Space-time audio-visual speech recognition with multiple multi-class probabilistic Support Vector Machines

Samuel Pachoud *Shaogang Gong* *Andrea Cavallaro*

School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK
{sgg,spachoud}@dcs.qmul.ac.uk, andrea.cavallaro@elec.qmul.ac.uk

Abstract

We extract relevant and informative audio-visual features using multiple multi-class Support Vector Machines with probabilistic outputs, and demonstrate the approach in a noisy audio-visual speech reading scenario. We first extract visual spatio-temporal features and audio cepstral coefficients from pronounced digit sequences. Two classifiers are then trained on a single modality to obtain confidence factors that are used to select the most appropriate fusion strategy. A final classifier is trained on the joint audio-visual feature space and used to recognize digits. We demonstrate the proposed approach on a standard database and compare it with alternative methods. The evaluation shows that the proposed approach outperforms the alternatives both in terms of recognition accuracy and in terms of robustness.

1 Introduction

Robust and accurate audio-visual automatic speech reading (AV-ASR) algorithms have to address three major problems, namely feature extraction, feature fusion, and recognition. Feature extraction is the process of selecting low-level perceptual information (lip movements, colour difference). Effective feature fusion leads to a robust integration of two (possibly) degraded or incomplete signal modalities. Recognition classifies the input signals into two or more semantic labels (words).

As in real-world environments audio and visual cues are likely to be degraded, it becomes essential to extract discriminative features, which provide robust information about the input signals. Traditional speech reading systems use video to assist low signal-to-noise ratio audio recognition. When signals are degraded, it becomes difficult to decide on which cue to rely. To that end, knowing the level of confidence of each signal and fusing them accordingly is an important feature of a classifier. An analysis of the entropy or of the tonality of the visual and audio cues can offer prior knowledge of this confidence. However, this confidence is highly dependent of the nature of the noise. To overcome this problem, the use of machine learning is desirable to provide a confidence factor for each modality. Most machine learning problems are modeled using the generative probabilistic distribution because it provides prior domain specific knowledge in terms of structure and parameter over the joint space of variables. For example, Bayesian networks [1] and Bayesian statistics [2] provide a rich and flexible language for specifying this knowledge and subsequently refining it with data and observations. Recently, discriminative learning algorithms, such as Logistic Regression, Conditional Random Field or Support Vector Machine adjust a possibly non-distributional model to data optimizing for a specific task, such as classification or prediction. This typically leads to superior performance, which can be obtained by avoiding generative modelling and focusing on the given task.

Generative approaches produce a probability density model over all variables in a system and manipulate it to compute classification and regression functions. Discriminative approaches provide a direct attempt to compute the input-to-output mappings for classification and regression and modelling of the underlying distributions. In AV-ASR, hidden Markov Models (HMM), multi-stream HMM [3] or coupled HMM approaches [1] are the methods of choice due to their probabilistic treatment of acoustic coefficients and the Markov assumptions necessary for time varying signals.

However, the extraction of accurate information is a challenge for lip reading as the size of the region of interest (mouth) and small perturbations resulting from lip movements necessitate a high dimensional feature space, which makes the generative model usually very hard to learn. Due to the difficulty to train such models with high dimensional spaces, in this paper we adopt a mixture of Support Vector Machines (SVM). We extract confidence factors from two single-modality classifiers and use those values to select the appropriate fusion strategy. Usually, the outcomes of SVMs are distances in a metric space, which have no simple interpretation and no calibration. To overcome this problem, we use SVMs with probabilistic outputs. We use two single modality classifiers to obtain confidence factors, which are generated from the probabilistic outputs. The latter are produced from a parametric form of a sigmoid, fitted using maximum likelihood estimation [4]. Then the confidence factors are used to select the relevant fusion strategy. The integration is performed using kernel Canonical Correlation Analysis. Finally, recognition is performed by a joint audio-visual SVMs.

The paper is organised as follows. A literature review is presented in Section 2. Section 3 outlines a brief description of multi-class SVMs and probabilistic outputs for SVMs. Section 4 details our system, audio-visual fusion with a kernel Canonical Analysis technique, combined with a multiple multi-class SVMs. The experimental results and the evaluation are presented in Section 5. Conclusions are drawn in Section 6.

