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

Class-specific classifiers for audio, visual and audio-visual speech recognition systems are developed and compared with traditional classifiers. We use state-of-the-art feature extraction methods and show the benefits of a class-specific classifier on each modality in speech recognition experiments.

1. INTRODUCTION

Visual information can improve the performance of audio-based Automatic Speech Recognition (ASR) systems, especially in the presence of noise. The improvement is due to the complementary nature of the audio and visual modalities, as visual information helps disentangle 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 approach for ASR in both visual and audio-visual modalities. 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 in ASR systems. 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. We report experiments with the CUAVE database [11] and show that the class-specific approach outperforms the traditional one for visual and audio-visual ASR. 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 the class-specific method and justifies the necessity of a dimensionality reduction transform. In section 3, we apply the statistical models used in ASR to the problem at hand and section 4 describes the experimental set-up, whose results are reported in section 5. Finally conclusions are drawn in section 6.

2. CLASSIFICATION 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 classification is based on

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

where H_j represents 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.

We assume, for the sake of simplicity, 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*.
The dimension of the feature space necessary to accurately estimate $p(X|H_j)$ for all possible classes is usually large. At the same time, the complexity and the amount of data necessary to estimate the *pdf* grow exponentially with the dimension of X. This is known as the curse of dimensionality [4]: we either loose information discarding features or suffer the problems of estimating high dimensional *pdf*s. That explains the necessity of a feature selection method and the interest of researchers in class-specific designs.

2.1 Class-specific method

The class-specific method [2, 1, 7] is a Bayes classification reformulated to use class-dependent features. A class-specific method identifies a set of statistics $z_j = T_j(X)$ that is "best" to statistically describe each class 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

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

where H_0 is any statistical hypothesis for which $p(X|H_0)$ and $p(z|H_0)$ are known. The *pdf* projection operator $J(X)$, called the J-function, is thus a function of the raw data X, the feature transformation $T(X)$, and the reference hypothesis H_0.
The *pdf* projection 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 $T(X)$. Kay [6] further shows optimality properties of $h(X)$.

Typically, 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 [7]. Note that the J-function is not estimated, but a fixed function of these choices, so it is not subject to the curse of dimensionality. Also, in general, H_0 can be a function of the classifier's hypothesis H_j, however, in this work, we use a common reference H_0.

2.2 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 finds the subspace whose basis vectors correspond to the maximum-variance directions in the original training space $\mathscr{S}$.

To fairly compare the class-specific method with a traditional approach, we define the same kind of transform for the whole training dataset $\mathscr{S}$ and for the subsets associated to each of the classes $\mathscr{S}_j \subset \mathscr{S}$. 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. To that purpose we split our training dataset into its classes, use $\mathscr{S}$ to determine the transform T leading to features z and each of the $\mathscr{S}_j$ to determine the class-specific transforms T_j and the corresponding features z_j, with z and z_j of the same dimension.

3. PROBABILITY ESTIMATION

We face two main problems building a class-specific Bayes classifier: defining the reduced feature sets to correctly estimate the class probabilities and choosing a common reference hypotheses H_0 providing a known expression for the J function. Thus, in the present section we first introduce Hidden Markov Models (HMM) as the statistical tools used in ASR [10, 9] to compute $p(H_j|z_j)$, we then define the reduced features based on PCA and we finally derive an analytical expression for the J function.

A single-stream HMM is the statistical model traditionally used in audio ASR. It has a hidden state variable evolving through time as a first order Markov process. A typical audio-visual extension is the coupled HMM [5], 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 Markovian assumption of the HMMs fails to model the correlation in time of the original speech and the correct statistical description of the observed features is just possible with a reduced dimensionality. To overcome those limitations, estimates of the derivatives are appended to the original features and the dimensionality of the vector is reduced before being input to the HMM. Those steps are included in the transforms applied to the original features.

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 Z.
Time derivatives being approximated 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 re-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

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

to form the final reduced features are $[z(t)\ \ r(t)]$ and $[Z\ \ R]$.

