Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.1, 2015

Audio-Video Detection and Fusion of Broad Casting Information

K. SUBASHINI [1] Dr. S. PALANIVEL[2] Dr. V. RAMALIGAM[3]
1.Research Scholar, Dept of Comp Sci and Engg., Annamalai University, India
Email:subashinikrishnaswamy@gmail.com
2.Professor, Dept of Comp Sci and Engg., Annamalai University,India
Email: spal_yughu@yahoo.com
3.Professor, Dept of Comp Sci and Engg., Annamalai University,India
Email: aucsevr@yahoo.com

Abstract
In the last few decade of multimedia information systems, audio-video data has become an glowing part in many digital computer applications. Audio-video classification has been becoming a focus in the research of audio-video processing and pattern recognition. Automatic audio-video classification is very useful to audio-video indexing, content-based audio-video retrieval and on-line audio-video distribution such as online audio-video shopping, but it is a challenge to extract the most similar and salient themes from huge data of audio-video. In this paper, we propose effective algorithms to automatically segmentation and classify audio-video clips into one of Six classes: advertisement, cartoon, songs, serial, movie and news. For these categories a number of acoustic and visual features that include Mel Frequency Cepstral Coefficients, Color Histogram are extracted to characterize the audio and video data. The autoassociative neural network model (AANN) is used to capture the distribution of the acoustic and visual feature vectors. The AANN model captures the distribution of the acoustic and visual features of a class, and the back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.
Keywords: - Audio and Video detection, Audio and Video fusion, Mel Frequency Cepstral Coefficient, Color Histogram, Autoassociative Neural Network Model(AANN)

1 Introduction

In this modern digital information technology is moving in the form of audio, video, text and audiovisual. Real time shopping as well as commercial broadcasters are enabled with devices to easily broadcast and store multimedia contents. Manual handling of this data is impractical for real time campaigning applications because of its increasingly huge visual information system.

Hence, it is important to have a method of automatically index multimedia data for targeting and commercial application based on multimedia contents. Detection and fusion of data into different categories is one important step for building such systems. Our main objective in this paper is to combine the results of audio-video for detection and fusion.

2 Related Work

Last few decades, there have been many studies on automatic audio and video classification and segmentation using several features and techniques. In [1], a generic audio classification approach for multimedia indexing and retrivel method is described. Unsupervised speaker segmentation with residual phase and MFCC features is given in [2]. The method described in [3] uses content-based audio classification and segmentation by using support vector machines. The work in [4] speech/music segmentation using entropy and dynamism features in a HMM classification framework. The technique described in [5] developed a reference platform for generic audio classification. In [6] audio classification system is proposed using SVM and RBFNN. The perceptual approach is used for automatic music genre classification based on spectral and cepstral features in [7]. A hierarchy based approach for video classification using a tree-based RBF network is described in [8]. In [9] a method is proposed for video classification using normalized information distance. Visual database can be preceptual and categorized into different genres in [10].

The technique described in [11] uses combining multiple evidences for video classification. In [12] the authors address the problem of video genres classification for the five classes with a set of visual features, and SVM is used for classification. Huge literature reports can be obtained for automatic video classification in [13]. Several audio-visual features have been described in [14] for characterzing semantic content in multimedia. The edge based feature, namely, the percentage of edge pixels, is extracted from each key frame for classifying a given sports video into one of the five categories, namely, badmiton, soccer, basket ball, tennis and figure skating techniques in [15]. A feature, called motion texture, is derived from motion field between video frames, either in optical flow field or in motion vector field in [16]. In [17] GMM is used to model low level audio/video feature for the classification of five different categories namely, sports, cartoon, news, commercial, and music. An average correct classification rate of 86.5% is achieved with one hour of records per genre, consisting of

continuous sequences of five minutes each and 40 second decision window. Combining the evidence obtained from several complementary classifiers can improve performance based on the literature shown in [18] and [19]. In [20] a survey of audio based music classification and annotation is described. Then, in [21] a survey on visual content based video indexing and retrieval shows huge information on video. A effective algorithm for unsupervised speaker segmentation using AANN is described in [2]. In [22] a robust speaker change detection algorithm is proposed. Evaluation of classification techniques for audio indexing is described in [23]. In [24] a hybrid approach is presented for audio segmentation. Acoustic, strategies for automatic segmentation are described in [25].