2 Background

Most existing AV-ASR approaches use a generative model, such as hidden Markov Models (HMM) [3, 5], for capturing temporal information explicitly. The drawback of generative models is that they estimate a distribution over all (input and output) variables, which can become difficult to compute for real-time applications and with high-dimensional feature spaces. Moreover, because all output probabilities should be computed for each incoming feature vector, it is useful to reduce often huge amount of computations which increase with the dimension of the feature space and also with the number of Gaussians. Alternatively, discriminative models, such as Support Vector Machines (SVM), have been used in speech recognition systems. Smith and Gales [6] and Shimodaira *et al.* [7] first investigated the use of Support Vector Machines for speech recognition. They pointed out the dual problem of using discriminative learners for speech recognition: (a) managing the variation of the time duration of the utterances (or words) and (b) dealing with multi-class decisions. Both papers investigate the former issue: Smith and Gales used an extension of the Fisher Kernel, whereas Shimodaira *et al.* used a dynamic time warping algorithm. Gurban and Thiran [8] apply SVMs within the framework of HMM-based speech recognition. However, they simply concatenate audio and visual features to feed a SVM classifier, which do not optimize the use of such a learner. Therefore they obtain better results with a decision fusion strategy.

As mentioned in the previous section, the values produced by SVMs are uncalibrated and do not give any assessment of the quality of the prediction. An extensive research corpus deal with audio-only speech reading. Golowich and Sun [9] use a combination of Support Vectors Classifiers (SVC) and HMM for phoneme recognition. An interpretation of the multiple SV classification as an approximation to multiple logistic smoothing spline regression allow them to recover conditional class probabilities, which are required as inputs to an HMM. Then Ganapathiraju *et al.* [10] extended this approach by creating a hybrid SVM/HMM architecture for speech reading. Gordan *et al.* [11] employ multiple SVM classifiers and integrate them into a Viterbi decoding lattice. Each class trained one SVM and each of their output is converted to a posterior probability. Then the SVM with probabilistic outputs are integrated into Viterbi lattices as nodes. This approach is performed on visual speech recognition only. Other combinations of SVMs and HMMs are employed in [12] and in [13], where a set of SVMs is used to calculate the class posterior probabilities and to share these probabilities among all HMMs. A similar approach is taken in [14], where a parallel mixtures of SVMs is integrated within a HMM framework. The output of the SVM mixtures, used to classify the phonemes, is used to estimate the emission probabilities of the HMMs, which perform the speech recognition. A summary of the speech reading system using SVM is shown in Table 1.

To conclude, only the Gurban and Thiran's technique uses audio and visual modalities for recognition. However, in this approach the two types of features are simply concatenated and therefore the high correlation between the two modalities is not exploited. Moreover, to the best of our knowledge, none of the previous approaches integrates audio and visual signals to extract and use information in degraded conditions. This is the contribution of our work that we present in this paper.

3 Support vector machines for multi-class problems

As discussed in the previous section, generative models can suffer from high dimensional feature spaces, whereas discriminative models could cope in this case. We now first briefly describe SVMs for binary and multi-class problems and then the theory of probabilistic outputs for SVMs.

3.1 Multi-class SVMs

SVMs are supervised learning methods based on the structural risk minimization principle and was initially defined for classifying linearly separable object classes [15]. For any particular set of two-class objects, a SVM finds the unique hyperplane having the maximum margin, which separated *+1* objects and *-1* objects. The hyperplane can be seen as a classifier decision surface. Since most classes is rarely separable, the coordinates of the objects can be mapped into a higher dimensional feature space (a Hilbert space of finite or infinite dimension), where a linear separation is sought (see Figure 1). This mapping is done by one (or several) kernel function ϕ. Polynomials and radial basis functions [16] kernels are the most used. The only difficulty is to identify, for a particular dataset, the correct set of non-linear functions than can perform such a mapping.

