CLASS-SPECIFIC CLASSIFIERS IN AUDIO-VISUAL SPEECH RECOGNITION

Virginia Estellers[1], *Paul M. Baggenstoss*[2], *Jean-Philippe Thiran*[1]

[1] Ecole Polytechnique Fédérale de Lausanne
Signal Processing Laboratory (LTS5)
Switzerland

[2] Naval Undersea Warfare Center
Newport RI, USA

ABSTRACT

In this paper, class-specific classifiers for audio, visual and audio-visual speech recognition systems are developed and compared with traditional Bayes classifiers. We use state-of-the-art feature extraction methods and develop traditional and class-specific classifiers for speech recognition, showing the benefits of a class-specific method on each modality for speaker dependent and independent set-ups. Experiments with a reference audio-visual database show a general increase in the systems performance by the introduction of class-specific techniques on both visual and audio-visual modalities.

1. INTRODUCTION

Visual information can improve the performance of audio-based Automatic Speech Recognition (ASR) systems, especially in the presence of noise [16]. The improvement is due to the complementary nature of the audio and visual modalities, as visual information helps discriminate sounds easily confusable by ear but distinguishable by eye.

ASR systems are composed of a feature extraction and a classification block. In this paper, we investigate how to apply a class-specific classifier for ASR in both single and multimodal systems, namely audio, visual and audio-visual systems. We use state-of-the-art feature extraction systems for the audio and video signals and focus on the requirements of a class-specific classifier for speech recognition purposes. Our design allows different feature sets and dimensionality reducing transforms for each class of interest, providing a more flexible system and avoiding the *Curse of dimensionality* [4]. We report experiments with the CUAVE database [12] and show that the class-specific approach outperforms the traditional one for visual and audio-visual recognition tasks. Previous studies on class-specific audio ASR have been conducted [3], but none considered the visual domain or the fusion of modalities.

The paper is organized as follows. Section 2 presents ASR as a classification task, justifying the necessity of a dimensionality reduction transform and how the class-specific method faces the problem. In section 3, we apply that method with the statistical models used in ASR and in section 4 we describe the experimental set-up, whose results are reported in section 5. Finally conclusions are drawn in section 6.

2. CLASSIFICATION AND CLASS-SPECIFIC METHOD IN SPEECH RECOGNITION

ASR systems are designed to assign to each utterance X the most probable word, phoneme or sentence within its vocabulary and grammar rules $\mathscr{L}$. The problem can be formulated as a M classification problem, that is, assigning a multidimensional sample of data X to one of M possible classes.

This work is supported by the Swiss National Science Foundation through the IM2 NCCR

The optimal Bayes classifier chooses the most likely class given the observed data X, that is

$$\arg\max_{H_j \in \mathscr{L}} p(H_j|X)$$

where p represents the probability density function (*pdf*) and H_j the hypothesis that class j is true. Making use of the Bayes rule, we can rewrite the classifier as

$$\arg\max_{H_j \in \mathscr{L}} p(X|H_j)p(H_j)$$

decomposing the problem in two: estimating $p(H_j)$ from the language model and $p(X|H_j)$ from a statistical model of H_j.
In this paper we assume equally probable classes and focus on the estimation of $p(X|H_j)$, that is, characterizing statistically X under each of the hypotheses by its *pdf*. We will define phonemes, the smallest unit of sound meaningful in terms of speech, as our classes of interest and propose a design for isolated speech recognition tasks. Such a system could be generalized considering the concatenation of phonemes to form words or sentences, the grammar rules and a decoder considering the possible paths through a lattice of all the concatenated phonemes (the Viterbi decoder, usually).

In that ASR classification task, the dimension of the feature space necessary to accurately distinguish all possible classes is usually large. At the same time, the complexity and the amount of data necessary to estimate those *pdf*s grow exponentially with the dimension of X. Therefore, we either loose information discarding features in order to obtain accurate estimates of $p(X|H_j)$, which might result in being unable to distinguish similar classes, or work in high-dimensional feature spaces and suffer the problems of estimating high dimensional *pdf*s. This is known as the *Curse of Dimensionality* [4] and explains the necessity of a dimensionality reducing transform and the interest of researchers in class-specific designs allowing the use of reduced feature sets for each class while obtaining a Bayes classification system.

