CEPSTRAL FEATURES FOR CLASSIFICATION OF AN IMPULSE RESPONSE WITH VARYING SAMPLE SIZE DATASET

Cyril Hory[♯]*, and William J. Christmas*[♮]

[♯]Département Traitement du Signal et des Images, CNRS/GET-ENST (Télécom Paris)
37-39 rue Dareau, 75014 Paris, France
hory@enst.fr

[♮]Center for Vision Speech and Signal Processing, University of Surrey
Guildford GU2 7XH, UK
W.Christmas@surrey.ac.uk

ABSTRACT

Cepstrum-based features have proved useful in audio and speech characterisation. In this paper a feature vector of cepstral polynomial regression is introduced for the detection and classification of impulse responses. A recursive algorithm is proposed to compute the feature vector. This recursive formulation is appealing when used in a sequential learning framework. The discriminative power of these features to detect and isolate racket hits from the audio stream of a tennis video clip is discussed and compared with standard cepstrum-based features. Finally, a new formulation of the Average Normalised Modified Retrieval Rank (ANMRR) is proposed that exhibits relevant statistical properties for assessing the performance of a retrieval system.

1. INTRODUCTION

In digital video analysis it is now recognised that audio cues extracted from a video clip, along with the visual cues can provide relevant information for the semantic understanding of the video content [14]. The audio-visual cooperation is insured through multi-modal conditional density estimation in [3]. Mel-Frequency Cepstral Coefficients (MFCC) are used in [7] to identify specific sounds in a baseball game in order to detect commercials, speech or music using the maximum entropy method. Most proposed methods combine audio and visual features within a HMM framework [2, 8]. The common approach is to consider a large set of audio and visual features and select the most relevant ones [5] during the training step.

Working with a high dimensionality feature space requires a large sample size training set [11]. However there are many situations where the sample size of the dataset cannot be chosen as large as desired. When a fast decision is to be taken a sequential classification scheme bearing optimality properties in terms of required sample size [13], is convenient. In the process of building up ground truth, the size of the dataset is small at first and increases in the course of annotation.

The scope of this paper is sequential learning monitored from the audio stream. In audio-based classification or indexation homomorphic features are widely used. Homomorphic processing was introduced to deconvolve the source and channel of a system in the cepstral domain [4]. The First Cepstral Coefficients (FCC) - similar to the MFCC's computed on the cognitively relevant mel-frequency scale - encode information about the spectral envelope of the analysed signal. In speech processing FCC's or MFCC's characterise the resonant system created by the mouth and lips of the speaker. Using the mel-frequency scale, it is usually found that 12 MFCC's are necessary for signal characterisation (see [14] for example). Seting the dimensionality of the feature space to 12 components requires a sufficiently large sample size learning set. Moreover in the problem of classifying waveforms extracted from a single audio stream, the characterization of the channel which models the same propagating medium and recording device is not relevant. Thus again, FCC's and MFCC's are not efficient features for such a classification problem.

In this paper we introduce a new set of cepstral features based on a polynomial regression of the cepstrum called Cepstral Regression Coefficients (CRC). We propose a recursive computation of the CRC's which potentially enables the convenient updating of the dimensionality of the feature space when the sample size of the training set increases. We focus on the cepstral characterisation of an impulse response with low dimensional extracted features since the impulse response waveform only encodes information about the system. The discriminating power of the proposed features is experimentally evaluated in an experiment of classification of racket hit waveforms extracted from the audio stream of a tennis video clip. Assessment is performed using ROC curves and a new formulation of the Average Normalised Modified Retrieval Ranking (ANMRR) [9, 10], introduced for its convenient statistical properties.

PART OF THIS WORK WAS CARRIED OUT DURING THE TENURE OF A MUSCLE INTERNAL FELLOWSHIP AND SUPPORTED BY INFOM@GIC PROJECT.

