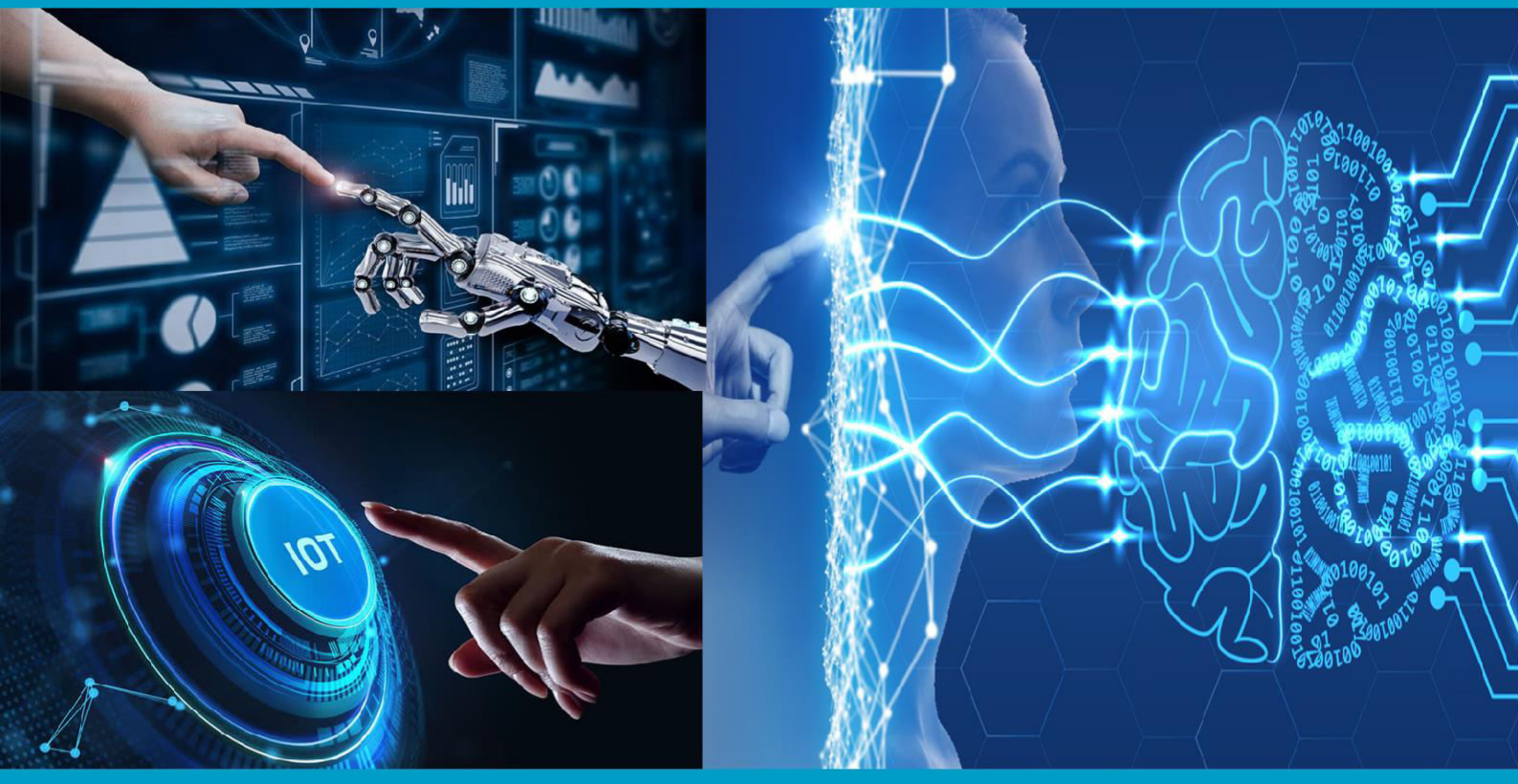
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 9, Issue 7, July 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.580

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

Experimental Study on Finding Audio-Visual Association by Maximizing Mutual Information

Mr.Paras, Mrs. Monica

M.Tech, Department of Electronics and Communication Engineering, S K I T M, Bahadurgarh, India

Assistant Professor, Department of Electronics and Communication of Engineering, S K I T M, Bahadurgarh, Haryana, India

ABSTRACT: In this thesis, one camera and one microphone are used to combine the audio and visual signals. Although there are several methods for combining audio and video, the issue is still unresolved.

Here, locating and identifying the speaker in a video is done by using as much of the shared information from the audiovisual context as possible. A signal-level fusion methodology is used in the experiment to find out whether audio and visual signals are coming from the same source. The probabilistic multi-modal generation statistical implicit model described in [1] is utilised to calculate the information theoretic measure of audio-visual correspondence in this context. By adhering to particular constraints for which entropy is maximised, non-parametric statistical density is able to provide superior results in the approach being used to characterise the mutual information between signals from distinct domains. We can identify who is uttering a certain speech and determine whether the audio is connected to the visual being observed by maximising the mutual information between the various pairs of signals.

