MULTIMODAL FUSION BY ADAPTIVE COMPENSATION FOR FEATURE UNCERTAINTY WITH APPLICATION TO AUDIOVISUAL SPEECH RECOGNITION

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

National Technical University of Athens
School of Electrical and Computer Engineering
Zografou 15773, Athens, Greece
E-mail: {nkatsam, gpapan, vpitsik, maragos}@cs.ntua.gr
Web: http://cvsp.cs.ntua.gr

ABSTRACT

In pattern recognition one usually relies on measuring a set of informative features to perform tasks such as classification. While the accuracy of feature measurements heavily depends on changing environmental conditions, studying the consequences of this fact has received relatively little attention to date. In this work we explicitly take into account uncertainty in feature measurements and we show in a rigorous probabilistic framework how the models used for classification should be adjusted to compensate for this effect. Our approach proves to be particularly fruitful in multimodal fusion scenarios, such as audio-visual speech recognition, where multiple streams of complementary features are integrated. For such applications, provided that an estimate of the measurement noise uncertainty for each feature stream is available, we show that the proposed framework leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We further show that previous multimodal fusion methods relying on stream weights fall under our scheme if certain assumptions hold; this provides novel insights into their applicability for various tasks and suggests new practical ways for estimating the stream weights adaptively. Preliminary experimental results in audio-visual speech recognition demonstrate the potential of our approach.

1. INTRODUCTION

Motivated by the multimodal way humans perceive their environment, complementary information sources have been successfully utilized in many pattern recognition tasks. Such a case is AudioVisual Automatic Speech Recognition (AV-ASR) [19], 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. Different streams provide complementary information and multimodal schemes should properly elevate the discriminative abilities of each of the modalities. Such schemes should adapt to the effective environmental conditions, which can dissimilarly affect the reliability of the separate modalities by contaminating feature measurements with noise. For example, the visual stream in AV-ASR should be discounted when the visual front-end loses track of the speaker's face.

Our work is supported in part by the European Network of Excellence MUSCLE, the European research project HIWIRE, and the Greek General Secretariat for Research & Technology under program PENED-2003.

A common theme in many stream integration methods is the utilization of stream weights to equalize the different modalities. These 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 [7, 17]. 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 audio-visual speech [15]. Although stream weighting has indisputable benefits as shown experimentally, it requires determining the weights for the different streams; various methods have been proposed for this purpose [9] but a rigorous approach to dynamically adapt the stream weights is still missing.

We choose to explicitly take observation uncertainty of the different modalities into account. Modeling observation noise has proven fruitful in the framework of single modality ASR [5], and has been further pursued for applications such as speech enhancement [22], speaker verification [23], multiband ASR [8, 14], soft decoding for wireless ASR [18], and recently in more advanced enhancement techniques based on noise modeling [4]. In our work, given an estimate of the feature measurement uncertainty, we show in a rigorous probabilistic framework how the models used for classification should be adjusted to compensate for this effect. The proposed scheme leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We also demonstrate that previous stream weight-based multimodal fusion formulations fall under our scheme under certain assumptions; this unveils their probabilistic underpinnings and provides novel insights into their applicability for various tasks. In this context, we further suggest new practical ways for estimating the stream weights adaptively. Preliminary experimental results in AV-ASR demonstrate the potential of our approach.

2. FEATURE UNCERTAINTY, ADAPTIVE COMPENSATION, AND INFORMATION 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. Normally, for each class we have trained a model that captures these statistics and represents the class-conditional probability distributions $p_X(x|c_i), i = 1 \dots N$. Our decision is then based on the so

called Maximum A Posteriori (MAP) rule:

$$\hat{c} = \operatorname{argmax} P(c_i|x) = \operatorname{argmax} p_X(x|c_i) \cdot P(c_i) \qquad (1)$$

One may identify three major sources of uncertainty that could perplex classification:

- *Inherent model ambiguity* due to improper modeling or limited capability of the feature to discriminate among the various classes. For instance, we cannot expect visual cues to help us distinguish between members of the same viseme class (e.g. /p/ and /b/), [19]. Proper choice of features and modeling schemes may lead to significant reduction of this kind of uncertainty [6].
- *Parameter estimation uncertainty* that mainly originates from insufficient training. Use of the Bayesian Predictive Classification rule instead of the MAP is a possible way to compensate for it [11].
- *Observation uncertainty* due to errors in the measurement process or contamination of the measured property with noise. This is the kind of uncertainty we mainly address in this paper.

