Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition

George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos
School of Electrical and Computer Engineering
National Technical University of Athens
Email: {gpapan, nkatsam, vpitsik, maragos}@cs.ntua.gr

***Abstract*— We study the effect of uncertain feature measurements and show how classification and learning rules should be adjusted to compensate for it. Our approach is particularly fruitful in multimodal fusion scenarios, such as audio-visual speech recognition, where multiple streams of complementary features whose reliability is time-varying are integrated. For such applications, by taking the measurement noise uncertainty of each feature stream into account, the proposed framework leads to highly adaptive multimodal fusion rules for classification and learning which are widely applicable and easy to implement. We further show that previous multimodal fusion methods relying on stream weights fall under our 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 our approach is demonstrated in audio-visual speech recognition experiments.**

I. Introduction

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) [1], where fusing visual and audio cues can lead to improved performance relatively to audio-only recognition, especially in the presence of audio noise. However, successfully integrating heterogeneous information streams is challenging, mainly because of the need for adaptation to dynamic environmental conditions, which dissimilarly affect the reliability of the separate modalities.

Using stream weights to equalize the different modalities is common to many stream integration methods. Stream weights operate as exponents to each stream's probability density and have been employed in fusion tasks of different audio streams [2] and audio-visual integration [3]. Despite its favorable experimental properties, the technique requires setting the weigths for the different streams; although various methods have been proposed for this purpose [4], a rigorous approach to stream weight adaptation is still missing.

In our work we approach the problem of adaptive multimodal fusion by explicitly taking feature measurement uncertainty of the different modalities into account. In single modality scenarios, modeling feature noise has proven fruitful for ASR [5], [6] and has been further pursued for applications such as speaker verification [7]and speech enhancement [8]. We show in a rigorous 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 EM algorithms for training are derived. Our approach leads to adaptive multimodal fusion rules which are widely applicable and easy to implement. This paper extends our previous work [9], [10] by considering the effect of uncertain features not only during decoding, but also during model training.

II. Multimodal Fusion by Uncertainty Compensation

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, application of Bayes' formula yields the class label probability given the full observation vector $x_{1:S} \equiv (x_1; \ldots; x_S)$:

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

In an attempt to improve classification performance, several authors have introduced stream weights w_s as exponents in Eq. (1), yielding

$$b(c|x_{1:S}) = p(c) \prod_{s=1}^{S} p(x_s|c)^{w_s}, \tag{2}$$

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 [2]) and inter-modal (*e.g.* audio-visual speech recognition [4], [11]) scenarios.

The stream weights formulation is however unsatisfactory in various respects as it has been discussed in [9], [10], where we have shown that accounting for feature uncertainty naturally leads to a highly adaptive mechanism for fusion of different information sources. More specifically, we consider a stochastic measurement framework where we do not have direct access to the features x_s and our decision mechanism depends on their noisy version $y_s = x_s + e_s$. The probability of interest is thus obtained by integrating out the hidden clean features x_s, *i.e.*

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \int p(x_s|c) p(y_s|x_s) dx_s. \tag{3}$$

In the common case of clean features modeled with a gaussian mixture model (GMM), $p(x_s|c) = \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(x_s; \mu_{s,c,m}, \Sigma_{s,c,m})$, and gaussian observation noise at each stream, *i.e.* $p(y_s|x_s) = N(y_s; x_s + \mu_{e,s}, \Sigma_{e,s})$ (extension to gaussian mixture noise model is trivial), we have shown in [9] that

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(y_s - \mu_{e,s}; \mu_{s,c,m}, \Sigma_{s,c,m} + \Sigma_{e,s}), \tag{4}$$

which means that in classification we should (1) use the *enhanced* feature estimate $y_s - \mu_{e,s}$, instead of the noisy feature y_s and (2) increase the model covariances $\Sigma_{s,c,m}$ by $\Sigma_{e,s}$. 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

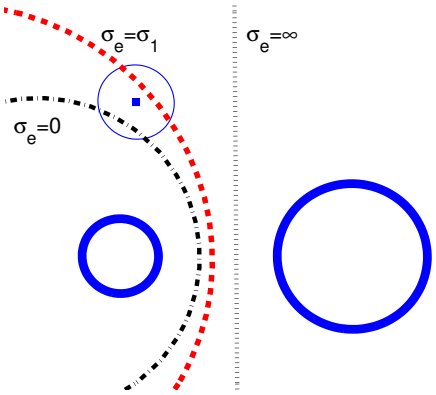

Fig. 1. Decision boundaries for classification of a noisy observation (square marker) in two classes, shown as circles, for various observation noise variances. Classes are modeled by spherical Gaussians of means μ_1, μ_2 and variances $\sigma_1^2 I$, $\sigma_2^2 I$ respectively. The decision boundary is plotted for three values of noise variance (a) $\sigma_e = 0$, (b) $\sigma_e = \sigma_1$, and (c) $\sigma_e = \infty$. With increasing noise variance, the boundary moves away from its noise-free position.

relative to the modeling uncertainty due to $\Sigma_{s,c,m}$. The effect of feature uncertainty compensation in a simple 2-class classification task is illustrated in Fig. 1.

