcal{Q}, \mathcal{X}, \mathcal{M}\}|\theta)|O, \theta']$ of the parameters θ in the EM algorithm's current iteration, given the previous guess θ', is obtained in the E-step as:

$$\begin{aligned} Q(\theta, \theta') = &\sum_{q \in \mathcal{Q}} \sum_{t=1}^{T} \log a_{q_{t-1}q_t} P(O, q|\theta') + \\ &\sum_{q \in \mathcal{Q}} \sum_{t=1}^{T} \int \log p(o_t|x_t, q_t, \theta') P(O, q, x_t|\theta') dx_t + \\ &\sum_{q \in \mathcal{Q}} \sum_{t=1}^{T} \sum_{m=1}^{M} \int \log p(x_t|m_t, q_t, \theta') P(O, q, m, x_t|\theta') dx_t + \\ &\sum_{q \in \mathcal{Q}} \sum_{t=1}^{T} \sum_{m=1}^{M} p(m|q_t, \theta') P(O, q, m|\theta') + \sum_{\mathbf{q} \in \mathcal{Q}} \log \pi_{q_0} P(O, q|\theta') \, . \end{aligned} \tag{4.16}$$

The responsibilities $\gamma_t(i, k) = p(q_t = i, m = k)$ are estimated via a forward-backward procedure [420] modified so that uncertainty compensated scores are utilized:

$$a_{t+1}(j) = P(o_{1:t}, q_t = j|\theta') = \Big[\sum_{i=1}^{N} \alpha_{ij} a_t(i)\Big] b'_j(o_{t+1}) \tag{4.17}$$

$$\beta_t(i) = P(o_{t+1:T}|q_t = i, \theta') = \sum_{j=1}^{N} \alpha_{ij} b'_j(o_{t+1}) \beta_{t+1}(j), \tag{4.18}$$

where $b'_j(o_t) = \sum_{m=1}^{M} \rho_m N(o_t; \mu'_{j,m} + \mu_{e_t}, \Sigma'_{j,m} + \Sigma_{e_t})$. Scoring is done similarly to the conventional case by the forward algorithm, *i.e.*, $P(O|\theta) = \sum_{i=1}^{N} a_T(i)$. The updated parameters θ are estimated using formulas similar to the GMM case in Section 4.3.1. For $\mu_{q,m}, \Sigma_{q,m}$ the filtered estimate for the observation is used as in (4.12).

4.3.3 Some Insights into Training Under Uncertainty

Focusing on the simpler GMM model and similarly to the analysis in Section 4.2, we can gain insight into the previous EM formulas by considering the special case of constant and model-aligned errors $\Sigma_{e,i} = \Sigma_e = \lambda_m \Sigma_m$. Then, after convergence, the covariance formula in Eq. (4.12) can be written as

$$\Sigma_m = \frac{1}{1+\lambda_m}\tilde{\Sigma}_m, \quad \text{or, equivalently,} \quad \Sigma_m = \tilde{\Sigma}_m - \Sigma_e \ , \tag{4.19}$$

where we just subtract from the conventional (non-compensated) covariance estimate $\tilde{\Sigma}_m = \frac{1}{r_m}\sum_{i=1}^{N} r_{i,m}(y_i - \mu_m)(y_i - \mu_m)^T$ the noise covariance Σ_e. The rule in Eq. (4.19) has been used before as heuristic for fixing the model covariance estimate after conventional EM training with noisy data (*e.g.*, [117]). We see that it is justified in the constant and model-aligned errors case; otherwise, one should use the more general rules in Eq. (4.12).

Another link of our training under uncertain measurements scenario is to neural network training with noise (or noise injection) [487], where an original training set is artificially supplemented with multiple noisy instances of it and the resulting enriched set is used for training. Monte-Carlo-based noise injection training should be contrasted to the analytic integration over the noise distribution suggested by our approach. Our interpretation thus shows that noise injection can be motivated under the noisy measurements viewpoint. Training with noise is also related to Tikhonov regularization [65] and is known to be relatively immune to over-fitting, thus leading to classifiers with improved generalization ability. Similar advantageous properties should be expected for our training under uncertain measurements technique.