3 Outline of the Work

In this paper, evidence from audio and video are combined for detection and fussion. The paper is organized as follows: Acoustic and visual feature extractions are described in Section 4. Modeling techniques used for detection and fussion are described in Section 5. Experimental results are reported in Section 6. Finally, conclusions is given in Section 7.

4 Feature Extraction

4.1 Acoustic Feature Extraction

MFCC is perceptually motivated representation defined as the cepstrum of a windowed short-time signal. A non-linear mel-frequency scale is used which approximates the behavior of the auditory system. The MFCC is based on the extraction of the signal energy with-in critical frequency bands by means of a series of triangular filters. Whose centre frequencies are spaced according to melscale. The mel-cepstrum exploits auditory principles as well as the decorrelating property of the cepstrum [2]. Fig. 2. illustrates the computation ofMFCC features for a Segment of audio signal which is described as follows: The mel-frequency cepstrum has proven to be highly effective in recognizing structure of music signals and in modeling the subjective pitch and frequency content of audio signals.

Psychophysical studies have found the phenomena of the mel pitch scale and the critical band, and the frequency scale-warping to the mel scale has led to the cepstrum domain representation. The mel scale is defined as $F_{mel} = \frac{c \log\left(1+\frac{f}{c}\right)}{\log(2)}$

Where F_{mel} is the logarithmic scale of f normal frequency scale. The mel- cepstral features, can be illustrated by the MFCCs, which are computed from the fast Fourier transform (FFT) power coefficients. The power coefficients are filtered by a triangular band pass filter bank. When c in (1) is in the range of 250 - 350, the number of triangular filters that fall in the frequency range 200 - 1200 Hz (i.e.,the frequency range of dominant audio information) is higher than the other values of c. Therefore, it is efficient to set the value of c in that range for calculating MFCCs. Denoting the output of the filter bank by Sk (k = 1, 2, · · · ,K), the MFCCs are calculated as

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\frac{\pi}{K}], \; n = 1,2.....,L$$

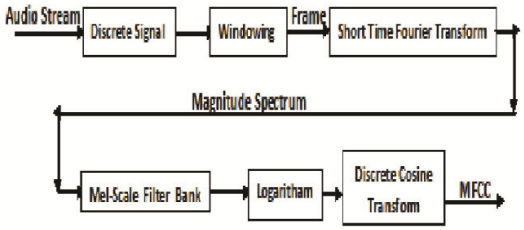

Fig. 2. Extraction of MFCC from audio signal

MFCCs are short-term spectral features as described above and are widely used in the area of audio and speech processing. To obtain MFCCs [2], the audio signals were segmented and windowed into short frames of n samples. Magnitude spectrum was computed for each of these frames using fast Fourier transform (FFT) and converted into a set of mel scale filter bank outputs.

Logarithm was applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. For each audio signal we arrived at 39 features. This number, 39, is computed from the length of the parameterized static vector 13, plus the delta coefficients 13, plus the acceleration coefficients

4.2 Visual Feature Extraction

A color histogram is a representation about distribution of colors in an image, derived by counting the number of pixels in each of the given set of color ranges in a typically two dimensional (2D) color space. A histogram of an image is produced first by dicretization of the colors in the image into a number of bins, and counting the number of image pixels in each bin. The histogram provides a compact summarization of the distribution of data in an image. The color histogram of an image is relatively invariant with translation and rotation about the viewing axis, and may vary very slowly with the view angle. Further, they are computationally trivial to compute. Hence, they are used to compare images in many applications. This work uses color histogram as visual feature. The RGB color space is quantized into 64 bins.

5.Modelling Techniques

5.1 Auto associate Neural Network

Auto associative neural network models are feed forward neural networks performing an identity mapping of the input space, and are used to capture the distribution of input data. The distribution capturing ability of the AANN model is described in this section.

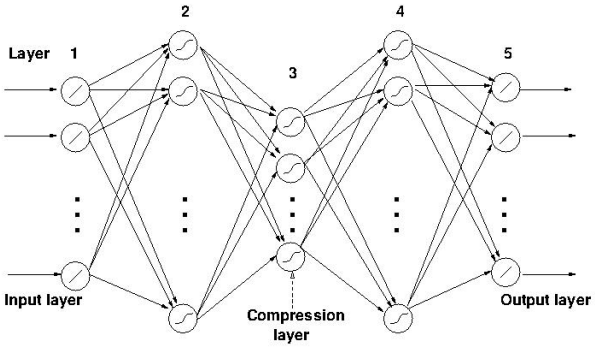

Fig 2(a) A Five Layer AANN model

Let us consider the five layer AANN model shown in Fig.2(a).,which has three hidden layers. The processing units in the first and third hidden layers are non-linear, and the units in the second compression/hidden layer can be linear or non-linear.