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 reference hypothesis and the transform determines how the J-function is calculated. Therefore, defining the reference hypothesis usually means choosing a simple *pdf* for $p(X|H_0)$ trying to simplify the determination of $p([Z\ \ R]|H_0)$. 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 $p([Z\ \ R]|H_0) = p(Z|H_0)p(R|H_0)$.

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. 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 thus obtained a closed form for the J-function based on chi-squared and Gaussian distributions with known mean and variance.

The structure of the classification system is the following: given the original feature stream X we apply a different transforms T_j for each class and obtain the corresponding features z_j. We use then the usually trained HMMs to compute $p(H_j|z_j)$ for each class and, finally we evaluate the J-function on both the original input X and transforms $\{T_j\}_{j=1...M}$ to project the obtained probabilities $\{p(H_j|z_j)\}_{j=1...M}$ to the original feature space $\{p(H_j|X)\}_{j=1...M}$, where the traditional Bayes classifier can be used.
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 as correctly choosing the reference hypothesis and transforms, the computations involved in the J-function can be simplified and 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 .

4. EXPERIMENTAL SET-UP

We perform speechreading experiments on the CUAVE database [11]. 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.

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 image of the speaker's mouth, normalized for size, centered and rotated. The 2-dimensinal DCT of the images are computed, from which we take the first 16x32 coefficients and remove their even columns to exploit face symmetry [12].

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 consonant visually distinguishable {n,r,t,v}, consonants {v,w,r,s} visually confusable [8] and a set including vowels {ah,eh,n,uw}.

For the single-modality experiments, the phoneme models are made of 3-state HMMs with their observed features described by Gaussian mixtures. In the audio-visual experiments, a coupled HMM from 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 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 and 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. 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. 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. That behaviour

Audio	spkr dep			spkr ind		
Class sets	MFCC	cs-pca	pca	MFCC	cs-pca	pca
{n,r,t,v}	92.35	100	98.89	98.3	96.72	86.69
{v,w,r,s}	87.77	97.79	95.67	87.77	83.43	71.88
{ah,eh,n,uw}	95.63	100	100	95.63	95.02	91.88

Table 1: Percentage of correctly recognized phonemes in audio only experiments

Visual	spkr dep		spkr ind	
Class sets	cs-pca	pca	cs-pca	pca
{n,r,t,v}	97.29	85.53	58.9	55.17
{v,w,r,s}	89.68	72.55	46.68	38.96
{ah,eh,n,uw}	91.35	82.21	54.79	53.01

Table 2: Percentage of correctly recognized phonemes in video only experiments

can be explained by the fact that MFCC are features already designed for human speech recognition, while the original

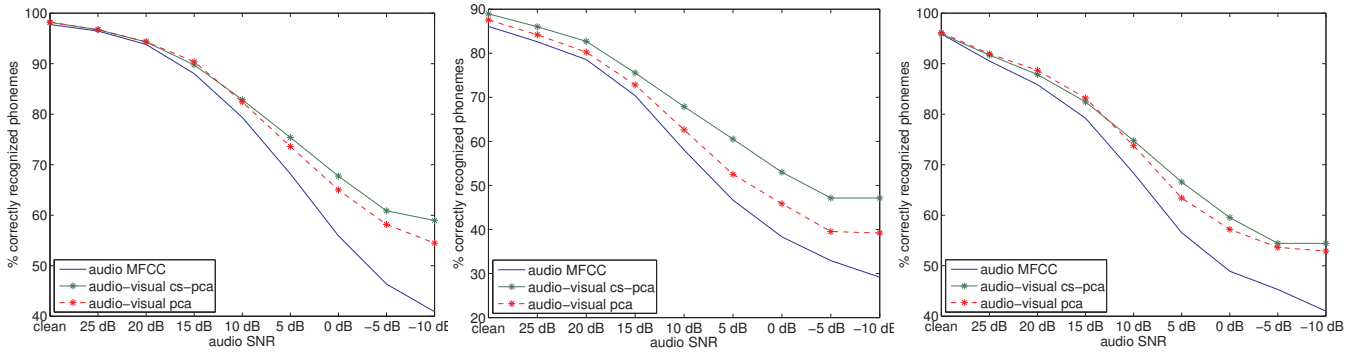

Figure 1: Recognition in the speaker independent audio-visual task for sets {n,r,t,v}, {v,w,r,s} and {ah,eh,n,uw}