Although Eq. (4) 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. (2), we examine a particularly illuminating special case of our result, when:

1) The measurement noise covariance is a scaled version of the model covariance, *i.e.* $\Sigma_{e,s} = \lambda_{s,c,m}\Sigma_{s,c,m}$ for some positive constant $\lambda_{s,c,m}$ interpreted as the *relative measurement error*.
2) For every stream observation y_s the gaussian mixture response of that stream is dominated by a single component m_0.

Under these conditions, Eq. (4) can be written as [9]

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \Big[\tilde{\rho}_{s,c,m_0} N(y_s - \mu_{e,s}; \mu_{s,c,m_0}, \Sigma_{s,c,m_0}) \Big]^{w_{s,c,m_0}} \tag{5}$$

$$w_{s,c,m_0} = 1/(1 + \lambda_{s,c,m_0}), \tag{6}$$

with w_{s,c,m_0} being *effective stream weights*; $\tilde{\rho}_{s,c,m_0}$ is a properly modified mixture weight, independent of the observation y_s. Note that these effective stream weights are between 0 (for $\lambda_{s,c,m_0} \gg 1$) and 1 (for $\lambda_{s,c,m_0} \approx 0$) 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 and unveils the probabilistic assumptions under stream weight-based formulations. It shows further that our fusion rule in Eq. (4) acts as effectively selecting for each new measurement y_s and uncertainty estimate $(\mu_{e,s}, \Sigma_{e,s})$ corresponding stream weights fully adaptively with respect to both class label c and mixture component m.

III. EM Training Under Uncertainty

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.

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 [12] 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:

$$\begin{aligned} Q(\theta, \theta') = &\sum_{i=1}^{N}\sum_{m=1}^{M} \log \pi_m p(m|y_i, \theta') + \\ &\sum_{i=1}^{N}\sum_{m=1}^{M} \int \log p(y_i|x_i) p(x_i, m|y_i, \theta') dx_i + \\ &\sum_{i=1}^{N}\sum_{m=1}^{M} \int \log p(x_i|m, \theta) p(x_i, m|y_i, \theta') dx_i \end{aligned} \tag{7}$$

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

$$\begin{aligned} &r_m = \sum_{i=1}^{N} r_{i,m}, \quad \pi_m = \frac{r_m}{N}, \quad \mu_m = \frac{1}{r_m}\sum_{i=1}^{N} r_{i,m}\hat{x}_{i,m}, \\ &\Sigma_m = \frac{1}{r_m}\sum_{i=1}^{N} r_{i,m}\big(\Sigma_{x_{i,m}} + (\hat{x}_{i,m} - \mu_m)(\hat{x}_{i,m} - \mu_m)^T\big), \end{aligned} \tag{8}$$

where (the prime denotes previous-step parameter estimates)

$$r_{i,m} = p(m|y_i, \theta') \propto \pi'_m N(y_i - \mu_{e,i}; \mu'_m, \Sigma'_m + \Sigma_{e,i}) \tag{9}$$

$$\hat{x}_{i,m} = \Sigma_{x_{i,m}}\big((\Sigma'_m)^{-1}\mu'_m + (\Sigma_{e,i})^{-1}(y_i - \mu_{e,i})\big), \tag{10}$$

$$\Sigma_{x_{i,m}} = \big((\Sigma'_m)^{-1} + (\Sigma_{e,i})^{-1}\big)^{-1}. \tag{11}$$

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. (9); *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 $r_{i,m} \propto \pi'_m \prod_{s=1}^{S} N(y_{s,i} - \mu_{s,e,i}; \mu'_{s,m}, \Sigma'_{s,m} + \Sigma_{s,e,i})$, which generalizes Eq. (9) and introduces interactions among modalities. Analogous EM formulas for HMM parameter estimation are given in the Appendix.