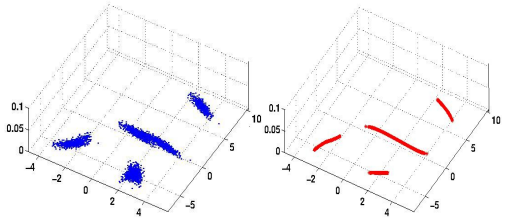

Fig 3(c) Probability Surface. Fig 3(b) Two dimensional output

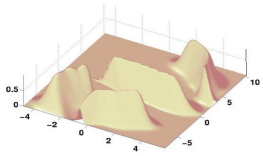

Fig 3(a) Artificial two dimensional data

Fig. 3. Distribution capturing ability of AANN model

As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hyper surface obtained by the projection onto the lower dimensional space. Fig. 3(b) shows the space spanned by the one-dimensional compression layer for the two-dimensional compression layer for the two dimensional data shown in Fig.3(a) for the network structure 2L 10N 1N 10N 2L, where L denotes a linear unit and N denotes a nonlinear unit. The inter value indicates the number of units used in that layer. The non-linear output function for each unit is tanh(s), where s is the activation value of the unit. The network is trained using back propagation learning algorithm. The solid lines shown in Fig. 3(b) indicate mapping of the given input points due to the one-dimensional compression layer. Thus, one can say that the

AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network.

In order to visualize the distribution better, one can plot the error for each input data point in the form of same probality surface as shown in Fig. 3 (c). The error E_j for the data point i in input space is plotted as $p_j = \exp^{(-E_i/\alpha)}$, where ∞ is a constant. Note that p_i is not strictly a probability density function, but we call the resulting surface as probability density function. The plot of the probability surface shown a large amplitude for smaller error E_i, indicating better match of the network for that data point. The constraints imposed by the network can be seen by the shape the error surface taken in both the cases. Once can use the probability surface to study the characteristics of the distribution of the input data capture by the structure of the network, Ideally, one would like to achieve the best probability surface, best defined in terms of same measure corresponding to a low average error.

The five auto associative neural network models as described in section 4 is used to capture the distribution of the acoustic and visual feature vectors. The structure of AANN model used in our study is 39L 38N 4N 38N 39L for MFCC, 64L 32N 8N 32N 64L for color histogram, for capturing the distribution of the acoustic and visual features of a class, where l denotes a linear units, and N denotes nonlinear unite. The nonlinear units use is tanh(s) as the activation, where s is the activation value of the unit. The network is trained using back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

6. Experimental Solution

For conducting experiments, audio and video data are recorded using a TV tuner card from various television regional language channel(sun tv, jaya tv, raj tv, kalinar tv, sutitv tv, K tv, vijay tv and cartoon network) at different timings to ensure quality and quantity of data stream. The training data test includes 5-min of audio stream for each genres, 5-min of video stream for each genres. Audio stream is recorded at 8 KHz with mono channel and 16 bits per sample. Video clips are recorded with a frame resolution of 320×240 pixels and frame rate of 25 frames per second. Training data is segmented into fixed overlapping frames (in our experiments we used 160 ms frames with 80ms overlapping). The sample features extraction process for repeated for audio and video data of varying durations.

6.1 Audio and Video detection

The proposed AANN based audio and video segmentation is shown in Fig. 4. The AANN model is trained to capture the distribution of the feature vectors in the left half of the window, and the feature vectors in the right half of the window are used for testing as shown in Fig. 4. The output of the model(O) is compared with the input to compute the normalized squared error (ek) for the test feature vector X is given by

$$e_k = \frac{\|x - o\|^2}{\|x\|^2}$$

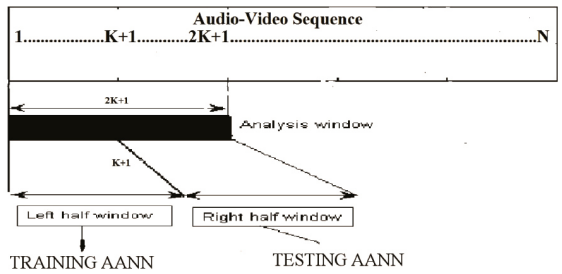

Fig.4 AANN based segmentation algorithm

where o is the output vector given by the model. The error e_k is transformed into a confidence score s using