We may represent observation uncertainty as a random variable E independent of any class i. For simplicity, it is regarded to be an additive, zero-mean Gaussian variable with probability distribution $p_E(e) = N(e;0,\Sigma_e)$. In this case, the measurement y is actually a realization of the random variable Y:

$$Y = X + E \qquad (2)$$

So, for the MAP rule (1) it would be desirable to use the distributions $p_Y(\cdot|c_i)$ in order to account for observation uncertainty. However, we only have $p_X(\cdot|c_i), i = 1 \ldots N$ available. In this framework, we may refer to the measurements of the variable X as clean training data.

2.1 Adaptive Compensation

To determine these distributions we assume that X and E are independent. Then the probability $p_Y(y|c_i)$ of the uncertain observation y given the class i may be expressed as convolution of $p_X(x|c_i)$ and $p_E(e)$: $p_Y(y|c_i) = \int_{-\infty}^{\infty} p_X(x|c_i) p_E(y-x)\,dx$. If $p_X(x|c_i) = N(x;\mu_i,\Sigma_i)$, then $p_Y(y|c_i)$ is also a normal distribution with the same mean μ_i and variance Σ_i+Σ_e:

$$p_Y(y|c_i) = \int_{-\infty}^{\infty} N(x;\mu_i,\Sigma_i) N(y-x;0,\Sigma_e)\,dx \qquad (3)$$

$$p_Y(y|c_i) = N(y;\mu_i,\Sigma_i+\Sigma_e) \qquad (4)$$

The above result indicates that it is possible to compensate for the observation uncertainty. The variances Σ_i of the trained models, namely the class-conditional probability distributions of the clean training data may be adjusted by adding the variance Σ_e of the measurement noise. A similar approach has been previously followed in [4, 23].

To further illustrate this point, we discuss how observation uncertainty influences decision in a simple 2-class classification task. The two classes are modeled by 2D spherical Gaussian distributions, $N(\mu_1,\sigma_1^2 I), N(\mu_1,\sigma_2^2 I)$ and they have equal prior probability. If our observation y is corrupted by zero mean spherical Gaussian noise with covariance matrix $\sigma_e^2 I$ then the modified decision boundary is described by the following equation [6]:

$$\log \frac{N(y;\mu_1,\sigma_1^2 I+\sigma_e^2 I)}{N(y;\mu_2,\sigma_2^2 I+\sigma_e^2 I)} = 0 \qquad (5)$$

If 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. So, neglecting uncertainty in the decision may easily lead to misclassifications. As uncertainty increases, decision becomes even more difficult since the observation is even less informative. For infinite uncertainty we have just to pick the class whose mean is closer to the observation, which is also intuitively expected. The above example is demonstrated in Fig. 1.

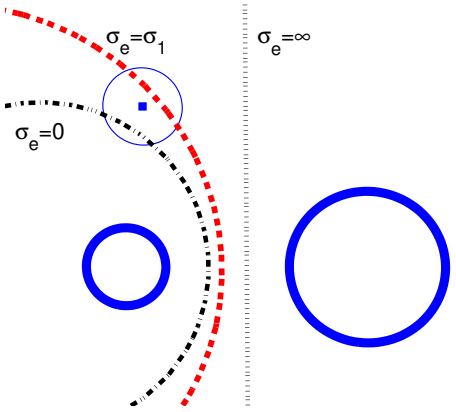

Figure 1: Decision boundaries for classification of a noisy observation (square marker) in two classes, shown as circles, for various observation noise variances. Classes are modeled by spherical Gaussians of means μ_1, μ_2 and variances $\sigma_1^2 I$, $\sigma_2^2 I$ respectively. The decision boundary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