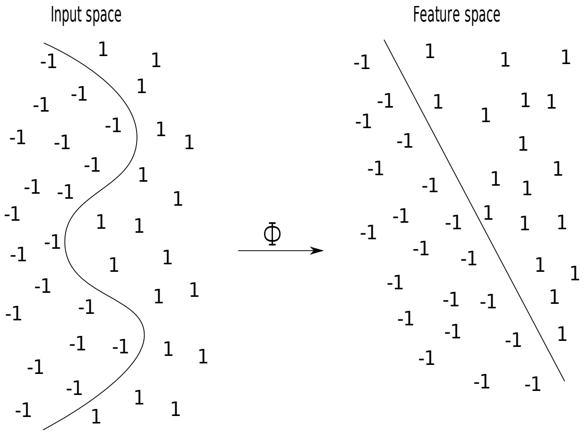

Figure 1: Linear separation of patterns in a two-dimensional feature space

Let us have a set of m training patterns $\{x_m, y_m\}$, where $x = \{x_1, x_2 \ldots, x_n\}$ is a n dimensional pattern and $y_m \in \{+1, -1\}$ represents the labels associated to each pattern. Given a set of feature functions as ϕ_i, such that $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_h(x))$, the class of a pattern x_k is:

$$f(x_k) = \text{sign}[w\phi(x) + b] = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \phi(x_i)\phi(x_k) + b\right) \quad (1)$$

where w are the support vectors, α_i the Lagrange multipliers and b the threshold parameter. However, as most classification problems contain more than two categories, several methods have been proposed to create a *multi-class SVMs* by combining several binary classifiers. Examples of such methods are *one-versus-the-rest*, *one-versus-one* and DAGSVM [17] classifiers. *One-versus-the-rest* trains a 2-class SVM model for all possible pairs of classes from the training set, which for a k-class problem results in a $\frac{K(K-1)}{2}$ SVM models. On the other hand, *one-versus-one* constructs K separate SVMs. The k^{th} SVM classifier is trained with all patterns from the k^{th} class labeled *+1* and all other patterns labeled *-1*. Finally, DAGSVM organises the pairwise classifiers into a direct acyclic graph. Alike *one-versus-the-rest*, for an K-class problem, DAGSVM contains $\frac{K(K-1)}{2}$ classifiers, one for each pair of classes. Other authors also consider all classes at once

Ref	Features		Fusion	Recognition	Database
	audio	visual			
[10]	MFCC Δ MFCC	-	-	SVM/HMM	36 words 10000 sentences (6 words per sentence)
[9]	MFCC	-	-	SVM/HMM	TIMIT
[11]	-	pixel intensities	-	parallel SVM	4 digits (English)
[8]	MFCC Δ MFCC	pixel intensities	straight concatenation	SVM/HMM	-
[14]	cepstral coeff. Frame energy	-	-	mixture of SVM, HMM	40 sentences 72 speakers
[13]	LPC	-	-	SVM/HMM	10 digits (Chineses) 400 utterances
[7]	MFCC Δ MFCC	-	-	DTAK-SVM	6 phonemes 2500 samples
[6]	MFCC Δ MFCC	-	-	Fisher Kernel + SVM	26 letters 300 utterances
Our approach	MFCC + SCF PCA	2D + time SIFT descriptors	kCCA	Multiple SVM	10 digits (Englsih) 2500 utterances

Table 1: A summary of audio-visual speech reading approaches using Support Vector Machines. DTAK: Dynamic Time-Alignment Kernel; HMM: Hidden Markov Model; LPC: Linear Predictive Coding; MFCC: Mel Frequency Cepstral Coefficients; Δ MFCC: 1^{st} and 2^{nd} derivative of Mel Frequency Cepstral Coefficients; kCCA: kernel Canonical Correlation Analysis; SCF: Spectral Crest Factor; SIFT: Scale-Invariant Feature Transform; SVM: Support Vector Machines

[18, 19]. Crammer and Singer [18] described an efficient fixed-point algorithm for solving a quadratic optimization problem in the context of output coding. Weston and Watkins [19] define a single objective function for training all K SVMs simultaneously, based on maximising the margin from each to remaining classes. However the results presented suggest that it performs no better than the more ad-hoc methods of building multi-class classifiers from sets of two-class classifiers.