$S = exp\ (-e_k)$

Average confidence score is obtained for the right half of the window. A low confidence score indicates that the characteristics of the audio (video) signal in the right half of the window are different from the signal in the left half of the window, and hence, the middle of the window is a category change point. The above process is repeated by moving the window with a shift of the about 80 msec until it reaches the right end of the audio

(video) signal. The evidence from audio and video segmentation is combined using weighted sum rule.

6.2 Audio and Video Fusion

Performance of the proposed audio-video classification system is evaluated using the Television broadcast audio database collected from various channels and various genres consists of the following contents: 100 clips of advertisement in different languages, 100 clips of cartoon in different languages, 100 clips on news, 100 clips of movie from different languages, and 100 clips of songs. Audio samples are of different length, ranging from one min to five min, with a sampling rate of 8 kHz, 16-bits per sample, monophonic and 128 kbps audio bit rate. The waveform audio format is converted into raw values (conversion from binary into ASCII). Silence segments are removed from the audio sequence for further processing 39 MFCC coefficients are extracted for each audio clip as described in Section 4.1. The training data is segmented into fixed-length and overlapping frames (in our experiments we used 20 ms frames with 10 ms frame shift.)

The distribution of 39 dimensional MFCC feature vectors in the feature space and 64 dimensional feature vectors is capture the dimension of feature vectors of each class. The acoustic feature vectors are given as input to the AANN model and the network is trained of 2000 epochs. One epoch of training is a single presentation of all training vector. The training takes about 2 mints on a pc with second generation dual core 2.2 GHz CPU. For evaluating the performance of the system, the feature vector is given as input to each of the model. The output of the input to compute the normalized squared error. The normalized squared error E for the feature vector y is given by $E = \frac{\|y - 0\|^2}{\|y\|^2}$, where 0 is the output vector given by the model. The error E is transformed into a confidence score c using c=exp (-E). Similarly the experiments are conducted using histogram as features in video classification. In our experiment video streams has been recorded and digitized at a resolution of 320×240 pixels and frame rate of 25 frames per second. From television broadcast video data base 5 min video genres stream are taken for training and testing. Each frame will have 64 dimensional vectors in which top 16 values are used as features. Compared to audio classification, video classification is more complicated. The memory used for video classification is twice that used for audio classification.

6.3 Combining Audio and Video Fusion using AANN

In this work, combining the modalities has been done at the score level. The methods to combine the two levels of information present in the audio signal and video signal have been proposed. The audio based scores and video based scores are combined for obtaining audio-video based scores as given equation (9). It is shown experimentally that the combined system outperforms the individual system, indicating complementary nature. The weight for each modality is decided empirically. $s = \frac{w}{n}\sum_{i=1}^{n} S_i^a + \frac{(1-w)}{p}\sum_{i=1}^{p} S_i^v$

$$1 \le j \le c \qquad (9)$$

Where

$$a_j = \sum_{i=1}^{n} x_i^j \qquad 1 \le j \le c$$

$$v_j = \sum_{i=1}^{p} y_i^j \qquad 1 \le j \le c$$

$$x_i^j = \begin{cases} 1, & \text{if } c_i^a = j \\ 0, & \text{otherwise} \end{cases}$$

$$1 \le i \le n \ , \quad 1 \le j \le c$$

$$y_i^j = \begin{cases} 1, & \text{if } c_i^v = j \\ 0, & \text{otherwise} \end{cases}$$

$$1 \le i \le p \ , \quad 1 \le j \le c$$

s -is the combinined audio and video confidence score.
s_i^a -is the Confidence score rate of the i^{th} audio frame.
s_i^v -is the Confidence score rate of the i^{th} video frame.

v_j - Video based score for j^{th} frame.
a_j -Audio based score for j^{th} frame.
m_j - Audio-video based score for j^{th} frame.
c -number of classes.
n -number of audio frames.
p -number of video frames.
w -weight.