2. FEATURE EXTRACTION

2.1 Data model

Consider the $N \times 1$ data representation $c = [c_1, c_2, \ldots, c_N]^T$ indexed by the $N \times 1$ vector $q = [q_0, q_1, \ldots, q_{N-1}]^T$ such that

$$c = Q_{(p)} a_{(p)} + \varepsilon, \tag{1}$$

where $a_{(p)}$ is a $(p+1) \times 1$ vector, $Q_{(p)}$ is a $N \times (p+1)$ matrix with element $g_{j-1}(q_{i-1})$ on the ith row and jth column where g_j are polynomials of order j, and ε is a $N \times 1$ vector of random perturbations.

2.2 Proposed features

The minimum mean-square estimation of the so-called vector of regressive coefficients $a_{(p)}$ is:

$$\hat{a}_{(p)} = R_{(p)}^{-1} Q_{(p)}^T c, \tag{2}$$

where $R_{(p)} = Q_{(p)}^T Q_{(p)}$ is a $(p+1)\times(p+1)$ matrix with elements

$$r_{ij} = \sum_{k=0}^{n-1} g_{i-1}(q_k) g_{j-1}(q_k), \tag{3}$$

on the ith row and jth column.
Let the data representation c be the vector of cepstral magnitudes of a time sequence e [4]:

$$c = |\mathrm{FT}^{-1}\{\log(|\mathrm{FT}\{e\}|)\}|, \tag{4}$$

where $\mathrm{FT}\{.\}$ is the discrete Fourier transform. We propose to take the coefficients $\hat{a}_{(p)}$ in (2) as descriptors of the cepstrum content of the detected events.
Figure 1 shows the cepstral magnitudes of a racket hit and crowd applause extracted from the audio stream of a broadcasted tennis game. A racket hit can be seen as the impulse response of a system since the source (the ball hitting the racket) can be modeled by a Dirac impulse although the source of crowd applause can be modeled by a train of Dirac impulses. The cepstra are computed over 256 quefrency bins. Regressions of order 2, 5 and 10 are superimposed. The slope at the origin of the polynomial of order 2 of the racket hit is stronger than for the crowd applause. Applause generates a cepstrum content resembling that of white noise with a sharp peak centered at the null quefrency. Low order regressions such as the one presented in this figure are not able to encode this low quefrency information. However, the regression coefficients of the crowd applause are more sensitive to local variations in the mid-range quefrencies.

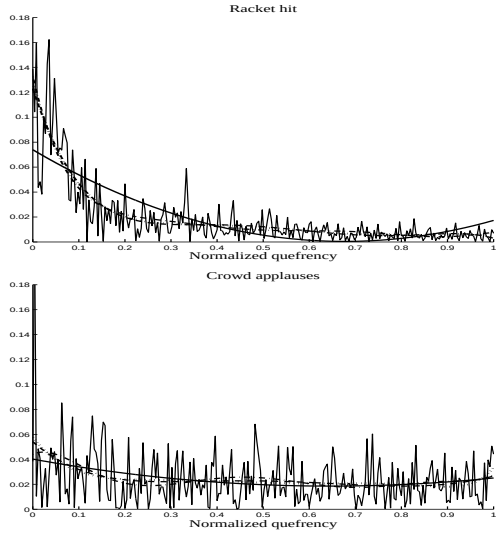

Figure 1: Cepstral regression of order 2 (plain line), 5 (dashed line) and 10 (dash-dotted line) of two typical events of a tennis rally: a racket hit (a), and crowd applause (b).