One-versus-the-rest is less complex and, based on empirical analysis, perform appropriately for our purpose.

3.2 Probabilistic outputs for SVMs

Given a test sample x, the output of SVMs, $f(x)$, provides the distance of x from the separating hyperplane. While the sign of the SVM output determines the class prediction, the magnitude of the SVM output can indicate the confidence level of that prediction. However, as the SVM output is an uncalibrated value, it might not translate directly into a probability value that is useful for estimating confidence. Vapnik [20] mapped the outputs of SVMs to probabilities by decomposing the feature space. However, this approach requires a solution of a linear system for every evaluation of the SVM. Hastie and Tibshirani [21] model probabilities to the output of a SVM by using Gaussians to fit the class-conditional densities $p(f(x)|y = +1)$ and $p(f(x)|y = -1)$, where y is a semantic label. The posterior probability is then computed with the Bayes rule:

$$P(y = 1|f(x)) = \frac{p(f(x)|y = 1)P(y = 1)}{\sum_{i=-1,1} p(f(x)|y = i)P(y = i)}, \qquad (2)$$

where $P(y = i)$ are prior probabilities that are computed from the training set. The posterior probability function in Equation 2 can be seen as a sigmoid with the following analytic form:

$$P(y = 1|f(x)) = \frac{1}{1 + exp(af(x)^2 + bf(x) + c)}. \qquad (3)$$

However, since a SVM is trained to separate the positive samples from the negative ones, we can assume $P(y = 1|f(x))$ to be monotonic in $f(x)$, which is not the case in Equation 3. The reason for this contradiction could be due to the assumption of Gaussian class-conditional probabilities, an assumption that may not always be valid. To overcome this issue, Platt [4] used a parametric model to fit the posterior $P(y = 1|f(x))$ directly, without having to estimate the conditional density $p(f(x)|y)$ for each semantic label y. The Bayes rule from Equation 2 on two exponentials suggests using a parametric form of a sigmoid:

$$P(y = 1|f(x)) = \frac{1}{1 + exp(Af(x) + B)}. \qquad (4)$$

This model assumes that the SVM outputs are proportional to the log odds of a positive example. The parameters A and B of Equation 4 are fitted using maximum likelihood estimation from a training set. More precisely, A and B are obtained by minimizing the negative log likelihood of the sigmoid training data using a model-trust minimization algorithm.

4 Multiple multi-class probabilistic SVMs

We aim to integrate the audio and the visual signals to extract and use information in degraded conditions. To that end, we fuse audio-visual features with a kernel Canonical Correlation Analysis (kCCA) technique [22], combined with a multiple multi-class SVMs. To detect corrupted signals, two classifiers are trained on each modality separately in order to extract confidences factors. The confidences factors are then used to select the most effective strategy to integrate audio and visual features. Finally a last SVMs classify the joint audio-visual space and perform the recognition. We use *one-versus-the-rest* for the three multi-class SVMs as we empirically found that that it performed better than the other approaches discussed above. A linear kernel is used for the audio classifier. The visual and the final classifier are both trained using a radius basis functions kernel. Let A_{test} and V_{test} denote the testing audio and visual feature space, respectively.

$A_{test} = \{a_{test} | a_{test} \in \mathbb{R}^{m_2}\}$ and $V_{test} = \{v_{test} | v_{test} \in \mathbb{R}^{n_2}\}$. Let $P_a(k|A_{test})$ and $P_v(k|V_{test})$ denote the probability estimates that each a_{test} and v_{test} belong to class $k | k \in \{1, ..., K\}$, provided by their respective single-modality SVM. The audio confidence factor, CF_a, is then

$$CF_a = \frac{\sum_i^{M_2} \text{argmax}_{1 \leq k \leq K}(P_a(k|A_{test}))}{M_2}. \quad (5)$$

The visual confidence factor, CF_v, is

$$CF_v = \frac{\sum_i^{N_2} \text{argmax}_{1 \leq k \leq K}(P_v(k|V_{test}))}{N_2}. \quad (6)$$