The class-specific method [1, 2, 7] is a Bayes classification system reformulated to use class-dependent features. It identifies a set of statistics $z_j = T_j(X)$ that is "best" to statistically describe each class H_j and explains how to project the estimated $p(z_j|H_j)$ to the original feature space $p(X|H_j)$.
The *pdf* projection theorem [1] states that any probability density function $g(z)$ defined on a feature space z where $z = T(X)$, can be converted into a *pdf* $h(X)$ defined on X using the formula

$$h(X) = \frac{p(X|H_0)}{p(z|H_0)}\, g(z) = J(X)\, g(z), \tag{1}$$

where H_0 is any statistical hypothesis for which $p(X|H_0)$ and $p(z|H_0)$ are known. In fact, the theorem states that $h(X)$ not only is a *pdf* and integrates to 1, but that it is a member of the class of *pdf*s that generate $g(z)$ through transformation $T(X)$. The optimality and other properties of $h(X)$ are presented and can de found in [6].

The *pdf* projection operator $J(X)$, called the J-function, is

 ISSN 2076-1465

thus a function of the raw data X, the feature transformation $T(X)$, and the reference hypothesis H_0. It projects the estimated probabilities from the reduced feature set z to the original feature space X, enabling the use of different features for each class of interest. It is important to note that this function is not estimated, but a fixed function of X, the transform $T(X)$ and reference hypothesis, so it is not subject to the *Curse of Dimensionality*. To obtain a closed form for the J-function, the distributions $p(X|H_0)$ and $p(z|H_0)$ need to be known either analytically, or by accurate approximation valid in the tail regions. These conditions have been met for some of the most useful feature transformations and reference hypothesis [7]. In general, H_0 can also be a function of the classifier's hypothesis H_j, however, in this work we will use a common reference H_0 and only adapt the transforms $T_j(X)$ applied to each class.

Compared to a traditional system, the class-specific approach involves the use of a different transform for each class and the computation of the J-function in order to project the estimated probabilities of the HMMs to the common feature space. The complexity of the system is not usually much increased, since by correctly choosing the reference hypothesis and transforms, the computations involved in the J-function can be simplified. On its turn, the cost of using several transforms for the feature stream instead of just one is negligible compared to the HMM computations when dealing with linear transforms and a reduced number of classes.

3. PROBABILITY ESTIMATION

The structure of the classification system is the following: given the original feature stream X we apply different transforms T_j for each class and obtain the corresponding features z_j. We use then statistical models to compute $p(H_j|z_j)$ for each class and, finally we evaluate the J-function on the original input X and transforms T_j under the reference hypothesis to project the obtained probabilities $p(H_j|z_j)$ to the original feature space $p(H_j|X)$, where the traditional Bayes classifier is used.

Following that structure, in that section we first present the dimensionality reducing transforms used in our system, we then introduce Hidden Markov Models as the statistical tools used in ASR to estimate the probabilities $p(H_j|z_j)$ in the reduced feature sets and we finally explain how to apply the selected transforms and derive an analytical expression for the J function with those models.

3.1 Dimensionality reduction and class-specific features

In order to reduce the dimensions of the samples X we apply a linear transform, so that the new features $z = W^T X$ retain as much of the information as possible of the original space. In our case, we want to preserve variance of the original space, or class-subspaces, and the transform we thus consider is Principal Component Analysis (PCA). PCA requires a training space $\mathscr{X}$, composed of enough samples X to characterize the original feature space, to find the subspace whose basis vectors correspond to the maximum-variance directions in $\mathscr{X}$. In practice, we use the same training set as the one used to train the classifiers and learn the models for each class.