4.4 Audio-Visual Speech Recognition

A challenging application domain for multimodal fusion schemes is Audio-visual 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 naturally integrated with multi-stream HMMs or other multimodal sequence processing techniques and clearly improve their performance in AV-ASR.

4.4.1 Visual Front-End

Salient visual speech information can be obtained from the shape and the texture (intensity/color) of the speaker's visible articulators, mainly the lips and the jaw, which constitute the *Region Of Interest* (ROI) around the mouth [413].

We use *Active Appearance Models* (AAM) [107] of faces to accurately track the speaker's face and extract visual speech features from it, capturing both

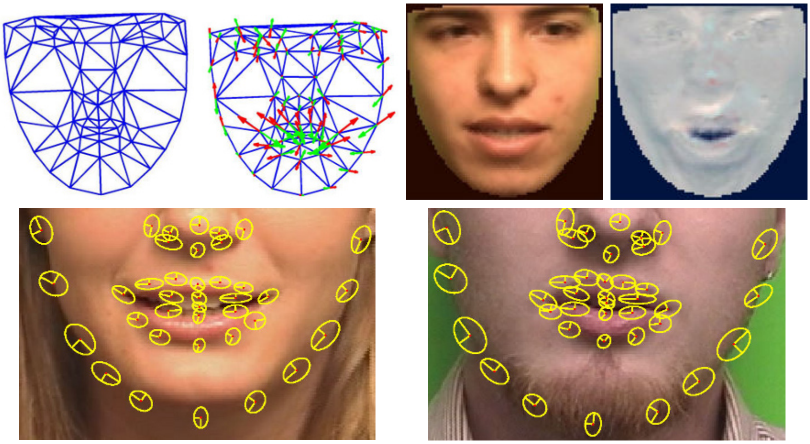

Fig. 4.4. Visual Front-End. *Upper-Left*: Mean shape s_0 and the first eigenshape s_1. *Upper-Right*: Mean texture A_0 and the first eigenface A_1. *Lower*: Tracked face shape and feature point uncertainty.

the shape and the texture of the face. AAM, which were first used for AV-ASR in [329], are generative models of object appearance and have proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. In the AAM scheme an object's shape is modeled as a wireframe mask defined by a set of landmark points $\{x_i, i = 1 \dots N\}$, whose coordinates constitute a shape vector s of length $2N$. We allow for deviations from the mean shape s_0 by letting s lie in a linear n-dimensional subspace, yielding $s = s_0 + \sum_{i=1}^{n} p_i s_i$. The deformation of the shape s to the mean shape s_0 defines a mapping $W(x;p)$, which brings the face exemplar on the current frame I into registration with the mean face template. After canceling out shape deformation, the face appearance (color values) registered with the mean face can be modeled as a weighted sum of "eigenfaces" $\{A_i\}$, *i.e.*, $I(W(x;p)) \approx A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, where A_0 is the mean texture of faces. Both eigenshape and eigenface bases are learned during a training phase. The first few of them extracted by such a procedure are depicted in Fig. 4.4.

Given a trained AAM, model fitting amounts to finding for each video frame I_t the parameters $\tilde{p}_t \equiv \{p_t, \lambda_t\}$ which minimize the squared texture reconstruction error $I_t(W(p_t)) - A_0 - \sum_{i=1}^{m} \lambda_{t,i} A_i$; efficient iterative algorithms for this non-linear least squares problem can be found in [107]. The fitting procedure employs a face detector [158] to get an initial shape estimate for the first frame. To extract information mostly related to visual speech, we utilize a hierarchy of two AAM. The first *ROI-AAM* spans only the area around the mouth and is used to analyze in detail the ROI's shape and texture; however, the ROI-AAM covers too small an area to allow for reliable tracking. To pinpoint the ROI-AAM we use a second *Face-AAM* which spans the whole face and can reliably track the speaker in long video sequences. As visual feature vector for speech recognition we use the parameters $\tilde{p}_t$ of the fitted

ROI-AAM. We employ as uncertainty in the visual features the uncertainty in estimating the parameters of the corresponding non-linear least squares problem [415, Chapter 15]; plots of the corresponding uncertainty in localizing the landmarks on the image for two example faces are illustrated in Fig. 4.4.