2.3 Recursive polynomial regression

Denote $u_i = \sum_{k=0}^{n-1} g_{i-1}(q_k) c_k$ the ith element of vector $u_{(p)} = Q_{(p)}^T c$. One have the simple recursive concatenation relation:

$$u_{(p)} = [u_{(p-1)}^T, u_p]^T. \tag{5}$$

Similarly, the matrix $R_{(p)}$ can be expressed in terms of $R_{(p-1)}$:

$$R_{(p)} = \begin{bmatrix} R_{(p-1)} & r_p \\ r_p^T & r_{p+1,p+1} \end{bmatrix}, \tag{6}$$

where r_p is the $1\times p$ vector $r_p = [r_{1,p+1}, r_{2,p+1}, \ldots, r_{p,p+1}]^T$ of coefficients (3).
The block-matrix inversion formula (see for example [12]) gives:

$$R_{(p)}^{-1} = \begin{bmatrix} R_{(p-1)}^{-1} + R_{(p-1)}^{-1} r_p s_{(p)}^{-1} r_p^T R_{(p-1)}^{-1} & -R_{(p-1)}^{-1} r_p s_{(p)}^{-1} \\ -s_{(p)}^{-1} r_p^T R_{(p-1)}^{-1} & s_{(p)}^{-1} \end{bmatrix}, \tag{7}$$

where $s_p = r_{p+1,p+1} - r_p^T R_{(p-1)}^{-1} r_p$ is the Schur complement of $R_{(p-1)}$. By inserting (5) and (7) into (2), the updated estimate of the regression coefficient vector takes the form:

$$\begin{aligned} \hat{a}_{(p)} &= R_{(p)}^{-1} u_{(p)}, \\ \hat{a}_{(p)} &= \begin{bmatrix} I + R_{(p-1)}^{-1} r_p s_{(p)}^{-1} r_p^T & -R_{(p-1)}^{-1} r_p s_{(p)}^{-1} \\ -s_{(p)}^{-1} r_p^T & s_{(p)}^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \hat{a}_{(p-1)} \\ u_p \end{bmatrix}. \end{aligned} \tag{8}$$

First, the order $p=1$ CRC's $\hat{a}_{(1)}$ are computed using (2) from the estimated magnitude cepstrum c defined in (4). This operation requires the inversion of the 2×2 matrix $R_{(1)}$. In a later step the computation of the order p CRC's is performed by means of the order $p-1$ CRC's using recursion (8).
The recursive computation of the CRC's is a fast method for extracting the proposed cepstral-based features. Moreover the recursive design is well-fitted to sequential schemes where data are available sequentially.
The next section is devoted to the evaluation of the capability of the CRCs to encode the characteristic feature of the magnitude cepstra of an impulse response.

3. EXPERIMENTAL EVALUATION

To evaluate the efficiency of the CRCs for discriminating impulse responses, a Biased Discriminant Analysis [15] has been performed within a supervised learning framework. We report below the result of the experiment carried on two training sets of different size in order to show the role played by the dimension of the feature space and the need to match the dimension of the feature space to the size of the training set.

3.1 Experimental setup

Two classification experiments were carried using the CRCs and FCCs. For each set of features the biased discriminant transform was computed from the training set and applied to the same set in order to estimate the likelihood function of the class of impulse response. A Gaussian model was assumed in the transformed feature space. The test set was then transformed onto the discriminative space and the likelihood function was computed. ROC curves were computed by successive thresholding of the likelihoods.

3.1.1 Dataset

The experiment has been carried on the second, third and fourth game of A. Agassi and R. Schuettler in the Australian

Open final tennis game of 2003. The match is played on a synthetic surface.
Events were detected by a CUSUM test of a change in the variance of the audio stream [1] assuming white Gaussian noise for silent segments (see [6] for details). Each segmented event starts at the local maximum of the waveform and is 20ms long.
In the first experiment the events extracted from the second game were used as the training set. A total of 203 events were detected including 20 actual racket hits. This is the *small size training set*. In the second experiment, the events detected in the second and third games were used as the *large size training set*. This set is composed of 759 events including 93 racket hits.
In both experiments the test data was collected from the fourth game. A total of 561 events were detected by the CUSUM test including 75 racket hits.
The events detected by performing the sequential test are *racket hits*, *echoes*, *ball bounces*, *shoe shuffles*, *voice (comments, player's shouts, umpire speaking)*, and *crowd noise*. Even though bounces and segmented echoes could be considered as impulse responses, the aim of the experiment was to classify racket hits.