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

$$p(x_{1:S}|c) = p(x_1,\ldots,x_S|c) = \prod_{s=1}^{S} p(x_s|c). \qquad (6)$$

Application of Bayes' formula yields the class label probability given the features:

$$p(c|x_{1:S}) \propto p(c) \prod_{s=1}^{S} p(x_s|c) \qquad (7)$$

This probability can then be used in classification, *e.g.* by the MAP rule $\hat{c} = \operatorname{argmax}_{c\in\mathscr{C}} p(c|x_{1:S})$.

In an attempt to improve classification performance, several authors have introduced stream weights w_s as exponents in 7, resulting to the modified score

$$b(c|x_{1:S}) = p(c) \prod_{s=1}^{S} p(x_s|c)^{w_s}, \qquad (8)$$

which can also 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 experimentally been proven beneficial for feature integration in both intra-modal (*e.g.* multiband audio [2]) and inter-modal (*e.g.* audio-visual speech recognition [7, 9, 15]) scenarios.

However we find the stream weights formulation unsatisfactory in many respects. From a theoretical viewpoint, the weighted score b in (8) ceases having the probabilistic interpretation of (7) as class label 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 audio-visual 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 [9], they are not rigorously justified and are difficult to generalize.

We will now show that our approach for model adjustment in the presence of 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

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \int p(x_s|c) p(y_s|x_s) dx_s, \qquad (9)$$

which is just a generalization of the convolution rule of Sec. 2.1 to the independent multiple streams case. In the common case that the clean feature emission probability is modeled as a mixture of gaussians (MOG), *i.e.* $p(x_s|c) = \sum_{m=1}^{M_{sc}} \rho_{scm} N(x_s; \mu_{scm}, \Sigma_{scm})$, and the observation noise at each stream is considered gaussian, *i.e.* $p(y_s|x_s) = N(y_s; x_s, \Sigma_{es})$, it directly follows from the analysis of Sec. 2.1 that Eq. (9) can be written as

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \sum_{m=1}^{M_{sc}} \rho_{scm} N(y_s; \mu_{scm}, \Sigma_{scm} + \Sigma_{es}), \qquad (10)$$

which simply means that we can proceed by considering our features y_s clean, provided that we increase the model covariances Σ_{scm} by Σ_{es}. Note that, although the measurement noise covariance factor Σ_{es} of each stream is the same for all classes c and all mixture components m, noise particularly affects the most peaked mixtures, for which the measurement noise uncertainty represented by Σ_{es} is substantial relative to the modeling uncertainty due to Σ_{scm}.

Although Eq. (10) is conceptually simple and easy to implement, provided that a good estimate of the measurement noise variance Σ_{es} of each stream is available, 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. (8), we examine a particularly illuminating special case of our result. More specifically, we make two simplifying assumptions:

1. The measurement noise covariance is a scaled version of the model covariance, *i.e.* $\Sigma_{es} = r_{scm} \Sigma_{scm}$ for some positive constant r_{scm} which can be considered as the relative measurement error. The simplest case that this is true is when all covariances are spherical.
2. For every stream observation y_s the gaussian mixture response of that stream is dominated by a single component m or, equivalently, there is little overlap among different gaussian mixtures.

Under these two conditions Eq. (10) can be approximated by

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \rho_{scm} N(y_s; \mu_{scm}, (1 + r_{scm}) \Sigma_{scm}). \qquad (11)$$

Using the power-of-gaussian identity $N(x; \mu, \Sigma)^w \propto N(x; \mu, w^{-1}\Sigma)$ we can write the last equation as