4.4.2 Audio Front-End

We use the Mel Frequency Cepstral Coefficients (MFCC) to represent audio, as it is common in contemporary ASR systems. Uncertainty is considered to originate from additive noise to the audio waveform. To get estimates of the clean features we employ the speech enhancement framework proposed in [130], adapted to work with MFCCs along the lines of [186]. The enhanced features are derived from the noisy ones by iteratively improving a guess based on a prior clean speech model and Vector Taylor Series approximation [171]. The uncertainty of the resulting clean feature estimates is assumed to be zero-mean Gaussian and for each such feature estimate a rough approximation of its uncertainty is also available at the output of the enhancement module. In this way, fusion by uncertainty compensation is facilitated. Alternative enhancement procedures could equivalently be applied provided that the variance of the enhanced features could also be roughly estimated.

4.4.3 Experiments and Discussion

The novel fusion approach proposed above is evaluated via classification experiments on the Clemson University Audiovisual Experiments (CUAVE) database [391]. Experiments are performed on the section of the database comprising audiovisual recordings of 36 speakers uttering 50 isolated digits each. The speakers are standing naturally still and they are framed including their shoulders and head, as shown in Fig. 4.5. Digit models are trained on data from 30 speakers who have been randomly selected. The rest of the data is held out for testing. For the tests in noise, the audio recordings in this testing subset have been contaminated with babble noise from the NOISEX-92 database at various SNR levels.

Mel frequency cepstral coefficients (MFCC) are extracted from 25 ms Hamming windowed frames of the preemphasized (factor: 0.97) audio stream at a rate of 100 Hz. Per audio frame, 13 coefficients are extracted. A visual feature vector is estimated per video frame, consisting of 6 shape and 12 texture features and the visual feature stream is upsampled from the video frame rate (29.97 FPS) to the audio rate of 100 Hz by linear interpolation. Mean Normalization is applied to both the audio and visual features.

To demonstrate the benefits of compensating for feature uncertainty for multimodal fusion we performed a series of digit classification experiments and the results are summarized in Fig. 4.6. For these experiments, the first derivatives of the audio and visual features have also been included in the corresponding feature vectors. Uncertainty estimates for the visual features

Fig. 4.5. Sample speaker images from the CUAVE database.

are acquired as discussed in Section 4.4.1. For the audio features, uncertainty is computed as the squared difference between each feature and the corresponding clean feature, which is considered to be available as well in this proof-of-concept scenario.

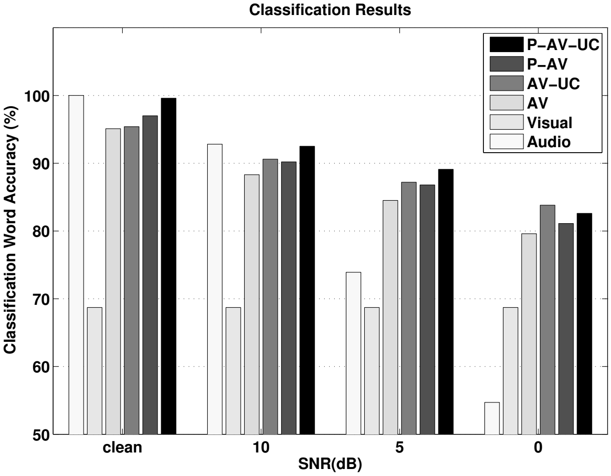

Fig. 4.6. Classification results with or without Uncertainty Compensation (UC) for fusion. Simple multistream models (AV) and product-HMMs (P-AV) have been evaluated at various SNR levels.