To fairly compare the class-specific method with a traditional approach, we define the same kind of transform for the whole training dataset $\mathscr{X}$ and for the subsets associated to each of the classes $\mathscr{X}_j \subset \mathscr{X}$. Comparing the performance of the system with features $\{z_j\}_{j=1\ldots M}$ in a class-specific design against $\tilde{z} = \bigcup_{j=1}^{M} z_j$, would just show the benefits of a class-specific approach against the *Curse of Dimensionality*, but not how class-specific features might outperform general ones for a given dimensionality and dimensionality reducing transform. To that purpose we split our training dataset into its classes, use $\mathscr{X}$ to determine the transform T leading to features z and each of the $\mathscr{X}_j$ to determine the class-specific transforms T_j and the corresponding features z_j, with z and z_j of the same dimension.

3.2 Hidden Markov Models

A single-stream Hidden Markov Model (HMM) is the statistical model traditionally used in audio ASR [15]. It has a hidden state and an observed variable associated to the feature streams, where the state variable evolves through time as a first order Markov process assigning different statistical distributions to the observed variable. A typical audio-visual extension is the coupled HMM [5, 9, 10], where the audio and video streams are synchronized at model boundaries and the joint audio-visual likelihood is a geometrical combination of the audio and visual ones.

However, the use of HMMs for ASR suffers two main limitations. First, the Markovian assumption of the HMMs fails to model the correlation in time of the original speech and estimates of the features derivatives must appended to the original observed features. The second constraint is due to the *Curse of Dimensionality*, as the correct statistical description of the observed features is just possible with a low dimensionality and the size of the vector has thus to reduced before being input to the HMM. Those steps are included in the transforms we apply to the original features and taken into account in the class-specific approach, as we describe in more detail in section 3.3.

3.3 Class-specific method with Hidden Markov Models

Let us denote x the original feature stream from which to define the observed features and $x(t)$ its value at time t. We first append the time derivatives to the features and obtain a new stream y defined as $y(t) = [x(t) \;\; \dot{x}(t)]$ with larger dimensionality than x. The final features z are obtained through a dimensionality reduction technique on y, in our case projecting each sample to the reduced PCA space $z(t) = W^T y(t)$.

For each utterance of length T, HMMs are used to estimate the likelihood of all the possible utterances given the observed features $Z = [z(1)\ldots z(T)]$. We will see that, in fact, we can apply a single linear transform to the original samples of the utterance $X = [x(1)\ldots x(T)]$ in order to obtain the previously defined Z.
Approximating the time derivatives by finite differences, we write Y as a linear transform B on the feature samples.

$$\dot{x}(t) = \frac{1}{2}\left(x(t+1) - x(t-1)\right) \;\; \rightarrow \;\; Y = B^T X$$

At the same time, PCA defines a fixed linear transform to be applied each time instant to the samples of y, $z(t) = W^T y(t)$. Thus, correctly applying the W matrix T times we create a new matrix C and rewrite the whole as a linear transform.

$$z(t) = W^T y(t) \;\; \rightarrow \;\; Z = C^T Y = C^T B^T X = A^T X$$

The matrix A defines a linear transform that combines PCA on the expanded features and time differencing of the original stream.

A first transform to be considered for the class-specific approach is thus $Z = A^T X$, with different A matrices for each class. Nevertheless, the dimensionality reduction implies that the subspace orthogonal to the columns of A will be absent from the output. If any data of certain class contains energy in the orthogonal space and the features for another class allow this energy to appear at the output, classification errors might take place. To avoid it, we adapt our linear transform appending a power estimate of the error introduced in the dimensionality reduction, that is, the energy lost on the orthogonal space to A.
First, we compute the error comparing X and its prediction based on Z, that is $\hat{X} = A\left(A^T A\right)^{-1} A^T X$, and look at the energy of

the error at each time step to form the final reduced features are $[z(t) \;\; r(t)]$ and $[Z \;\; R]$ as

$$r(t) = |x(t) - \hat{x}(t)| \;\; \rightarrow \;\; R = |(Id - A\left(A^T A\right)^{-1} A^T)X|$$

As we have already said, the J-function is a function of the original data sample X that depends on the feature transformation and the reference hypothesis. The choice of that hypothesis usually means choosing a simple *pdf* for $p(X|H_0)$ trying to simplify the determination of $p([Z \;\; R]|H_0)$.