CF_a and CF_v are used in the fusion process. kCCA provides the canonical factors pairs, W_a and W_v. If R_{va} and R_{av} are the regression matrices calculated from the training set, then we have the following fusion strategies:

$$\begin{aligned} \tilde{V} &= \left(\left(W_a^T A_{test} \right)^{-1} R_{av} \right)^T . \\ \tilde{A} &= \left(\left(W_v^T \tilde{V} \right)^{-1} R_{va} \right)^T \end{aligned} \quad (7)$$

when $CF_v < 0.5 < CF_a$ (noisy visual and clean audio features); or

$$\begin{aligned} \tilde{A} &= \left(\left(W_v^T \Phi(V_{test}) \right)^{-1} R_{va} \right)^T \\ \tilde{V} &= \left(\left(W_a^T \tilde{A} \right)^{-1} R_{av} \right)^T \end{aligned} \quad (8)$$

when $CF_a < 0.5 < CF_v$ (noisy audio and clean video features); and

$$\begin{aligned} \tilde{A} &= \left(\left(W_v^T \Phi(V_{test}) \right)^{-1} R_{va} \right)^T \\ \tilde{V} &= \left(\left(W_a^T A_{test} \right)^{-1} R_{av} \right)^T \end{aligned} \quad (9)$$

in all other conditions (both signals are either degraded or clean). Finally, the joint audio-visual feature vector Z_{test} is a combination of $\tilde{A}$ and $\tilde{V}$:

$$Z_{test} = \begin{pmatrix} \tilde{A} \\ \tilde{V} \end{pmatrix}. \quad (10)$$

Figure 2 shows the block diagram of the proposed framework.

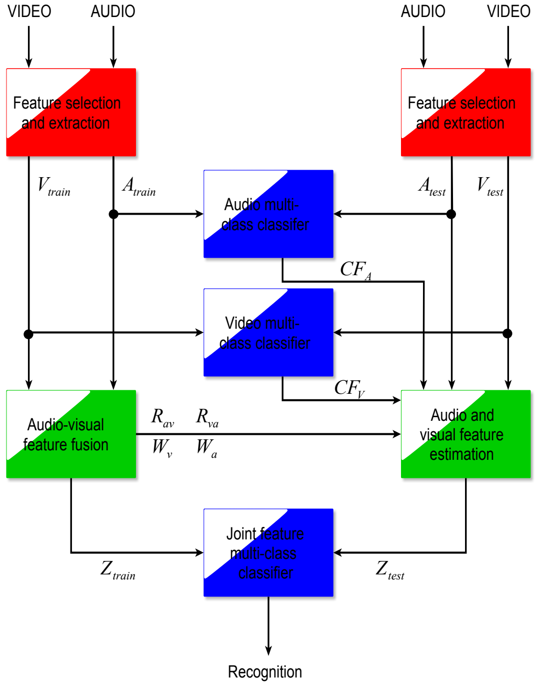
VIDEO
AUDIO
AUDIO
VIDEO
Feature selection and extraction
Feature selection and extraction
V_{train}
A_{train}
Audio multi-class classifer
A_{test}
V_{test}
CF_A
Video multi-class classifier
CF_V
Audio-visual feature fusion
R_{av} R_{va}
W_v W_a
Audio and visual feature estimation
Joint feature multi-class classifier
Z_{train}
Z_{test}
Recognition

Figure 2: Block diagram of the proposed audio-visual recognition system.

5 Experimental Results

5.1 Setup

We evaluate the proposed approach on the CUAVE database [23]. This database consists of 36 speakers pronouncing 10 connected or continuous digits. There are over 2500 utterances of single individuals facing the camera, either moving or still. The audio and visual speech signals are recorded as a sequence of acoustic waveforms (sampled at 16kHz mono) and MPEG-2 files (compressed at 5000 kbps).

We temporally extract the Region-of-Interest (ROI) using a colour Block Matching Algorithm. At initialisation, a manual selection of three points (the tip of the nose and the two corners of the lips) is performed on the first frame only. The ROI is then automatically extracted on the subsequent frames. Our visual representation consists of a visual space-time feature space, which embed the lip movements, using 2D + time SIFT descriptors [24]. As audio features, we use Mel-Frequency Cepstral Coefficients (MFCC) to model the human ear perception and the Spectral Flatness Measure (SFM) to measure the tonality of the signal. Thirteen MFCCs and five SFM coefficients are extracted from the FFT spectrum. Then dimensionality reduction is performed using Principal Component Analysis (PCA).