3.1.2 Extracted features

The estimated magnitude cepstra c are vectors of $N = 256$ normalised quefrency bins $q_i = i/2N$. The CRCs are computed with polynomials $g_j(q_i) = q_i^j$ indexed by normalised quefrencies.

3.1.3 Biased discriminant analysis

The objective of the experiment was to assess the capability of the Cepstral Regression Coefficients (CRC) to discriminate and identify racket hits among the different types of detected events. For this purpose it is relevant to split the data set into the class of racket hits and a single class of all other events. The class of racket hits can be considered as homogeneous whereas the other class is more likely to be a compound of heterogeneous subclasses. This is a "$1-x$" class problem.
Bias Discriminant Analysis is a modification of Linear Discriminant Analysis proposed by Zhou et al. in [15] to address $1-x$ class problems. Call the racket hit class the *target class*. The between-scatter matrix S_b and within-scatter matrix S_w are defined as:

$$S_w = \sum_{\mathscr{R}} (a^k - m_r)(a^k - m_r)^T \quad , \tag{9}$$

$$S_b = \sum_{\overline{\mathscr{R}}} (a^k - m_r)(a^k - m_r)^T \quad , \tag{10}$$

where a^k is the vector of CRCs of the kth event[1], $\mathscr{R}$ is the set of feature vectors belonging to the target class, $\overline{\mathscr{R}}$ is the set of all the other detected events and m_r is the mean vector of the CRCs of the target class.
Denote by Λ the diagonal matrix of generalised eigenvalues of the scatter matrices and V the matrix of the corresponding generalised eigenvectors defined by:

$$S_b V = S_w V \Lambda \quad . \tag{11}$$

[1]For the sake of simplicity of the notation the order p of the polynomial regression is dropped here.

The generalised eigenvalues are solutions of the optimization problem:

$$\Lambda = \arg\max_L \frac{|L^T S_b L|}{|L^T S_w L|} \quad . \tag{12}$$

The Biased Discriminant Transform (BDT) is defined by the operator:

$$W = V\Lambda^{1/2} \quad , \tag{13}$$

and the transformed feature vector $\tilde{a}^k$ of the feature vector a^k is:

$$\tilde{a}^k = V\Lambda^{1/2} a^k \quad . \tag{14}$$

Figure 2 shows an example of BDT performed on a 2-dimension feature space. One can see that the BDT tends to cluster the target class while keeping away elements of the other classes. In the transformed feature space the target class can relevantly be modelled by a single mode probability density function for further probabilistic or statistical processing.

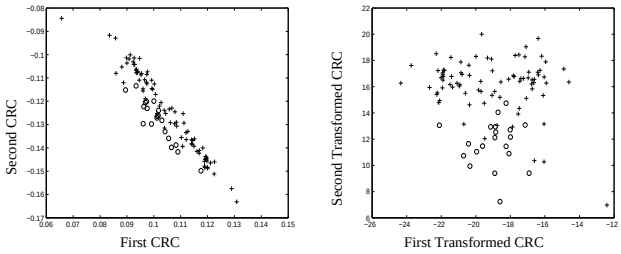

Figure 2: Bias Discriminant Analysis. Target class items are labeled with circles (○), other items are labeled with crosses (+). Original order one CRC space (left) and transformed CRC space using the BDT computed from the target class (right).