We choose as reference hypothesis X being independent identically distributed samples of normal Gaussian noise under H_0, so that under H_0 both $Z = A^T X$ and the error are also samples of Gaussian random variables with known mean and covariance. The chosen R being the lost energy on the projection $Z = A^T X$, assures the independence of Z and R and allows the factorization $p([Z \;\; R]|H_0) = p(Z|H_0)p(R|H_0)$. Under these circumstances, when the energy of the error $e(t)$ is added up in $r(t)$ for each time step, the result is a Chi-Square random variable with $N-P$ degrees of freedom, with N and P denoting the size of the original and transformed feature samples $x(t)$ and $z(t)$ respectively. We have then a closed form for the J-function based on chi-squared and Gaussian distributions, which we will use to project the reduced set probability estimates obtained with our HMMs to the original feature space and build a Bayes classifier.

4. EXPERIMENTAL SET-UP

We perform speechreading experiments on the CUAVE database [12]. We use the static portion of the 'individuals' section of the database, consisting of 36 speakers repeating the digits five times. We divide our experiments into speaker dependent and independent doing three-fold cross validation in every case, i.e training on two thirds of the all data for each speaker and testing on the remaining in speaker dependent experiments and training on two thirds of the speakers and testing on the rest for the speaker independent set-up. Changing the training and testing splits of the data, we can validate our results by three runs and ensure they do not depend on the training or testing data.

The audio features used are 13 mel-frequency cepstral coefficients with cepstral mean normalization and their first and second temporal derivatives. In testing, we artificially add babble noise to the audio stream with Signal to Noise Ratios (SNR) ranging from clean to -10db, at 5db steps. The visual features are selected from a pool of DCT coefficients on a 128x128 grayscale image of the speaker's mouth, normalized for size, centered and rotated in order to reduce speaker variability. The 2-dimensinal DCT of the images are then computed, from which we take the first 16x32 coefficients and remove their even columns to exploit face symmetry [13].

We define the phonemes as our classes of interest and propose different experiments in terms of complexity: 3 simpler experiments with only 4 phoneme classes and a final experiment with the 20 phonemes available in the database. The 3 subsets of classes are chosen in order to test the method in different conditions: distinguishing between consonants visually distinguishable {n,r,t,v}, consonants {v,w,r,s} visually confusable within the reduced set [8] and a set including vowels and consonants {ah,eh,n,uw}. The task examined is then isolated phoneme classification, which is the core of continuous speech recognition, where phoneme models are concatenated to recognize words or sentences taking into account vocabulary and grammar rules. Class-specific HMMs could also be defined at word level, but they would be limited to small vocabulary tasks while phoneme models are the natural choice for real-world systems. The number of phoneme classes depends on the language, 43 in english for instance, and the vocabulary associated to the speech recognition task, 20 classes in the case of the digits. Our experiments are thus limited by the size of the database, but the system could naturally be extended to more general tasks if more training data was provided.

For the single-modality experiments, the phoneme models are made of 3-state HMMs with their observed features described by one and three Gaussian mixtures for the speaker dependent and independent set-ups. In the audio-visual experiments, a coupled HMM from the previous 3-state audio and visual HMMs is built, where the contribution of each stream to the combined likelihood is geometrically weighted with λ_A, λ_V. During testing and for each SNR level, the best fixed weights are chosen from the possible combinations satisfying $\lambda_A + \lambda_V = 1$ and ranging from $\lambda_A = 1$ to $\lambda_A = 0$ at 0.05 steps.

5. EXPERIMENTAL RESULTS

A first set of audio and video-only experiments is performed in order to choose the number of reduced features leading to the best performance and whether or not a class-specific approach on each modality is useful. In fact, a class-specific approach has already been used for audio-only ASR outperforming the traditional system in a speaker-dependent set-up [3]. We focus, however, on the improvement we can obtain on the system's performance by adding the visual modality.
In the results presented, 'pca' stands for the traditional Bayes classifier using PCA as dimensionality reduction transform and 'cs-pca' for the class-specific one. Similarly, 'spkr-dep' and 'spkr-ind' correspond to the speaker dependent and independent set-ups. In the audio-visual experiments, we also report results of an audio-only system in order to measure the improvement obtained by the visual modality.