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \left[\tilde{\rho}_{scm} N(y_s; \mu_{scm}, \Sigma_{scm}) \right]^{w_{scm}}, \qquad (12)$$

where

$$w_{scm} = 1/(1 + r_{scm}) \qquad (13)$$

is the *effective stream weight* and $\tilde{\rho}_{scm}$ is a properly modified mixture weight which is independent of the observation y_s (the sum of these modified stream weights $\sum_{m=1}^{M_{sc}} \tilde{\rho}_{scm}$ needs not necessarily equal 1). Note that these effective stream weights are between 0 (for $r_{scm} \gg 1$) and 1 (for $r_{scm} \approx 0$) 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 $\sum_{s=1}^{S} w_{scm} = 1$, 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. (13) provides a rigorous mechanism to select for each new measurement value-bias-variance triplet $(y_s, \mu_{es}, \Sigma_{es})$ all involved stream weights *fully adaptively*, *i.e.* with respect to both class label c and mixture component m.

3. AUDIO-VISUAL SPEECH RECOGNITION

To demonstrate the applicability of the proposed fusion scheme we apply it in a practical problem for which proper information fusion is of critical importance. We show that Audiovisual Speech Recognition can clearly benefit from the suggested approach.

3.1 Visual Front-end

The role of the visual front-end in audiovisual speech recognition systems is to track the speaker's face and extract a low-dimensional feature vector which summarizes visual speech information in video. 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 *Region Of Interest* (ROI) around the mouth [19].

We use *Active Appearance Models* (AAM) [3] of faces to accurately track the speaker's face and extract visual speech

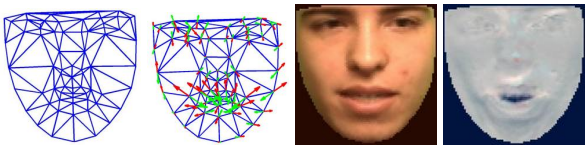

Figure 2: *Left*: Mean shape s_0 and the first eigenshape s_1. *Right*: Mean texture A_0 and the first eigentexture A_1.

features from it, capturing both shape and texture of the face. AAM, which were first used for AV-ASR in [13], 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 framework an object's shape is 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 of the shape s from the mean shape s_0 by letting s lie in a low-dimensional manifold. Typically a linear n-dimensional subspace is utilized, yielding $s = s_0 + \sum_{i=1}^{n} p_i s_i$. The deformation of the mean shape s_0 to another shape s defines a mapping of the landmark points. This mapping can be extended to the whole face area by imposing regularity constraints, utilizing e.g. thin plate splines. This procedure yields a deformation $W(x;p)$ mapping each pixel of the face template to a pixel on the exemplar face. The spatial deformation $W(x;p)$ brings the face exemplar I into registration with the face template A. After factoring out shape deformation, the face color texture $I(W(x;p))$ of a novel face image I registered with the mean face can be modeled as a weighted sum of "eigenfaces" $\{A_i\}$ as: $A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, where A_0 is the mean texture of faces. The mean shape and eigenshapes $\{s_i\}$ and their texture counterparts $\{A_i\}$ are learned during a training phase, using a representative set of hand-labelled face images [3]. The training set shapes are first aligned by means of Procrustes' Analysis and then a PCA of the aligned training set shapes yields the main modes of shape variation $\{s_i\}$. Similarly, the leading principal components of the training set texture vectors constitute the eigenface set $\{A_i\}$. The first eigenshape s_1 and eigenface A_1 extracted by such a procedure are depicted in Fig. 2.

Given a trained AAM and a novel image I, model fitting amounts to finding for each new image the parameters $\tilde{p} \equiv \{p, \lambda\}$ which minimize a measure of the discrepancy between the registered image $I(W(x;p))$ and the AAM appearance reconstruction, such as:

$$E(\tilde{p}) = \underset{p,\lambda}{\operatorname{argmin}} \sum_{x \in s_0} \frac{1}{\sigma^2} \left[A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(W(x;p)) \right]^2, \tag{14}$$

where σ^2 is the noise variance. A global similarity transform on the shape and a linear brightness correction on the texture (not included in eq. (14)) are also used to allow for scale and brightness invariance. Although this is a non-linear least-squares optimization problem if attacked straightforwardly, there are efficient real-time approximate algorithms for iteratively solving it [3] to obtain the visual feature vector $\tilde{p}$. The fitting procedure uses the output of a face detector as initial shape estimate for the first video frame and it is repeated for each new video frame using the converged solution at the previous frame as starting point. The variance of the visual features is computed as the uncertainty in estimating the parameters of the corresponding non-linear least squares problem [20]. This method generally yields satisfactory variance estimates; however, it tends to under-estimate the tracking error in case the AAM instantaneously mistracks the face; we are currently exploring alternatives to alleviate this problem.