3.2 ROC curves

Figure 3 shows the Receiver Operating Characteristic (ROC) curves of the classifier trained with the small training set for four feature spaces of 4, 8, 12, and 16 dimensions. The training set is classified in order to provide an upper bound on the performance of the classifier.
The performance of the classifier applied to the training set increases with the dimension of the feature space for both sets of features. Best performance is reached by the 16-dimension CRCs with a small size training set shown on Figure 3 where the probability of detection $P_d = 1$ with a probability of false alarm $P_{fa} = 0$. A higher dimensional feature space provides a better description of the data set. However the performance of the classifier applied to the test set decreases as the dimension increases.
These simultaneous behaviors show that the classifier is prone to over-fitting at high dimensions. The dramatic decrease of the performance whatever the features also show that a high dimensional feature space is too sparse in regard to the size of the training set. This is an illustration of the curse of dimensionality.
Figure 4 shows the ROC curves of the classifier trained with the large training set.
The performance of the classifier computed with the CRCs are stable while the dimension of the feature space increases

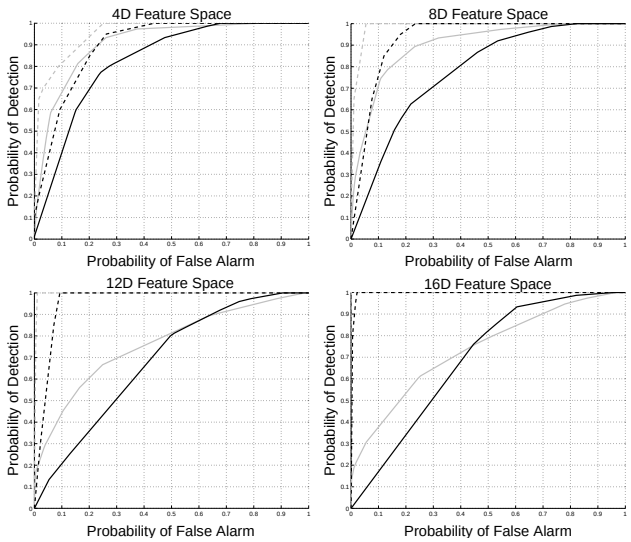

Figure 3: ROC curves of the Bias Discriminant Analysis trained with the small size training set for feature spaces of various dimensions. Feature vectors are CRC (grey line), and first cepstral coefficients (black line). The training set (dashed line) and the test set (plain line) are classified.

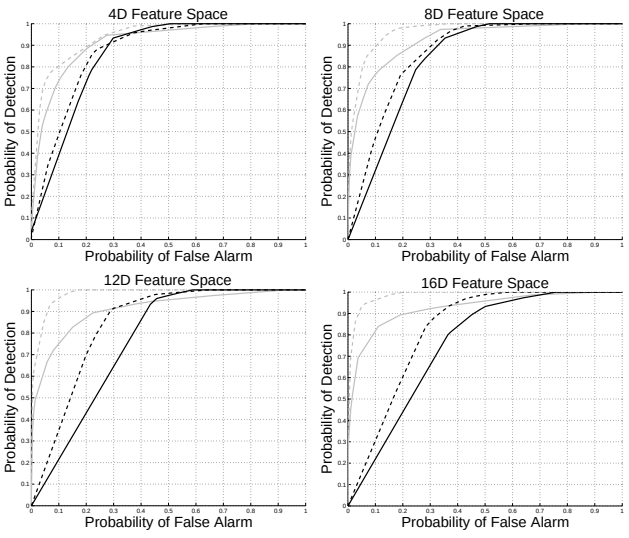

Figure 4: ROC curves of the Bias Discriminant Analysis trained with the large size training set. See Figure 3 for description.

although the performances of the classifier computed with the FCCs decreases as the dimension of the feature space increases. Note also that apart from the 4-dimension case where performances are similar irrespective of the size of the training set, the classifier performs better on the large training set than on the small training set. This highlights again an overfitting phenomenon since the large training set contains more diversity than the small training set.
The somewhat bad performances of the FCCs show that these features are less robust to noise and cepstral estimation error than the CRCs. It also shows that discriminating the impulse responses from other waveforms extracted from the same audio stream cannot be efficiently performed without taking into account information encoded in the high quefrency coefficients.
The CRCs offer better classification performances than the FCCs irrespective of the dimension of the feature space and the size of the training set. Moreover, the discrepancy between the performances of the classifier applied to the training set and to the test set for both sets of features shows that the CRCs are subject to less overfitting than the FCCs. A classifier based on CRCs is less dependent on the training set. Recursion (8) makes CRCs a good candidate for adaptive and sequential classification. The dimension of the CRC space can be updated to match the size of the ever increasing training set with no loss of classification performance and a reduced overfitting.