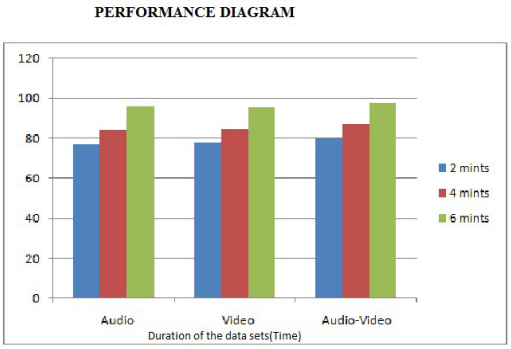

Fig.5 AANN Based Fusion Performance

The category is decided based on the highest confidence score various from 0 to 1. Audio and video frames are combined based on 4:1 ration of frame shifts. The weight for each of modality is decided by the parameter w is chosen such that the system gives optimal performance for audio-video based classification. Obtained experimental results 2 mints , 4 mints and 6 mints data sets are charted on performance diagram Fig. 5. Similarly the 6m ints data sets result is given in the form of confusion matrix in Table1.

Table 1
Confusion matrix for audio-video Fusion using AANN (in %)

Classifier	Modality	Adv	Cartoon	Songs	Serial	Movie	News	
Adv.	Audio	96.78	0.38	0.76	0.91	0.68	0.59	
	Video	95.17	0.98	1.26	0.74	0.99	0.86	
	Audio-video	97.48	0.38	0.42	0.39	0.62	0.71	
Cartoon	Audio	0.97	94.38	1.52	1.41	0.88	0.84	
	Video	0.96	95.32	0.74	1.31	0.69	0.98	
	Audio-video	0.32	98.12	0.48	0.26	0.35	0.47	
Song	Audio	0.73	0.89	94.71	1.42	0.98	1.17	
	Video	0.98	0.79	95.71	0.76	0.93	0.91	
	Audio-video	0.61	0.45	97.16	0.69	0.58	0.51	
Serial	Audio	1.07	0.99	0.86	95.38	0.76	0.94	
	Video	0.82	0.94	0.91	95.43	1.04	0.89	
	Audio-video	0.29	0.17	0.68	97.21	0.91	0.74	
	Audio	1.13	0.49	0.56	0.94	96.01	0.87	1.44
Movie	Video	0.93	0.3	0.96	0.88	95.98	0.95	1.47
	Audio-video	0.68	0.55	0.87	0.38	96.98	0.44	0.97
News	Audio	0.47	0.62	0.61	0.51	0.67	97.12	
	Video	0.97	0.90	0.44	0.86	0.79	96.04	
	Audio-video	0.27	0.32	0.43	0.13	0.84	98.01	

4 Conclusion

This paper proposed an automatic audio-video based classification using AANN. Mel frequency cepstral coefficients are used as features to characterize audio content. Color Histogram coefficients are used as features

to characterize the video content. A non linear support vector machine learning algorithm is applied to obtain the optimal class boundary between the various classes namely advertisement, cartoon, songs, serial, movie and news by learning from training data. Experimental results show that proposed audio-video classification gives an accuracy of 97.49%., using AANN.

References:

[1] S. Kiranyaz, A. F. Qureshi, M. Gabbouj, *A generic audio classification and segmentation approach for multimedia indexing and retrieval,* IEEE Trans. Audio, Speech and Lang Processing 14(3)(2006) 1062–1081.

[2] S.Jothilaskmi, S. Palanivel, V. Ramalingam, *Unsupervised speaker segmentation with residual phase and MFCC features,* Expert System With Applications 36 (2009) 9799–9804.

[3] L. Lu, H.-J. Zhang, S. Z. Li, *Content-based audio classification and segmentation by using support vector machines*,Springer-Verlag Multimedia Systems 8 (2003) 482–492.

[4] J. Ajmera, I. McCowan, H. Bourlard, *Speech/music segmentation using entropy and dynamism features in a HMM classification framework*, Speech Communication40(3) (2003) 351–363.

[5] R. Jarina, M. Paralici, M. Kuba, J. Olajec, A.Lukan,M. Dzurek,*evelopment of reference platform for generic audio classification development of reference plat from for generic audio classification*., IEEE Computer society, Work shop on Image Analysis for Multimedia Interactive, (2008) 239–242.