Similarly to the analysis in Section II, 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. (8) can be written as

$$\Sigma_m = \frac{1}{1 + \lambda_m}\tilde{\Sigma}_m, \quad \text{or, equivalently,} \quad \Sigma_m = \tilde{\Sigma}_m - \Sigma_e, \tag{12}$$

where we simply 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. (12) has been used before as

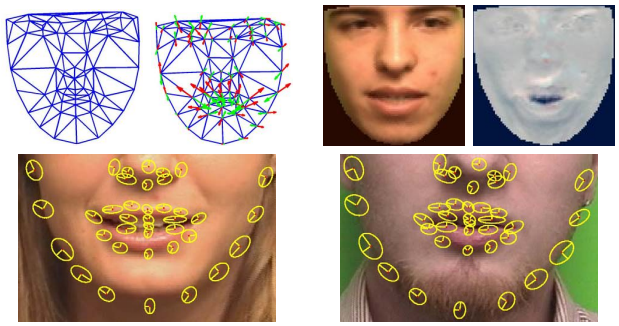

Fig. 2. Visual Front-End. *Upper-Left*: Mean shape s_0 and the first eigenshape s_1. *Upper-Right*: Mean texture A_0 and the first eigenface A_1. *Lower*: Tracked face shape and feature point uncertainty.

heuristic for fixing the model covariance estimate after conventional EM training with noisy data (*e.g.* [13]). We have shown that it is justified in the constant and model-aligned errors case, and only after convergence; otherwise, one should use the more general Eq. (8).

Another link of our training under uncertain measurements scenario is to neural network training with noise (or noise injection) [14], where an original training set is artificially supplemented with multiple noisy instances of it and the resulting enriched set is used for training. Training with noise is relatively immune to over-fitting and leads to classifiers with improved generalization ability.

IV. Audio-Visual Speech Recognition

To demonstrate the applicability of the proposed fusion scheme we apply it in Audio-Visual Automatic Speech Recognition (AV-ASR), a practical problem for which proper information fusion is important.

A. Visual Front-end and Visual Feature Uncertainty

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 [1].

We use *Active Appearance Models* (AAM) [15] of faces to accurately track the speaker's face and extract visual speech features from both its shape and texture. AAM, first used for AV-ASR in [16], 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 color texture 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; see Fig. 2. 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 [15]. The fitting procedure employs a face detector [17] to get an initial shape estimate for the first frame.

As *visual features* for speech recognition we use the parameters $\tilde{p}_t$ of the fitted AAM. We employ as *visual feature uncertainty* the uncertainty in estimating the parameters of the corresponding non-linear least squares problem [18, ch. 15]; plots of the corresponding uncertainty in localizing the landmarks on the image for two example faces are illustrated in Fig. 2.

B. Audio Front-end and Audio Feature Uncertainty

Our audio front-end is based on the Mel Frequency Cepstral Coefficient (MFCC) audio representation. To employ our fusion technique in the presence of auditory noise, we need (see Sec. II) (1) an estimate of the enhanced MFCC features $y_s - \mu_{e,s}$, to use instead of the noisy MFCC features y_s, and (2) the uncertainty $\Sigma_{e,s}$ of the MFCC enhanced estimate. A number of recent *audio-only* approaches to robust ASR have developed techniques which can be used for this purpose; in our experiments we employ the speech enhancement and uncertainty estimation framework proposed in [8]. Following [8], we utilize a prior clean speech model (in the form of GMM in the MFCC space) and a non-linear parametric model of the MFCC feature degradation under noise; based on these, we can iteratively improve an estimate of the enhanced speech and estimate its uncertainty; see [8] for further details. Alternative enhancement procedures could be used provided that they give variance estimates for the enhanced features.

C. Audio-Visual Speech Recognition Experiments

We evaluate our fusion approach in classification experiments on the CUAVE audiovisual database [19]; the considered task is word classification of isolated digits. By contaminating the clean audio signal with babble noise from the NOISEX collection we extended the database including its noisy version. We use MFCCs, along with their first and second order derivatives, as audio features, comprising a 39-dimensional audio vector in total; the corresponding audio feature uncertainty has been computed along the lines of Section IV-B. In the visual front-end, we form a 18-dimensional visual feature vector (6 shape and 12 texture features) and also add up to second derivatives, for a 54-dimensional visual feature vector, which, along with its variance, is computed as discussed in Section IV-A. Mean Normalization has been applied to both the audio and visual features.

For the acoustic and visual observations modeling we constructed 8-state left-right word multi-stream HMMs [1] with a single multi-dimensional Gaussian observation probability distribution per stream at each state. The proposed incorporation of feature uncertainty in the testing phase has been implemented in the HMM decoder by increasing the observation variance in the modified forward algorithm described in the Appendix. The models were trained on clean audio data, while for the visual training data their corresponding variances were taken into account into the modified EM algorithm of the Appendix in the corresponding experiment. The baseline audiovisual setup uses stream weights equal to unity for both streams.