3.3 Retrieval rank

The analysis of the ROC curves shows that the classification algorithm is subject to false positives. It can be of interest to investigate which events belonging to the negative class produce false positives. We propose here to evaluate the retrieval performance of the classifier by computing a measure of ranking of the likelihood of the events using the Average Normalised Modified Ranking Retrieval introduced by Manjunath et al. [9] in the context of MPEG7.
The ANMRR for the top T from a set $\mathscr{R}$ of N_r positives as proposed in [10] is:

$$\tilde{r} = \frac{1}{NN_r}\left[s(r) - \frac{N_r(N_r+1)}{2}\right], \tag{15}$$

where $s(r) = \sum_{k=1}^{N_r} r_k$, r_k is the rank of the kth positive instance and N is the size of the dataset. It is zero for a perfect ranking and maximum when the N_r positive items occupy the ranks $N-N_r+1$ to N. It can be shown that the expected value and variance of the ANMRR for a random uniform ranking r_k are:

$$\mathrm{E}\{\tilde{r}\} = \frac{(N-N_r)}{2N}, \tag{16}$$

$$\mathrm{Var}\{\tilde{r}\} = \frac{(T+1)^2(N-N_r)}{4N^2N_r}. \tag{17}$$

The statistics of the ANMRR are functions of the size of the dataset N, the number of positives N_r and the total number of retrieved events T. This is not convenient for comparing retrieval results in various configurations. Manjunath et al. [9] in the original formulation proposed to assign a fixed value to the rank of positive items higher than a given threshold. This approach allows for comparison but then statistical properties of a random ranking are not easy to derive. However the expected value of the uniform rank is a measure of comparison much like the coin toss line of ROC curves. Above this value the performances of a retrieval system are statistically worst than performing a random retrieval. We propose here to modify the ANMRR $\tilde{r}$ in order to overcome these drawbacks. Consider the normalised ANMRR:

$$\hat{r} = \frac{1}{s_{\max} - s_{\min}}\left[s(r) - s_{\min}\right], \tag{18}$$

Dim.	4		8		12		16	
racket	.11	.09	.13	.09	.25	.09	.29	.09
bounces	.36	.40	.40	.36	.45	.34	.42	.33
echoes	.30	.34	.35	.31	.30	.36	.34	.35
voice	.57	.63	.66	.68	.63	.71	.65	.75
shuffle	.60	.56	.57	.59	.55	.56	.54	.54
crowd	.60	.56	.57	.57	.62	.63	.56	.66
unknown	.61	.61	.55	.54	.50	.54	.48	.51

Table 1: Average Normalised Modified Retrieval Ranking (ANMRR) for the detected events. For each dimension the left column displays the ANMRR for the small size training set and the right column displays the ANMRR for the large size training set. Three top rows are impulse response type of events.

where $s_{\min} = \sum_{k=1}^{N_r} k$ and $s_{\max} = \sum_{k=N-N_r+1}^{N} k$. The expected value and variance of this new ANMRR are:

$$\mathrm{E}\{\hat{r}\} = \frac{1}{2} \quad \text{and} \quad \mathrm{Var}\{\hat{r}\} = \frac{(T+1)^2}{4N_r(N-N_r)}. \qquad (19)$$

The expected value of the proposed ANMRR $\hat{r}$ does not depend on the configuration of the retrieval experiment. Moreover $\hat{r}_{\min} = 0$ and $\hat{r}_{\max} = 1$ so the boundaries of the proposed ANMRR do not depend on the configuration of the experiment neither.