[6] P.Dhanalakshimi,S.Palanivel,V.Ramaligam, *Classification of audio signals using SVM and RBFNN*, Expert System With Applications 36 (2009) 6069– 6075.

[7] C. Lin, J. Shih, K. Yn, H. Lin, *Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features,* IEEE Transactions Multimedia 11 (4) (2009) 670–682.

[8] W.Gillespie,D. Nguyen, *Video classification using a tree-based RBF network,* IEEE International Conferance on image processing 3 (1) (2005) 465–468.

[9] K. Kaabneh, A. Abdullah, A. Al-Halalemah, *Video classification using normalized information distance.,* In Proceedings of the geometric modeling and imaging-new trends (2005) 34–40.

[10] M. Geetha, S.Palanivel, V.Ramaligam, *A novel block intensity code for video classification and retrieval*, Expert System With Applications 36 (2009) 6415–6420.

[11] V.Suresh, C. K. Mohan, R. Kumaraswamy, B. Yegnanarayana,*Combining multiple evidence for video classification,* In IEEE international conference intelligent sensing and information processing (2005) 187–192.

[12]V.Suresh,C.K.Mohan,R.Kumaraswamy,B. Yegnanarayana,*Content-based video classification using SVM.,* In International conference on neural information processing.

[13] D. Brezeale, D. J.Cook, *Automatic video classification a survey of the literature,* IEEE Transaction on System, Man, and cybernetic 38 (3) (2008) 416–430.

[14] Y. Wang, Z. Liu, J. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Process. Mag. 17 (2000) 12–36.

[15] Y. Yuan, C. Wan, *The application of edge features in automatic sports genre classification.,*In Proceedings of IEEE Conference on Cybernetics and Intelligent Systems (2004) 1133–1136.

[16] Y.-F. Ma, H.-J. Zhang, *Motion pattern based video classification using support vector machines.,*In Proceedings of IEEE International Symposium on Circuit and Systems 2 (2002) 69–72.

[17] L. Q. Xu, Y. Li, *Video classification using spactial-temporal features and PCA*, International Conference on Multimedia and Expo 3 (2003) 345–348.

[18] J. Kittler, M. Hatef, R. Duin, J.Matas, *On combining classifier,* IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[19] L. Xu, A. Krzyzak, C. Suen, *Methods of combining multiple classifiers and their applications to handwriting recognition,* IEEE Trans. Syst. Man, Cybern. 2 (1992) 418–435.

[20] Z. Fu, G. Lu, K. M. Ting, D. Zhang, *A survey of audio-based music classification and annotation,* IEEE Transactions Multimedia 13 (2) (2011) 303– 318.

[21] H. V. Weiming, N. xie, L. Li, X. L. Zeng, S. maybank, *A survey on visual content-based video indexing and retrival*, IEEE Transaction on System, Man, and cybernetic part c (2011) 1–3.

[22] J. Ajmera, I. McCowan, H. Bourland, *Robust speaker change detection*, IEEE Journal of Signal Process Letter 11 (8) (2004) 649–651.

[23] J. A. Arias, J. Pinquier, R. Ande-Obrecht, *Evaluation of classification techniques for audio indexing,* In proc. 13th Eropean conf. Signal Processing.

[24] S. Cheng, H. Wang, *Metric SEQDAC: A hybrid approach for audio segmentation*, Proc. 8th International conference on spoken language Process. (2004) 1617–1620.

[25] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, *Acoustic strategies for automatic segmentation of audiodata,*Proc.IEEE International conference on Acoust, Speech, Signal Process. (2000) 1423–1426.

[26] P. Lin, J. Wang, J. Wang, H. Sung, *Unsupervised speaker change detection using SVM training misclassification rate*, IEEEInt'l Conf. Acoustics, Speech and Signal Processing 14 (3) (2006) 1062–1081.
[27] M. Sieglar, U. Jain, B. Raj, R. Stern, *Automatic segmentation,classification and clustering of broadcast news audio*, Proc. DARPA Speech recognition workshop (1997) 97–99.
[28] V. Vapnik, *Statistical Learning Theory,* John Wiley and Sons, New York, 1998.
[29] S. Palanivel, *Person authentication using speech, face and visual speech*, Ph.D thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg (2004).
[30] B. Yegnanarayana, S. Kishore, *AANN: An alternative to GMM for pattern recognition,* Neural Networks 15.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar