A Temporal Network of Support Vector Machine Classifiers for the Recognition of Visual Speech

Mihaela Gordan[1], Constantine Kotropoulos[2], and Ioannis Pitas[2]

[1] Faculty of Electronics and Telecommunications
Technical University of Cluj-Napoca
15 C. Daicoviciu, 3400 Cluj-Napoca, Romania
mihag@bel.utcluj.ro
[2] Artificial Intelligence and Information Analysis Laboratory
Department of Informatics, Aristotle University of Thessaloniki
Box 451, GR-54006 Thessaloniki, Greece
{costas, pitas}@zeus.csd.auth.gr

Abstract. Speech recognition based on visual information is an emerging research field. We propose here a new system for the recognition of visual speech based on support vector machines which proved to be powerful classifiers in other visual tasks. We use support vector machines to recognize the mouth shape corresponding to different phones produced. To model the temporal character of the speech we employ the Viterbi decoding in a network of support vector machines. The recognition rate obtained is higher than those reported earlier when the same features were used. The proposed solution offers the advantage of an easy generalization to large vocabulary recognition tasks due to the use of viseme models, as opposed to entire word models.

1 Introduction

Visual speech recognition refers to the task of recognizing the spoken words based only on the visual examination of the speaker's face. This task is also referred as lipreading, since the most important visible part of the face examined for information extraction during speech is the mouth area. Different shapes of the mouth (i.e. different mouth openings, different position of the teeth and tongue) realized during speech cause the production of different sounds. One can establish a correspondence between the mouth shape and the phone produced, even if this correspondence will not be one-to-one, but one-to-many, due to the fact that invisible parts of the vocal tract are also involved in speech production as well. For small size word dictionaries, we can still perform good quality speech recognition using the visual information regarding the mouth shape only.

So far, many methods have been reported in the literature for solving the visual speech recognition problem. The different types of solutions adopted vary widely with respect to: 1) the feature types; 2) the classifier used; and 3) the class definition. For example, Bregler uses time-delayed neural networks (TDNN) for visual classification, and the outer lip contour coordinates as visual features [6].

Luettin uses active shape models for representing different mouth shapes, gray level distribution profiles (GLDPs) around the outer and/or inner lip contours as feature vectors, and finally builds whole-word hidden Markov models (HMMs) for visual speech recognition [7]. Movellan employs also HMM for building visual word models, but using as features directly the gray levels of the mouth images, after some simple preprocessing to exploit the vertical symmetry of the mouth [5].

Despite the big variety of existing strategies for visual speech recognition, there is still ongoing research in this area, attempting: 1) to find the most suitable features and classification techniques to discriminate efficiently between the different mouth shapes, but to keep the mouth shapes corresponding to the same phone produced by different individuals in the same class (i.e., to develop speaker independent techniques); 2) to require limited processing of the mouth image so that the implementation of the mouth shape classifier in real time is feasible; 3) to facilitate the easy integration of audio and video speech recognition.

In this paper, we aim to contribute to the first of the above mentioned aspects in visual speech recognition, by examining the suitability of a new type of classifiers for visual speech recognition tasks, the support vector machines (SVMs). We are motivated by the success of SVMs in various pattern recognition applications including visual classification tasks such as biometric person authentication, medical image processing, etc.

The use of SVMs as classifiers for automatic speech recognition is a new idea. Very good results in audio speech recognition using SVMs were recently reported in [1]. No attempts in applying SVMs for visual speech recognition have been reported so far, although a somehow closely related application is described in [11], where SVMs were applied for detecting the degree of opening/smile of mouth images in videosequences. This work uses SVMs for linear regression, not for classification task. Thus, according to the best of the author's knowledge, the use of SVMs as visual speech classifiers is a novel idea. Regarding SVMs applications as visual classifiers, there are some very good results in face detection and face recognition [2, 3] and in dynamical object detection in videosequences [13].

One of the reasons for not using SVMs in automatic speech recognition so far is the fact that they are inherently static classifiers, whilst speech is a dynamic process, where the temporal information is essential for recognition. This means one cannot use directly SVMs for speech recognition. A solution to this problem is presented in [1], where a combination of HMM and SVM is proposed. In this paper we adopt a similar strategy for modeling the visual speech dynamics with the difference that we shall use only the Viterbi algorithm to create dynamical visual word models.

Another novel aspect in the visual speech recognition approach proposed here refers to the strategy adopted for building the word models: while most of the applications presented in the literature [1, 7, 5] build whole word models as basic visual models, our basic visual models are mouth shape models (viseme models), and the visual word model is obtained by the combination of these basic models

into a temporal dynamic sequence. This approach offers the advantage of an easy generalization to large vocabulary word recognition tasks without a significant increase in storage requirements by maintaining the dictionary of basic visual models needed for word modeling to a reasonable limit.

The visual speech recognition results obtained are very promising as compared to similar approaches reported in the literature. This shows that SVMs are a promising alternative for visual speech recognition and encourages the continuation of the research in this direction.

The outline of the paper is as follows. Section 2 details the proposed visual speech recognition using SVMs. The modeling of temporal speech dynamics is described in Section 3. Experimental results are presented in Section 4 and conclusions are drawn is Section 5.

2 Description of the proposed visual speech recognition approach using support vector machines

The problem of discriminating between different shapes of the mouth during speech production, the so-called *visemes*, can be viewed as a pattern recognition problem. In this case the feature vector comprises a representation of the mouth image, either low-level at pixel-level, or by extracting several geometric parameters, or by applying some linear transform of the mouth image. The different pattern classes are the different mouth shapes occurred during speech. For example, in the case of producing the sound "o", the mouth will have an open-rounded shape, while for example in the case of sound "f", the mouth will have an almost closed position, not rounded, the upper teeth will be visible and the lower lip will be moved inside.

Obtaining the phonetic description of each word from a possible dictionary is a simple task, and there are currently many publicly available tools to do this. Correlations can be established between the different phones produced during speech and the visemes corresponding to them. However, this correspondence is not one-to-one, since non-visible parts of the vocal tract are also involved in speech production, and even more, it depends on the nationality of the different speakers given the fact that the pronunciation of the same word varies and is not always according to the "standard" one. Furthermore, although there are phoneme-to-viseme correspondence tables available in the literature [4], currently there is not a universally accepted mapping, as in the case of phonemes (cf. [12]). The solution adopted here is to define the viseme classes and the viseme-to-phoneme mapping dependent on the application (i.e., the recognition of the first four digits in English, as spoken by the different individuals in the Tulips1 database [5]). The viseme classes defined and their corresponding phonemes are presented in Table 1.

Once we have defined the mapping between the classes of visemes needed in our application and their corresponding phonemes based on the phonetic description of each word from the dictionary, we can build the *visemic models* of the words as sequences of mouth shapes which could produce the phonetic

Table 1. Viseme-to-phoneme mappings for the first four digits.

Phoneme	Corresponding viseme classes
W	w (small rounded open mouth state)
	ao (larger rounded open mouth state)
	wao (medium rounded open mouth state)
AH	ah (medium ellipsoidal mouth state)
N	n (medium open, not rounded, mouth state; teeth visible)
T	t (medium open, not rounded, mouth state; teeth and tongue visible)
UW	SAME AS W
TH	$th_{1,2}$ (medium open, not rounded)
R (context C-**C**-V)	w (small rounded open mouth state)
	ao (larger rounded open mouth state)
IY	iy (longitudinal open mouth state)
	ah (medium ellipsoidal mouth state)
F	$f_{1,2,3}$ (almost closed position; upper teeth visible; lower lip moved inside)
AO	SAME AS W

realizations of the words. Thus, for the small four word dictionary of the first four digits in English from our application, we have the phonetic and the visemic models given in Table 2.

SVMs is a principled technique to train classifiers that stems from statistical learning theory [8, 9]. Their root is the optimal hyperplane algorithm. They minimize a bound on the empirical error and the complexity of the classifier at the same time. Accordingly, they are capable of learning in sparse high-dimensional spaces with relatively few training examples. Let $\{\mathbf{x}_i, y_i\}$, $i = 1, 2, \ldots, N$, denote N training examples where $\mathbf{x}_i$ comprises an M-dimensional pattern and y_i is its class label. Without any loss of generality we shall confine ourselves to the two-class pattern recognition problem. That is, $y_i \in \{-1, +1\}$. We agree that $y_i = +1$ is assigned to positive examples, whereas $y_i = -1$ is assigned to counterexamples.

The data to be classified by the SVM might be linearly separable in their original domain or not. If they are separable, then a simple linear SVM can be used for their classification. However, the power of SVMs is demonstrated better in the nonseparable case, when the data cannot be separated by a hyperplane in their original domain. In the latter case, we can project the data into a higher dimensional Hilbert space and attempt to linearly separate them in the higher dimensional space using kernel functions. Let Φ denote a nonlinear map $\Phi : \mathcal{R}^M \rightarrow \mathcal{H}$ where $\mathcal{H}$ is a higher-dimensional Hilbert space. SVMs construct the optimal separating hyperplane in $\mathcal{H}$. Therefore, their decision boundary is of the form:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i \; y_i \; K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{1}$$

Table 2. Phonetic and visemic description models of the four spoken words from Tulips1 database.

Word	Phonetic model	Visemic models
"one"	W-AH-N	w-ah-n
		ao-ah-n
		wao-ah-n
"two"	T-UW	t-w
		t-wao
		t-ao
"three"	TH-R-IY	$th_{1,2}$-w-iy
		$th_{1,2}$-w-ah
		$th_{1,2}$-ao-iy
		$th_{1,2}$-ao-ah
		$th_{1,2}$-iy
"four"	F-AO-R	$f_{1,2,3}$-ao
		$f_{1,2,3}$-w
		$f_{1,2,3}$-wao
		$f_{1,2,3}$-ao-ah

where $K(\mathbf{z}_1, \mathbf{z}_2)$ is a kernel function that defines the dot product between $\Phi(\mathbf{z}_1)$ and $\Phi(\mathbf{z}_2)$ in $\mathcal{H}$, and α_i are the nonnegative Lagrange multipliers associated with the quadratic optimization problem that aims to maximize the distance between the two classes measured in $\mathcal{H}$ subject to the constraints

$$\begin{aligned} \mathbf{w}^T \Phi(\mathbf{x}_i) + b \geq 1 \ \text{ for } y_i = +1 \\ \mathbf{w}^T \Phi(\mathbf{x}_i) + b \leq 1 \ \text{ for } y_i = -1. \end{aligned} \tag{2}$$

The sign function in the decision boundary (1) simply makes the optimal separating hyperplane an indicator function. In the following we will omit this sign function and use as the output of the SVM classifier the real valued function:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \ y_i \ K(\mathbf{x}, \mathbf{x}_i) + b, \tag{3}$$

as a measure of confidence in the class assignment.

A single SVM can recognize a single mouth shape. To recognize all the mouth shapes we shall need to define and train one SVM classifier for each mouth shape and to arrange the SVMs in a parallel structure. The input mouth image is simultaneously presented to the input of all the SVMs and each of them gives a real output value showing the confidence in assigning the mouth shape in the corresponding class. Figure 1 depicts the topology of SVM network built.

The selection of the type of feature vector to be classified by the SVMs takes into account that by their nature SVMs have the ability of separating the input data into classes even when the correlation among the data and the dimensionality of the feature vector is high, due to the projection of the data into a higher dimensional space performed inside the SVM. This allows us to

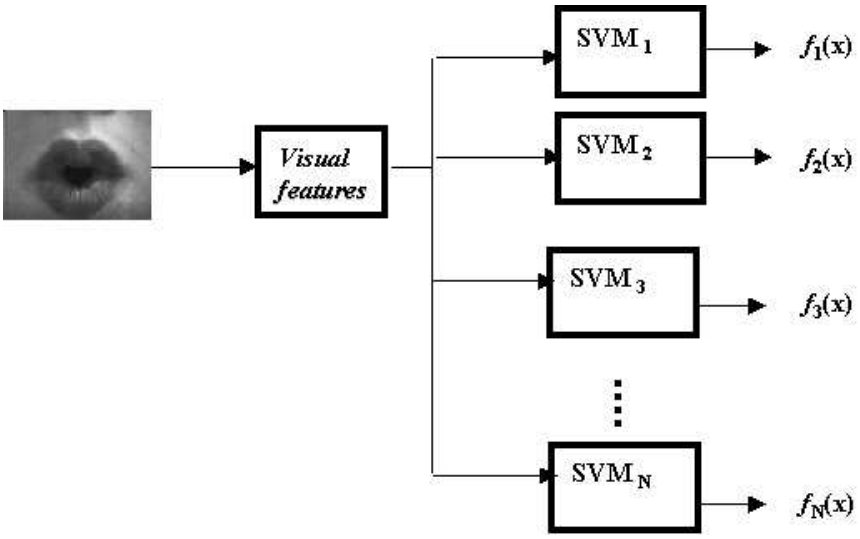

Fig. 1. Topology of SVM network used for visual speech recognition

use very simple features to represent the mouth image, e.g. pixel-level features. As a consequence, we decided to use as feature vector for the mouth image whose shape we want to recognize, the vector comprising the gray levels of the pixels from the mouth image, scanned in row order. The labeling of the mouth images is done manually. To ensure a good training, only the unambiguous positive and negative examples are included in the training set of each SVM. Preprocessing of the mouth images from Tulips1 was needed due to the fact that the mouth has different scale, position in the image and orientation towards the horizontal axis from utterance to utterance, varying with the position of the subject in front of the camera. To compensate for these variations we applied the normalization procedure of mouth images with respect to scale, translation and rotation described in [7].

3 Modeling the temporal dynamics of visual speech

In every audiovisual speech sequence, a word is described as a sequence of phonemes in the audio domain and visemes in the video domain covering a number of frames. The symbolic phonetic/visemic models show only the sequence of the different symbols in a word realization without specifying the duration of each symbol, as this is strongly person-dependent.

The most natural way of representing the word models in the temporal domain, starting only from the symbolic visemic model and from the total number of T frames in the word pronunciation, is to assume that the duration of each viseme in the word pronunciation can be whatever, but *necessarily not zero.*

Thus, we can create a temporal network of models corresponding to the different possible durations of the visemes in the model, containing as many states as many frames we have in the videosequence, that is, T. The most straightforward way to represent such a network of models is the Viterbi algorithm [14]. One of the possible visemic models and the resulting Viterbi lattice are shown in Figures 2 and 3 for the example of the word "one", where the visemes present in the word pronunciation have been denoted according to Table 1. The paths formed by the solid lines in the Vitterbi lattice from Figure 3 show the possible model realizations. Each node of the Vitterbi lattice in Figure 3 signifies the realization of the corresponding viseme at that particular time instant. Each visemic word model from the set of D visemic description models of the four words in the dictionary, given in Table 2, w_d, $d = 1, 2, \ldots, D$, will have its own Viterbi lattice model. In the current application, $D = 15$.

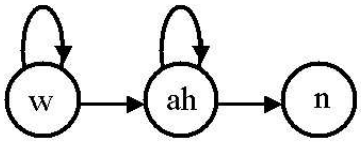

Fig. 2. Temporal sequence for the pronunciation of the word "one"

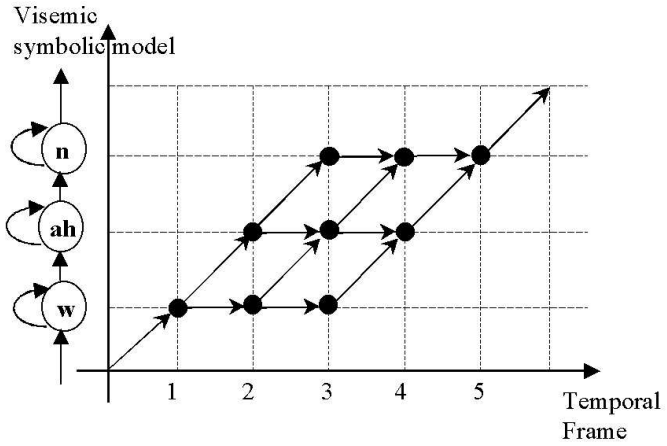

Fig. 3. The temporal Viterbi lattice for the pronunciation of the word "one" in a videosequence of 5 frames

Let us interpret each node in the lattice of Figure 3 as a measure of confidence that the corresponding symbol o_k is emitted at the time instant k. We denote this

measure of confidence by $c_{o_k k}$. Each solid line between the nodes corresponding to the symbol o_k at time instant k and o_{k+1} at time instant $k+1$ represents the transition probability from the state that is responsible for the generation of o_k to the state that generates the symbol o_{k+1}. We denote the latter probability by $a_{o_k o_{k+1}}$, where o_k and o_{k+1} may be different or not. To a first approximation, we assume equal transition probabilities $a_{o_k o_{k+1}}$ between whatever two symbol emission states. Thus, they do not contribute to differentiate between the costs of following different paths in the Viterbi lattice.

Having a videosequence of T frames for a word pronounced and such a Viterbi model for each visemic word model w_d, $d = 1, 2, \ldots, D$, we can compute the confidence for the visemic word model w_d to be produced following a path ℓ in the Viterbi lattice as:

$$c_{d,\ell} = \sum_{k=1}^{T} c_{o_k k} \,|d, \ell \ , \tag{4}$$

independent of $a_{o_k o_{k+1}}$, and the confidence score that the visemic word model w_d was produced is the maximum over all possible $c_{d,\ell}$. Among the words that can be produced following all the possible paths in all the D Viterbi lattices, the most plausible word, that is, the one corresponding to the visemic model with the maximum confidence score c_d, $d = 1, 2, \ldots, D$, is finally recognized. In the visual speech recognition approach discussed in this paper, the symbol emission measures of confidence $c_{o_k k}$ are given by the corresponding SVMs, SVM_{o_k}.

4 Experimental results

To evaluate the recognition performance of the proposed SVM-based visual speech recognizer, we choose to solve the task of recognizing the first four digits in English. As experimental data we used the small audiovisual database Tulips1 [5], frequently used in similar visual speech recognition experiments. The phonetic and visemic description of the four words and the phoneme to viseme mapping for this application are given in Tables 1 and 2. The visual speech recognizer requires the training of 12 different SVMs, one for each distinct mouth shape considered in the Table 1. We used for our experiments SVMs with a polynomial kernel of degree 3. For the training of the SVMs we used the publicly available SVMLight toolkit [10]. The complete visual speech recognizer was implemented in C++ programming language. In the module implementing the Viterbi decoder for all the possible visual word models, the SVM classifiers in the nodes of a Viterbi decoder were implemented using the classification module of the SVMLight toolkit. We performed speaker-independent visual speech recognition tests, using the leave-one-out testing strategy for the 12 subjects in the Tulips1 database. More precisely, the testing strategy was as follows: we trained the system 12 times separately, each time using 11 subjects in the training set and leaving the $12th$ subject out for testing. In this way, we obtained actually 24 test sequences per word, due to the fact that Tulips1 database contains 2 pronunciations per subject for each word (Set1 and Set2). This gives a total of 24×4 words $= 96$ video test sequences.

We examine the overall word recognition rate (WRR) comparing this result with those reported in literature under similar conditions (i.e., using the same features, the same database and the same testing procedure) [7, 5] in Table 3.

Table 3. The overall WRR of the proposed system of SVM classifiers as compared to other techniques (without delta features)

Method	Dynamic SVM network (our method)	Stochastic networks [5]	AAM and HMM shape model inner+ outer lip contour [7]	AAM and HMM intensity model outer lip contour [7]
WRR [%]	76	60	75	65.6

We can see that, for similar features used, our system achieves a slightly higher word recognition performance than those reported in the literature. The WRR is lower than the best rate reported without delta features in [7], i.e., 87.5 %, where the shape + intensity information is used with the inner and outer lip contour model. In the latter model, the intensity is sampled in the exact subregion of the mouth image comprising the lips and not including the skin areas. However, the computational complexity of this method is higher to that of our solution, due to the need for re-definition of the region of interest at each frame.

To assess the statistical significance of the rates observed, we model the ensemble {test patterns, recognition algorithm} as a source of binary events, 1 for correct recognition and 0 for an error, with probability p of drawing a 1 and $(1-p)$ of drawing a 0. These events can be described by Bernoulli trials. Let us denote by $\hat{p}$ the estimate of p. The exact ϵ confidence interval of p is the segment between the two roots of the quadratic equation [15]:

$$(p-\hat{p})^2 = \frac{z^2_{(1+\epsilon)/2}}{K}\, p\,(1-p) \tag{5}$$

where z_u is the u-percentile of the standard Gaussian distribution having zero mean and unit variance, and $K = 96$ is the total number of tests conducted. We computed the 95% confidence intervals ($\epsilon = 0.95$) for the WRR of the proposed approach and also for the WRRs reported in literature [7, 5], as summarized in Table 4.

5 Conclusions

In this paper we examined the suitability of SVM classifiers in visual speech recognition. Due to the inherent temporal dependency of the speech, we also propose a solution to build a dynamic SVM-based classifier. We tested the proposed method on a small visual speech recognition task, namely, the visual recognition

Table 4. Confidence interval for the WRR of the proposed system of SVM classifiers as compared to other techniques (without delta features)

Method	Dynamic SVM network (our method)	Stochastic networks [5]	AAM and HMM shape model inner+outer lip contour [7]	AAM and HMM intensity model outer lip contour [7]
Confidence interval [%]	[66.6%;83.5%]	[49.9%;69.2%]	[65.5%;82.5%]	[55.6%;74.3%]

of the first four digits in English. The features used are the simplest possible: directly the raw gray level values of the mouth image. Under these circumstances, we obtained good word recognition rates as compared to the similar results from the literature. This shows that SVMs are promising classifiers for visual speech recognition tasks. Another advantage of the viseme-oriented modeling method proposed here is the possibility of easy generalization to large vocabularies. The existing correlation between the phonetic and visemic models can also lead to an easy integration of the visual speech recognizer with its audio counterpart. In our future research, we will try to enhance the performance of the visual speech recognizer by including delta features in the feature vector, by using other type of kernel functions and by including the temporal constraints at symbol level in the temporal word models trough the learning of the state transitions probabilities for the Vitterbi decoding lattice.

Acknowledgement

This work has been supported by the European Union funded Research Training Network on "Multi-modal Human-Computer Interaction" (HPRN-CT-2000-00111).

References

1. Ganapathiraju, A., Hamaker, J., Picone, J.: Hybrid SVM/HMM architectures for speech recognition. Proc. of Speech Transcription Workshop. College Park, Maryland, USA (May 2000).
2. Yongmin, Li, Shaogang, Gong, Liddell, H.: Support vector regression and classification based multi-view face detection and recognition. Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition. Grenoble, France (March 2000) 300–305.
3. Terrillon, T.J., Shirazi, M. N., Sadek, M., Fukamachi, H., Akamatsu, S.: Invariant face detection with support vector machines. Proc. 15th Int. Conf. Pattern Recognition. Barcelona, Spain. **4** (September 2000) 210–217.
4. Chen, T.: Audiovisual speech processing. IEEE Signal Processing Magazine. **18(1)** (January 2001) 9–21.
5. Movellan, J. R.: Visual speech recognition with stochastic networks. In: Tesauro, G., Toruetzky, D., Leen, T. (eds.): Advances in Neural Information Processing Systems. **7**. MIT- Press, Cambridge, MA (1995).

6. Bregler, C., Omohundro, S.: Nonlinear manifold learning for visual speech recognition. Proc. IEEE Int. Conf. Computer Vision (1995) 494–499.
7. Luettin, J., Thacker, N. A.: Speechreading using probabilistic models. Computer Vision and Image Understanding. **65(2)** (February 1997) 163–178.
8. Vapnik, V.N.: Statistical Learning Theory. J. Wiley, N.Y. (1998).
9. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, U.K. (2000).
10. Joachims, T.: Making large-scal SVM learning practical. In: Schoelkopf, B., Burges, C., Smola, A. (eds.): Advances in Kernel Methods - Support Vector Learning. MIT-Press (1999)
11. Kumar, V. P., Poggio, T.: Learning-based approach to real time tracking and analysis of faces. Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition. Grenoble, France (March 2000) 96–101.
12. Ezzat, T., Poggio, T.: MikeTalk: A talking facial display based on morphing visemes. Proc. Computer Animation Conference. Philadelphia, Pennsylvania (June 1998).
13. Papageorgiou, C., Poggio, T.: A pattern classification approach to dynamical object detection. Proc. IEEE Int. Conf. Computer Vision. **(2)** (1999) 1223–1228.
14. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev V., Woodland, P.: The HTK Book. HTK version 2.2. Edition. Entropic, Ltd., Cambridge, UK (1999).
15. Papoulis, A.: Probability, Random Variables, and Stochastic Processes. 3^{rd} Edition. McGraw-Hill (1991)