Table 1 shows the ANMRR of each class of events for both experiments performed with the CRCs. The ANMRR of the racket hits slowly increases when the classifier is trained with the small sample size training set and is stable when the classifier is trained with the large sample size training set. This tendency was already shown with the ROC curves in the previous section. For all the other events the behavior of the ANMRR is rather erratic. This can be explained by the high variability of this measure due to the sensitivity to a change in the rank of one item.

However for racket hits, bounces and echoes, the ANMRR is lower than 0.5 whatever the experiment although the ANMRR of the other events is higher than 0.5. This illustrates that false positives are mainly due to bounces and echoes. As already pointed out bounces and echoes can be considered as impulse responses. This confirms that CRCs are valid features for discriminating impulse responses.

4. CONCLUSION

We have introduced a set of features extracted by cepstral regression for the classification of impulse responses. Experiments have shown that for a small number of coefficients, the proposed Cepstral Regression Coefficients (CRC) outperform the standard First Cepstral Coefficients (FCC). Moreover we have shown that CRC extraction can be implement recursively. If data are processed sequentially, the growing of the population of the dataset requires the simultaneous increasing of the dimension of the feature space. The proposed recursive computation of the CRCs can thus be conveniently integrated into a sequential and adaptive learning scheme.

The discriminating power of the CRCs has been assessed from the interpretation of ROC curves. Such a representation fails to provide information about the non-stationarity of the data even though this is a crucial feature in an adaptive framework. A statistically motivated modification of the Average Normalized Modified Ranking Retrieval (ANMRR) measure has been proposed to evaluate the performance of the proposed features on a sliding short-time interval. This is the first step towards the development of a method for assessing non-stationarity in pattern recognition problems.

REFERENCES

[1] M. Basseville and I. Nikiforov. *Detection of abrupt changes. Theory and applications*. Prentice Hall, 1993.

[2] C.-C. Cheng and C.-T. Hsu. Fusion of Audio and Motion Information on HMM-Based Highlights Extraction for Baseball Games. *IEEE trans. on multimedia*, 8(3):585–599, June 2006.

[3] R. Dahyot, A. Kokaram, A. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *Proceedings of IEEE ICASSP'03*, pages 561–564, 2003.

[4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. MacMillan, 1993.

[5] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE trans. on audio speech and language processing*, 14(1):68–80, January 2006.

[6] C. Hory, A. Kokaram, and W. J. Christmas. Threshold learning from samples drawn from the null hypothesis for the GLR CUSUM test. In *Proc. IEEE MLSP*, pages 111–116, 2005.

[7] W. Hua, M. Han, and Y. Gong. Baseball scene classification using multimedia features. In *Proceedings of IEEE ICME'02*, pages 821–824, 2002.

[8] J. Huang, Z. Liu, and Y. Wang. Joint scene classification and segmentation based on hidden markov model. *IEEE trans. on multimedia*, 7(3):538–550, June 2005.

[9] B. S. Manjunath, J. Rainer Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE trans. on circuits and systems for video technology*, 11(6):703–715, June 2001.

[10] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[11] S. J. Raudys and A. K. Jain. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE trans. on PAMI*, 13(3):252–264, 1991.

[12] L. L. Scharf. *Statistical Signal Processing*. Addison Wesley, 1991.

[13] A. Wald. *Sequential Analysis*. Wiley and Sons, New-York, 1947.

[14] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia Content Analysis Using Both Audio and Visual Clues. *IEEE Signal Processing Magazine*, 17(6):12 – 36, November 2000.

[15] X. S. Zhou and T. S. Huang. Small Sample Learning during Multimedia Retrieval using BiasMap. In *Proceedings of IEEE CVPR'01*, pages 11–17, December 2001.