Ultimately a sequence of visual speech features $V_t \equiv \{p_t, \lambda_t\}$, along with their respective variances Σ_{V_t}, is extracted for each frame t.

3.2 Audio Front-end

Exploiting audio information is crucial for speech recognition. However, contamination of speech with noise can degrade ASR performance dramatically [1]. In this case apart from robust feature extraction methods [10], the role of the visual cues becomes prominent. Adaptive fusion of multiple modalities in the presence of uncertainty in the audio observations requires methods that account for the audio observation error, as shown in Sec. 2.

Standard methods employed for the representation of the audio stream do not include explicitly the description of the measurement error. This error can be estimated in the case of noisy speech by noise estimation methods. These include spectral subtraction and speech enhancement techniques [23], or statistical modeling of the error e.g. by expectation maximization and iterative Taylor series [4] or by integrated speech/noise HMMs [21]; other implicit techniques include the use of voicing criteria [8] and sub-band ASR [14].

3.3 Audio-Visual Speech Recognition Experiments

The novel fusion approach proposed above is evaluated via classification experiments on the CUAVE audiovisual database [16]; the considered task is word classification of isolated digits. By contaminating the clean audio signal with babble noise from the NOISEX database we extended the database including its noisy version. Mel frequency cepstral coefficients (MFCC) have been utilized as observations for the audio stream, constructing a 13-dimensional vector. In our preliminary experiments, we have utilized the squared differences between clean and noisy features as uncertainty variances of the audio cues. As far as the visual front-end is concerned, we form a visual feature vector by concatenating 6 shape and 12 texture features along with their variances, computed as discussed in Section 3.1. Mean Normalization has been applied to both the audio and visual features.

For the acoustic and visual modeling of the observations we constructed 8-state left-right word multistream HMMs [19] with a single mutlidimensional Gaussian observation probability distribution per stream at each state. The models were trained on clean data. Incorporation of feature uncertainty in the testing phase has been implemented in the HMM framework by increasing the observation variance as presented in Section 2 (Eq. 10).

Our experimental results summarized in Fig. 3 show that accounting for uncertainty in the case of audiovisual fusion (AV-UC, Audiovisual with Uncertainty Compensation) improves AV-ASR performance. For the baseline audiovisual setup we used multistream HMMs with stream weights equal to unity for both streams. For comparison, we also provide results with stream weights as exponents (AV-W) fixed at certain values that maximize classification accuracy in a held out data set for each noisy condition. For this scheme we assume that the audio and visual weights sum to unity and the applied audio stream weights for various SNRs are given in Table 1.

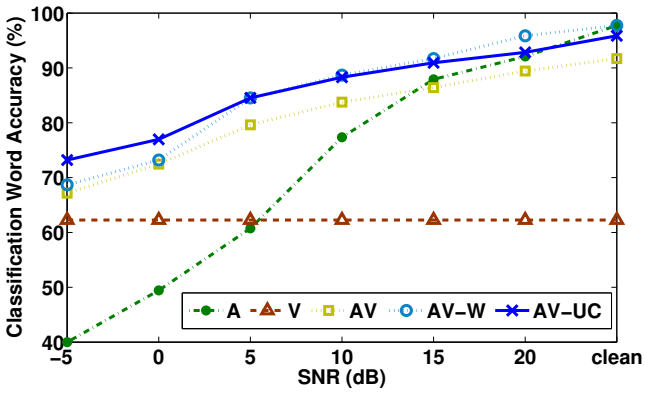

Figure 3: Classification Word Accuracy (%) on the CUAVE database; Audio (A), Visual (V), baseline Audio-Visual (AV), baseline Audio-Visual with Stream Weights as Exponents (AV-W) and the proposed Audio-Visual Fusion with Uncertainty Compensation (AV-UC) scores in various levels of babble noise.