Audiovisual observations are modeled by digit left-right multistream Hidden Markov Models (AV), each with 8 states and with a single multidimensional Gaussian observation probability distribution per stream and per state. Single modality 8-state digit HMMs have also been evaluated for reference. Further, to better account for asynchrony between the modalities, these single-modality HMMs have been merged in product-HMMs (P-AV) as described in

[312]. Asynchrony has been limited to two states only, while stream weights are assumed to be equal to unity in all cases (multistream or product-HMMs). The multimodal models have been evaluated both with and without uncertainty compensation. Compensation has been implemented in the HMM decoder by increasing the observation variance in the modified forward algorithm described in Section 4.3.3.

Models with uncertainty compensation in general outperform those without. The best overall performance is demonstrated by the uncertainty compensated product HMMs (P-AV-UC), which at 5 dB SNR yields 89.1% accuracy, an absolute 2.3% over the conventionally decoded product HMM. The corresponding results for state-synchronous multi-stream HMMs are 87.2% for uncertainty compensated decoding and 84.5% for conventional decoding. We see that accounting for uncertainty clearly favors multimodal fusion, by approximately 2.5% absolute, and has a cumulative beneficial effect when combined with asynchrony modeling through product HMMs, which give another 2% absolute accuracy improvement. As expected, the beneficial effect of uncertainty compensation gets increasingly important for decreasing audio SNR.

In a separate series of experiments we evaluate uncertainty compensation for fusion in the training phase. The compensated models are trained on clean audio data, while for the visual training data their corresponding variances are taken into account into the modified EM algorithm of Section 4.3.3. This time, both the first and the second derivatives of the audiovisual features are also utilized. Testing with uncertainty compensation is implemented as before. In this case however we have utilized more realistic estimates of the uncertainty of the audio features following the procedure sketched in Section 4.4.2 Our experimental results summarized in Table 4.1 show that accounting for uncertainty in the case of audiovisual fusion, either solely in testing or both in training and testing, AV-UC and AV-UCT, respectively, improves AV-ASR performance in most cases. Again, for the baseline audiovisual setup we used multistream HMMs with stream weights equal to unity for both streams. The proposed approach (AV-UC, AV-UCT) seems particularly effective at lower SNRs.

4.5 Conclusions

The chapter has shown that taking the feature uncertainty into account constitutes a fruitful framework for multimodal feature analysis tasks. This is especially true in the case of multiple complementary information streams, where having a good estimate of each stream's uncertainty at a particular moment facilitates information fusion, allowing for proper training and fully adaptive stream integration schemes. In order for this approach to reach its full potential, reliable methods for dynamically estimating the feature observation uncertainty are needed. Ideally, the methods that we employ to extract

SNR	A	V	AV	AV-UC	AV-UCT
clean	99.3	75.7	90.0	-	-
15 dB	96.7	-	88.0	88.3	88.0
10 dB	91.3	-	88.3	88.7	87.7
5 dB	82.0	-	87.0	88.0	87.7
0 dB	62.7	-	84.3	87.0	87.3
-5 dB	40.3	-	81.7	82.0	83.0

Table 4.1. Word Percent Accuracy (%) of classification experiments on CUAVE database for various noise levels on the audio stream; experiments have been conducted for: Audio (A), Visual (V) and Audio-Visual (AV) features, with stream weights equal to unity, with Uncertainty Compensation in the testing phase (UC), and with Uncertainty Compensation both in the testing and training (UCT).

features in pattern recognition tasks should accompany feature estimates with their respective errorbars. Although some progress has been done in the area, further research is needed before we fully understand the quantitative behavior under diverse conditions of popular features commonly used in pattern analysis tasks such as speech recognition.