The results for the single modality experiments are presented in tables 1 and 2. As expected, the performance of the audio system is superior than the video one, 96% against 90% of recognition rates in speaker dependent set-up and 94% against 53% in the speaker-independent task. In both modalities, the class set {v,w,r,s} proves more challenging than the others, who perform similarly. In speaker dependent experiments, the class-specific method outperforms the traditional approach on both audio and visual modalities, increasing the recognition rate around 2% in the audio and 10% in the video case. However, for the speaker independent set-up, the class-specific design only improves the recognizer's performance of the visual modality system, while using the original audio features obtains better results than any PCA reduced set. In that case, the improvement of the class-specific approach is also more limited, around 4% improvement in the recognition rate in the visual modality and none at all for the audio. Such a behaviour can

Audio	spkr dep			spkr ind		
Class sets	MFCC	cs-pca	pca	MFCC	cs-pca	pca
{n,r,t,v}	98.3	100	98.9	98.3	96.7	86.7
{v,w,r,s}	92.7	97.8	95.7	87.8	83.4	71.9
{ah,eh,n,uw}	97.1	100	100	95.6	95.0	91.9

Table 1: Percentage of correctly recognized phonemes in audio only experiments with traditional and class-specific classifiers. The original MFCC features and the extracted PCA from all the classes are used in a traditional Bayes classifier, while the features obtained by class-specific PCA are used in the class-specific design.

be explained by the fact that MFCC are features already designed for human speech recognition, where speaker independence has already been taken into account in their definition, while the original visual features correspond to a standard representation of images, not aimed to the representation of the mouth area for ASR. In that sense, the results with the original visual features,

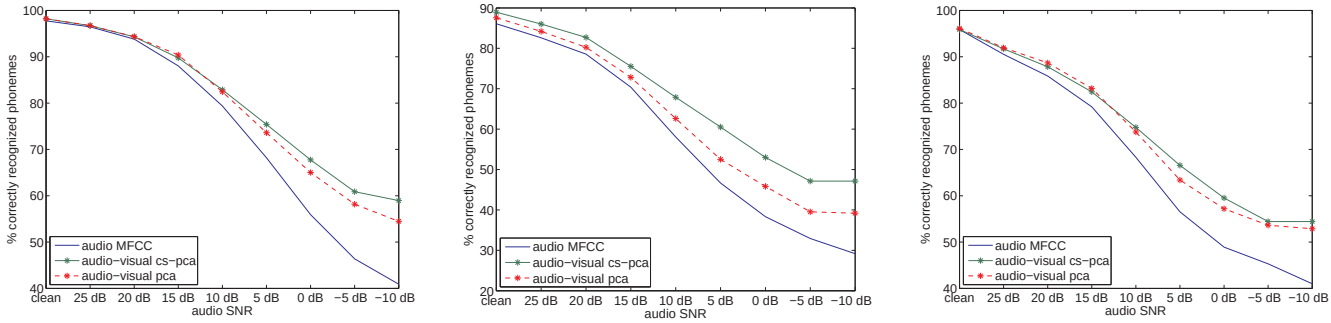

Figure 1: Recognition in the speaker independent audio-visual task for sets {n,r,t,v}, {v,w,r,s} and {ah,eh,n,uw} compared to an audio-only recognizer. The audio-visual classifier use the original MFCCs as audio features and PCA or cs-PCA for the video modality.