Our experimental results, summarized in Table I, show that: (1) Accounting for feature uncertainty in the case of audiovisual fusion consistently improves accuracy (AV-UC/AV-ACT *vs.* AV). (2) While in heavy noise (SNR under roughly 10 dB) the proposed audio-visual fusion approach outperforms audio-only recognition, in clean conditions (SNR greater than 10 dB) audio-only rates are better (AV-UC/AV-ACT *vs.* A). This could be remedied by using discriminatively set stream weights, in addition to our current uncertainty compensation technique (which would then act as stream adaptation mechanism). Such a combined scheme would clearly further increase speech recognition performance, but we currently cannot justify it theoretically; we leave its further experimental and theoretical study for future work. (3) In agreement with the existing audio-visual

TABLE I

WORD PERCENT ACCURACY (%) OF CLASSIFICATION EXPERIMENTS ON CUAVE DATABASE FOR VARIOUS NOISE LEVELS ON THE AUDIO STREAM: AUDIO (A), VISUAL (V), AUDIO-VISUAL FEATURES, WITH STREAM WEIGHTS EQUAL TO UNITY (AV), WITH UNCERTAINTY COMPENSATION IN TESTING (AV-UC), AND WITH UNCERTAINTY COMPENSATION IN BOTH TESTING AND TRAINING (AV-UCT).

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

ASR literature [1], using visual features clearly leads to noise-robust ASR results; in our experiments audio-visual ASR results show little degradation when audio SNR drops from 15 dB down to 0 dB.

V. CONCLUSIONS

The paper 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 good estimates of each stream's uncertainty facilitates information fusion, allowing for proper training and fully adaptive stream integration schemes. In order this approach to reach its full potential, methods for reliably estimating the feature observation uncertainty are needed. Ideally, the methods that we employ to extract 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.

Acknowledgments: This work was supported by the European Network of Excellence MUSCLE, by the European research programs HIWIRE and ASPI, and by grants ΠΕΝΕΔ-2003 ΕΔ865 & ΕΔ866 [co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)]. We thank A. Potamianos for providing an initial experimental setup and J. N. Gowdy for the CUAVE database. We also appreciate feedback by G. Gravier, I. Kokkinos, and the anonymous reviewers.

APPENDIX

For the HMM, similarly to the GMM covered in Sec. III, 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 EM's current iteration given the previous guess θ' is (E-step):

$$\begin{aligned} Q(\theta,\theta') = &\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\log a_{q_{t-1}q_t}P(O,q|\theta')+ \\ &\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\int \log p(o_t|x_t,q_t,\theta')P(O,q,x_t|\theta')dx_t+ \\ &\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\sum_{m=1}^{M}\int \log p(x_t|m_t,q_t,\theta')P(O,q,m,x_t|\theta')dx_t+ \\ &\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|q_t,\theta')P(O,q,m|\theta') + \sum_{\mathbf{q}\in\mathcal{Q}}\log\pi_{q_0}P(O,q|\theta') \end{aligned} \quad (13)$$

The responsibilities $\gamma_t(i,k) = p(q_t = i, m = k)$ are estimated via a forward-backward procedure [20] modified so that uncertainty compensated scores are utilized:

$$a_{t+1}(j) = P(o_{1:t}, q_t = j|\theta') = \Big[\sum_{i=1}^{N}\alpha_{ij}a_t(i)\Big]b'_j(o_{t+1}) \quad (14)$$

$$\beta_t(i) = P(o_{t+1:T}|q_t = i,\theta') = \sum_{j=1}^{N}\alpha_{ij}b'_j(o_{t+1})\beta_{t+1}(j), \quad (15)$$

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 III. For $\mu_{q,m},\Sigma_{q,m}$ the filtered estimate for the observation is used as in (11).

REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

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

[3] A. Potamianos, E. Sanchez-Soto, and K. Daoudi, "Stream weight computation for multi-stream classifiers," in *Proc. ICASSP*, 2006.

[4] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. ICASSP*, 2001.

[5] 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 TSAP*, pp. 431–442, 1993.

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

[7] N. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE TSAP*, vol. 10, no. 3, pp. 158–166, 2002.

[8] L. Deng, J. Dropo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE TSAP*, vol. 13, no. 3, pp. 412–421, 2005.

[9] A. Katsamanis, G. Papandreou, V. Pitsikalis, and P. Maragos, "Multimodal fusion by adaptive compensation for feature uncertainty with application to audiovisual speech recognition," in *EUSIPCO*, 2006.

[10] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation," in *ICSLP*, 2006, pp. 2458–2461.

[11] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Tr. on Mult.*, vol. 2, no. 3, pp. 141–151, 2000.

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

[13] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," vol. 46, no. 4, pp. 886–902, 1998.

[14] J. Sietsma and R. Dow, "Creating artificial neural networks that generalize," *Neural Networks*, vol. 4, pp. 67–79, 1991.

[15] T. Cootes, G. Edwards, and T. C.J., "Active appearance models," *IEEE PAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[16] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE PAMI*, vol. 24, no. 2, pp. 198–213, 2002.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, vol. I, 2001, pp. 511–518.

[18] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*. Cambridge Univ. Press, 1992.

[19] 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. ICASSP*, 2002.

[20] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.