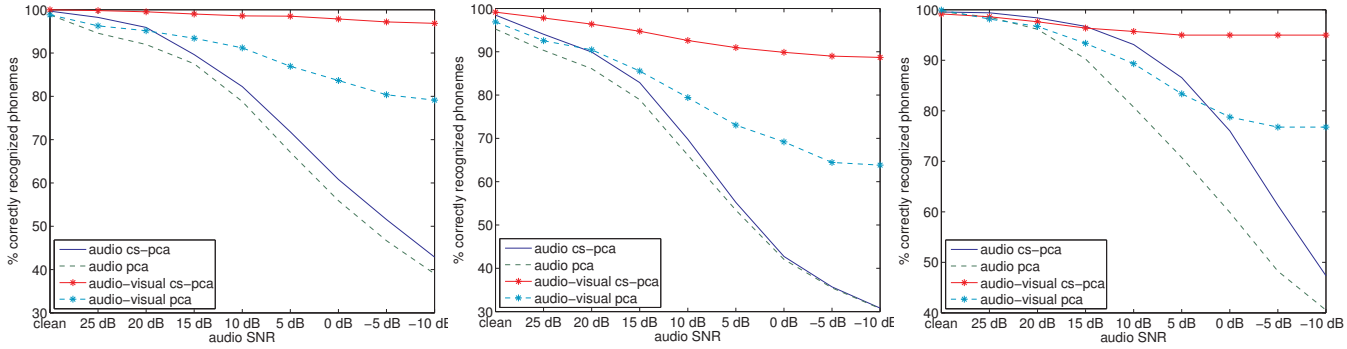

Figure 2: Recognition in the speaker dependent audio-visual task for sets {n,r,t,v}, {v,w,r,s} and {ah,eh,n,uw}

visual features correspond to a standard representation of images, not aimed to the representation of the mouth area for ASR.

We choose different approaches for each set-up when performing audio-visual experiments. The chosen approaches obtained the best results in single-modality experiments, as we wanted to compare the performance of the best traditional classifier we could build with 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 for the speaker independent experiments are presented in figure 1 for the reduced classes sets, , showing the advantage of the class-specific approach also in a multimodal domain. As expected, we observe that the improvement obtained from the visual modality is more relevant when the phonemes are visually distinguishable. The results with the speaker-dependent set-up, see figure 1, are similar and, as in the single modality experiments, the experiments show a clear improvement on the systems performance when the class-specific technique is used.

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.

	spkr-dep				spkr-ind		
SNR	audio		audio-visual		audio	audio-visual	
	cs-pca	pca	cs-pca	pca	MFCC	cs-pca	pca
clean	97.73	79.51	98.75	88.03	74.54	76.39	74.99
25db	89.46	62.96	96.22	77.2	63.64	67.04	64.92
20db	82.47	52.87	93.67	71.36	54.76	60.36	56.38
15db	71.55	42.29	90.97	64.11	46.07	51.06	46.78
10db	56.04	31.78	87.76	58.1	35.36	41.71	35.81
05db	39.2	22.31	85.34	52.21	24.84	32.21	25.24
00db	24.73	14.45	83.51	47.85	15.89	26.06	17.09
-05db	15.49	9.55	82.85	44.71	11.37	22.88	12.87
-10db	10.19	6.52	82.85	44.27	8.77	21.89	10.66

Table 3: Percentage of correctly recognized phonemes

6. CONCLUSIONS

In the present paper we prove 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 made on ASR.

From our experiments, we conclude that for the speaker independent set-up more work is to be done on the definition of 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 def-

inition of different features for each class through all noise levels.
Future work includes, therefore, the introduction of speaker adaptation techniques, the study of other class-specific transforms for the video domain and the application of the explained method to continuous speech recognition.

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

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