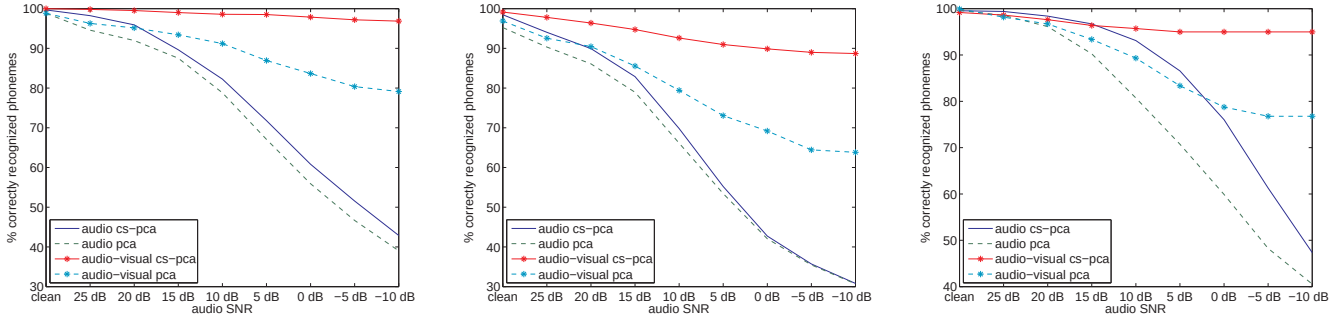

Figure 2: Recognition in the speaker dependent audio-visual task for sets {n,r,t,v}, {v,w,r,s} and {ah,eh,n,uw} compared to an audio-only recognizer. The audio-visual classifier use PCA and cs-PCA features for the both the audio and video modalities.

Visual	spkr dep		spkr ind	
Class sets	cs-pca	pca	cs-pca	pca
{n,r,t,v}	97.3	85.5	58.9	55.2
{v,w,r,s}	89.7	72.5	46.7	39.0
{ah,eh,n,uw}	91.3	82.2	54.8	53.0

Table 2: Percentage of correctly recognized phonemes in video only experiments with class-specific and traditional Bayes classifier for the corresponding cs-PCA and PCA features.

i.e 512 DCT features plus their first time derivatives, lead to poor recognition performances around 30%, which is little better than the 25% chance of correctly performing in a 4 class subset. This is due to both the curse of dimensionality and the non specificity of the features for ASR. More advanced visual features for speech recognition include speaker normalization techniques [14, 11] leading to more speaker invariant features. The class-specific method should also be useful in conjunction with those feature sets, as the speaker dependent results from our experiments show.

In our experiments, the dimension of the 'pca' and 'cs-pca' features leading to best performance for each class and reduced sets were chosen. For both speaker dependent and independent set-ups the audio dimensionality ranged from 3 to 4 for 'cs-pca' and from 3 to 6 for 'pca', depending on the class sets. For the video modality, the best performance correspond to PCA dimensions of 4 or 5 in speaker dependent experiments and 3 for speaker independent ones. The dimensionality reduction obtained with PCA is therefore considerable, specially in the visual modality, and does not show much phoneme or speaker specificity.

We choose different approaches for each set-up when performing audio-visual experiments. In the audio-visual systems, we combined the approaches obtaining the best results in single-modality experiments because we wanted to compare the performance of the best traditional classifier we could build against a class-specific one. Therefore, we have used a class-specific design on both modalities in the speaker dependent experiments while on the speaker independent set-up, the original audio stream was kept and the class-specific method was just applied to the video stream.

The results of the audio-visual experiments are presented in figures 1 and 2 for the reduced classes sets, showing the advantage of the class-specific approach also in a multimodal domain. As expected, the reduced set {v,w,r,s} proves more challenging than the others, both due to its lower recognition rates in audio and visual modalities. The results with the speaker-dependent and independent set-ups show different gains when using the class-specific method compared to the traditional approach. In the speaker dependent task, both modalities profit from the class-specific design and improve the recognition rates between 2% and 10% depending on the audio SNR, while in the class-independent set-up only the video modality benefits from the class-specific design and increases the recognition between 0.5% and 4% for the different audio SNRs

In the more realistic experiments when all the classes are considered, see table 3, we observe a clear gain on both the incorporation of the visual modality and the class-specific approach, not limited to the speaker dependent set-up. Indeed, in those experiments the visual modality always enhances the audio-only results and the class-specific method proves beneficial in both speaker dependent and independent experiments. Even though the resulst with noisy audio might seem poor, the classifier do better than the 5% chance of recognition in a 20 class set.