5.2 Evaluation

To evaluate the use of the residual information available in a degraded or incomplete signals, we tested the proposed approach with two types of *visual degradation*, partial occlusion and salt and pepper noise; and one type of *audio degradation*, additive Gaussian noise. The visual occlusion consists of three fingers covering the frames from the top left corner of the ROI to the bottom right. Different sizes of occlusion, from 8 to 19% of the ROI, are applied. Salt and pepper represents a noise density added to the frames from 0.01 to 0.55. Finally, degraded audio is categorized by signal-to-noise ratios (SNR) from -5dB (very degraded) to 25dB (clean audio). Table 2 summarises the degradation of the recognition rates of the proposed approach when increasing the audiovisual noise. One can observe that the occlusion is less disturbing than the salt and pepper noise. This is due to the spatio-temporal visual features, which can effectively cope with missing data. In heavy noisy conditions in both audio and visual inputs, the recognition rate is still acceptable. The canonical space generated by the training set gives a strong and robust support to the testing set.

Due to the lack of available studies on both audio and visual degraded signals, a complete comparative evaluation is not possible. However we compare our approach with the works presented

		degraded video													
		occlusion (%)							**salt and pepper**						
		8.4	9.7	11.2	12.9	14.5	16.3	18.9	0.01	0.02	0.06	0.10	0.23	0.50	0.55
	25	95.2	95.0	94.9	94.8	93.5	93.3	93.3	95.1	95.1	91.2	91.1	90.9	90.1	89.5
	20	95.1	95.0	94.7	94.5	93.1	93.7	93.9	95.1	92.6	92.2	88.3	87.2	71.8	67.8
degraded audio	15	94.2	94.0	94.7	93.5	93.1	93.7	93.9	96.1	92.6	92.2	88.3	87.2	69.4	60.5
SNR (dB)	10	94.2	94.1	94.8	94.6	94.2	92.8	93.0	95.3	92.6	92.2	88.3	86.5	59.4	50.4
	5	93.4	92.7	92.0	91.6	91.1	89.7	89.6	93.4	92.6	92.2	84.1	72.0	49.4	40.4
	0	93.4	92.7	92.0	91.2	91.1	89.7	89.5	92.6	91.2	91.2	83.5	63.3	43.5	39.0
	-5	93.4	92.7	92.0	91.0	90.3	89.4	89.0	91.6	90.7	88.4	77.0	52.8	38.0	28.3

Table 2: Recognition rate (%) over 10 digits using kCCA using our multiple multi-class strategy

in Section 2 and summarize in Table 1. Most algorithms do not make use of visual cues and none of them are evaluated in noisy conditions. However it is interesting to observe how they perform using a SVM or SVM/HMM recognition system. Table 3 shows six audio speech recognition systems, one visual lip-reading system and two AV-ASR systems. We can observe how the confidence factors (which allow a detection of noisiness) remove the dependency of the results from the degraded audio signal.

	degraded audio - SNR (dB), clean video						
	25	20	15	10	5	0	-5
Ganapathiraju, 2000 [10]	88.4	-	-	-	-	-	-
Golowich, 1998 [9]	54.9	-	-	-	-	-	-
Gordan, 2002 [11]	89.33 (visual feature only)						
Gurban, 2005 [8]	93	92	91	91	83	80	80
Kruger, 2006 [14]	92.23	-	-	-	-	-	-
Qu, 2006 [13]	89	-	-	-	-	-	-
Shimodaira, 2001 [7]	92.3	-	-	-	-	-	-
Smit, 2002 [6]	95.9	-	-	-	-	-	-
Our model	97.3	97.3	97.3	97.3	97.3	97.3	97.3

Table 3: Recognition rate (%) of SVM-based approaches.

6 Conclusions

In this work we have addressed the problem of using residual information in degraded audio-visual signals and have shown the viability of a multiple multi-class SVMs strategy for speech reading. The proposed approach first extracts visual spatio-temporal features and audio cepstral coefficents. Then two classifiers are trained on a single modality to obtain confidence factors are exploited to select the fusion strategy. A third classifier is trained on the joint audio-visual feature space and used to perform digit recognition.

Experimental results demonstrate that our system can efficiently recognize digits in degraded conditions, both in the audio and the visual signals. Moreover, a visual occlusion is less disturbing than salt and pepper noise, thanks to the visual extraction technique and the canonical space generated by the training set. A comparative evaluation also attested how the confidence factors (which allow a detection of noisiness) remove the dependency of the results from the degraded audio signal.

As future work we will explore other types of audio degradation, such as compression and reverberation.

References

[1] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, “Dynamic bayesian networks for audio-visual speech recognition,” *EURASIP*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[2] J. Bernardo, *Encyclopedia of Life Support Systems (EOLSS).* Oxford, UK: UNESCO, 2003, ch. Bayesian statistics.

[3] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus,” *EURASIP*, vol. 2002, no. 11, pp. 1189–1201, 2002.

[4] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[5] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, “A coupled HMM for audio-visual speech recognition,” in *ICASSP*, 2002.

[6] N. Smith and M. Gales, “Speech recognition using SVMs,” in *NIPS*, 2002.

[7] H. Shimodaira, K. Noma, M. Naka, , and S. Sagayama, “Support vector machine with dynamic time-alignment kernel for speech recognition,” in *NIPS*, 2001.

[8] M. Gurban and J. Thiran, “Audio-Visual Speech Recognition with a Hybrid SVM-HMM System,” in *EUSIPCO*, 2005.

[9] S. E. Golowich and D. X. Sun, “A support vector/hidden markov model approach to phoneme recognition,” in *ASA Proceedings of the Statistical Computing Section*, 1998.

[10] A. Ganapathiraju, J. Hamaker, and J. Picone, “Hybrid SVM/HMM architectures for speech recognition,” in *STW*, 2000.

[11] M. Gordan, C. Kotropoulos, and I. Pitas, “A support vector machine-based dynamic network for visual speech recognition applications,” *EURASIP*, vol. 2002, no. 1, pp. 1248–1259, 2002.

[12] J. Stadermann and G. Rigoll, “A hybrid SVM/HMM acoustic modeling approach to automatic speech recognition,” in *INTERSPEECH*, 2004.

[13] Z. Qu, Y. Lui, L. Zhang, and M. Shao, “A speech recognition system based on a hybrid HMM/SVM architecture,” in *ICICIC*, 2006.

[14] S. Krüger, M. Schaffoner, M. Katz, E. Andelic, and A. Wendemuth, “Mixture of support vector machines for HMM based speech recognition,” in *ICPR*, 2006.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, November 1995.

[16] M. J. D. Powell, “Radial basis functions for multivariable interpolation: a review,” *Algorithms for Approximation*, pp. 143–167, 1987.

[17] J. Platt, N. Cristianini, and J. Shawe-taylor, “Large margin DAGs for multiclass classification,” in *NIPS*, 2000.

[18] K. Crammer and Y. Singer, “On the learnability and design of output codes for multiclass problems,” in *CCLT*, 2000.

[19] J. Weston and C. Watkins, “Multi-class support vector machines,” 1998.

[20] V. Vapnik, “An overview of statistical learning theory,” *Neural Networks*, vol. 10, pp. 988–999, 1999.

[21] T. Hastie and R. Tibshirani, “Classification by pairwise coupling,” in *NIPS*, 1998.

[22] S.-Y. Huang, M.-H. Lee, and C. K. Hsiao, “Kernel canonical correlation analysis and its applications to nonlinear measures of association and test of independence,” 2006.

[23] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “Cuave: a new audio-visual database for multimodal human-computer interface research,” in *ICASSP*, 2002.

[24] S. Pachoud, S. Gong, and A. Cavallaro, “Macro-cuboïd based probabilistic matching for lip-reading digits,” in *CVPR*, 2008.