The proposed approach (AV-UC) seems particularly effective at lower SNRs. In higher SNRs the fact that the variance of the visual features is underestimated is possibly responsible for the slightly lower results than those of the AV-W scheme.

SNR (dB)	-5	0	5	10	15	20	clean
w_A	0.2	0.5	0.7	0.8	0.8	0.9	1

Table 1: Audio stream weight as applied in the AV-W (Audiovisual with Stream Weights as Exponents) scheme.

4. PERSPECTIVE

The paper has shown that taking the feature uncertainty into account constitutes a fruitful framework for pattern analysis tasks. This is especially true in the case of multiple complementary measurement streams, where having a good estimate of each stream's uncertainty at a particular moment allows for fully adaptive stream integration schemes, greatly facilitating information fusion.

However, in order 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 features in pattern recognition tasks should accompany feature estimates with their respective errorbars. Although various authors have done progress in the area, much remains to be done before we fully understand the quantitative behavior of popular features commonly used in speech recognition under various environmental conditions.

Considering possible asynchronicity between information streams may also be beneficial to multistream fusion [7, 12]. Early experimentation with product HMMs in the proposed Uncertainty Compensation framework has demonstrated additional improvement of ASR performance.

5. ACKNOWLEDGMENTS

We thank A. Potamianos for discussions and for providing the initial experimental setup for AV-ASR, I. Kokkinos for visual front-end discussions, K. Murphy for using his HMM toolkit, and J.N. Gowdy for the use of the CUAVE database.

REFERENCES

[1] J. B. Allen. How do humans process and recognize speech? *IEEE TSAP*, 2(4):567–577, 1994.

[2] H. Bourlard and N. Morgan. Connectionist speech recognition: A hybrid approach. *Kluwer Academic Publishers*, 1994.

[3] T.F. Cootes, G.J. Edwards, and Taylor C.J. Active appearance models. *IEEE PAMI*, 23(6):681–685, 2001.

[4] 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*, 13(3):412–421, 2005.

[5] V. Digalakis, J.R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE TSAP*, pages 431–442, 1993.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2nd edition, 2000.

[7] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Tr. Multimedia*, 2(3):141–151, 2000.

[8] H. Glotin and F. Berthommier. Test of several external posterior weighting functions for multiband full combination ASR. In *Proc. Int'l Conf. on Spoken Language Processing*, pages 333–336, 2000.

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

[10] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *Proc. ICASSP*, volume 1, pages 121–124, 1992.

[11] Q. Huo and C. Lee. A bayesian predictive approach to robust speech recognition. *IEEE TSAP*, 8(3):200–204, 2000.

[12] J. Luettin, G. Potamianos, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *Proc. ICASSP*, 2001.

[13] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE PAMI*, 24(2):198–213, 2002.

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

[15] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 11:1–15, 2002.

[16] 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.

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

[18] A. Potamianos and V. Weerackody. Soft-feature decoding for speech recognition over wireless channels. In *Proc. ICASSP*, 2001.

[19] G. Potamianos, C. Neti, G. Gravier, and A. Garg. Automatic recognition of audio-visual speech: Recent progress and challenges. *Proc. of the IEEE*, 91(9):1306–1326, 2003.

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

[21] 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*, 2(2):245–257, 1994.

[22] N. B. Yoma, F. McInnes, and M. Jack. Weighted matching algorithms and reliability in noise canceling by spectral subtraction. In *Proc. ICASSP*, volume 2, pages 1171–1174, 1997.

[23] N.B Yoma and M. Villar. Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. *IEEE TSAP*, 10(3):158–166, 2002.