For the reduced class sets, the improvement of the class-specific

design is more important in the speaker-dependent task, between 10% to 38% increase in the recognition rate for the different SNR levels, than in the speaker independent one, where that increase ranges from 2% to 10%.

	spkr-dep				spkr-ind		
SNR	audio		audio-visual		audio	audio-visual	
	cs-pca	pca	cs-pca	pca	MFCC	cs-pca	pca
clean	97.7	79.5	98.7	88.0	74.5	76.4	75.0
25db	89.5	63.0	96.2	77.2	63.6	67.0	64.9
20db	82.5	52.9	93.7	71.4	54.8	60.4	56.4
15db	71.5	42.3	91.0	64.1	46.1	51.1	46.8
10db	56.0	31.8	87.8	58.1	35.4	41.7	35.8
05db	39.2	22.3	85.3	52.2	24.8	32.2	25.2
00db	24.7	14.4	83.5	47.8	15.9	26.1	17.1
-05db	15.5	9.5	82.8	44.7	11.4	22.9	12.9
-10db	10.2	6.5	82.8	44.3	8.8	21.9	10.7

Table 3: Percentage of correctly recognized phonemes considering all classes. The audio-visual classifiers use MFCCs for the audio features and PCA or cs-PCA for the video in speaker independent experiments and PCA or cs-PCA for both the audio and video modalities in speaker dependent experiments.

6. CONCLUSIONS

In the present paper we demonstrate that a class-specific approach improves the performance of audio-visual ASR systems. Compared to previous work, we consider the effects of multiple modalities on class-specific methods and the effects of appending the derivatives to the HMM features in order to comply with the Markovian assumption of the HMMs used in ASR.
From our experiments, we conclude that for the speaker independent set-up more work is to be done on the definition of general video features, while the audio MFCC features already suit the task. In those situations, the performance of the audio-visual system, can be boosted with a class-specific approach on the video modality, specially improving the results in noisy conditions. On the other hand, in speaker dependent set-ups, both audio and video modalities profit from the definition of different features for each class through all noise levels. The apparent increased benet of the class-specic approach in the speaker-dependent experiments agrees with the results of previous work on a different data set [3] and may result from the reduced dimensionality needed to accurately model a given phoneme spoken by a given speaker.

REFERENCES

[1] Baggenstoss. The pdf projection theorem and the class-specific method. *Transactions on Signal Processing*, 2003.

[2] Baggenstoss. The class-specific classifier: Avoiding the curse of dimensionality. *Aerospace and Electronic Systems Magazine*, 2004.

[3] Baggenstoss. Iterated class-specific subspaces for speaker-dependent phoneme classification. In *EUSIPCO proceedings*, 2008.

[4] Bellman. Adaptive control processes: a guided tour. *Princeton University Press*, 1962.

[5] Brand, Oliver, and Pentland. Coupled hidden markov models for complex action recognition. In *CVPR proceedings*, 1997.

[6] Kay. Sufficiency, classification, and the class-specific feature theorem. *Transactions on Information Theory*, 2000.

[7] Kay, Nuttall, and Baggenstoss. Multidimensional probability density function approximation for detection, classification and model order selection. *Transactions on Signal Processing*, 2001.

[8] Lucey, Martin, and Sridharan. Confusability of phonemes grouped according to their viseme classes in noisy environments. In *ASSTA proceedings*, 2004.

[9] Nefian, Liang, Pi, Liu, and Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 2002.

[10] Neti, Potamianos, Luettin, et al. Audio-visual speech recognition. In *Final Workshop Report, Johns Hopkins CLSP*, 2000.

[11] G. Papandreou, A. Katsamanis, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audio-Visual Speech Recognition. *Multimodal Processing and Interaction*, pages 1–15, 2006.

[12] Patterson, Gurbuz, Tufekci, and Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *ICASSP proceedings*, 2002.

[13] Potamianos and Scanlon. Exploiting lower face symmetry in appearance-based automatic speechreading. In *AVSP proceedings*, 2005.

[14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, pages 1306–1326, 2003.

[15] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, 53(3):267–296, 1990.

[16] D. Stork and M. Hennecke. *Speechreading by humans and machines: models, systems, and applications*. Springer Verlag, 1996.