No inferences about the user's looks or voice are formed in the process. Additionally, this approach doesn't call for corporate training. Utilizing the CUAVE database, the experimental findings in this thesis are presented.

I. INTRODUCTION

Background

The way we think about the world is influenced by many aspects of nature. In order to promote appropriate perception, humans constantly blend information from various sensory modalities. For example, we can estimate the taste of food based on its appearance and scent without actually tasting it. Any listener can take advantage of the relationship between produced sound and lip movement.

When a human sees conflicting audio and visual cues, the received sound may not exist in either modality, according to the McGruk effect. This effect serves as the foundation for modelling audio and visual speech in the field of audio-visual signal processing. As a result, based on the cues from audio and visual signals for perception, we can claim that speech is fundamentally bimodal. The inclusion of audio and visual signals to the speaker recognition process adds another modality.

Researchers have attempted to merge knowledge from several scientific domains as a result of this observation. Reservations, ticket booking, traffic information, radio, FM, and database access are all areas where audio alone recognition is practical. These conversational speech systems, on the other hand, are designed for a single user and need tethered engagement, which necessitates the use of a telephone headset or a microphone. This limits the performance of a dialogues system, since there must be circumstances where

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

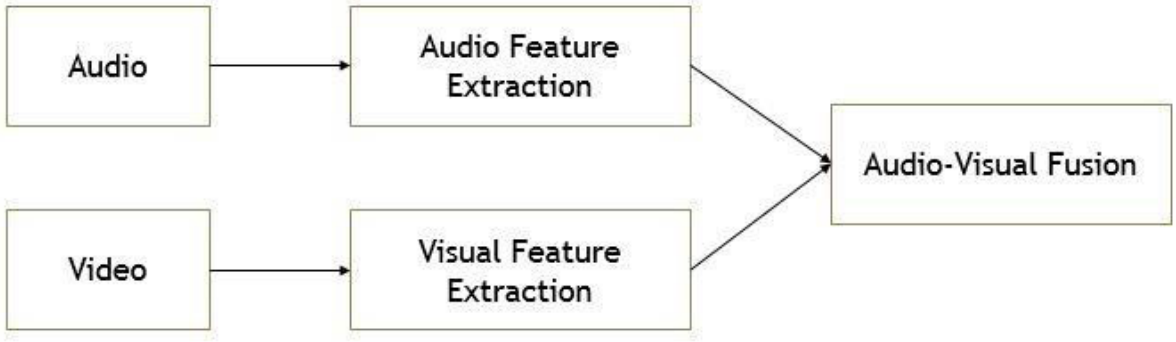


Figure 1.1: Generalized structure of audio-visual fusion

users might expect to freely communicate with a device e.g. perceptual user interfaces. Where user wish to directly command the system. So we need to localize the speaker, who can give command to the system, and it can verify that both modality belongs to the same event.

II. LITERATURE REVIEW

Hershey and Movell [2] have finished the basic work for fusing audio and video using mutual information. They calculated the Pearson correlation coefficient between pixel intensity and average acoustic energy to show connection between audio and visual. First, they assumed the densities were Gaussian, but later they integrated the correlation coefficient with mutual information, including dividing the mutual information between each pixel and acoustic energy audio/video temporal alignment. Their approach may be used to determine whether two signals are from the same individual.

To assess the consistency of speech and facial expression, Nock [5] evaluated two mutual information and one HMM-based technique. The majority of them first use the discrete cosine transform of optical flow to capture the characteristics of video, then utilise the cepstral coefficient for voice signal and optical flow. Another approach for audio-visual correspondence has been developed by J. Fisher and T. Darrell [1, 3] employing mutual information and information theoretic learning. They used projection to output subspace to translate high-dimensional audio and visual signals into low-dimensional subspace. Then demonstrate how their system may capture complicated relationships between audio and video using information theoretic learning and a non-parametric density estimator.

Canonical correlation is used by Slaney and Covell [4].

III. AUDIO FEATURE EXTRACTION

Energy of Audio Signal

Hershey and Movellan [2] were the first to utilise the average acoustic energy of an audio signal in a specific audio frame as an audio feature, and they attempted to determine the mutual information between audio energy and pixel intensity. The energyof audio signal is calculated as,

$$\text{Energy} = \frac{1}{N}\sum_{i=1}^{N} x(n)^2 \qquad (2.1)$$

where N is the number of audio samples in a given frame and x(n) is the sample amplitude of the audio stream. To differentiate between voiced and unvoiced speech, these are most often utilised. Additionally suggested in were log energy and the root mean square of audio amplitude [6] [7].

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is used in speech and audio visual speaker identification systems because it represents a person's voice spectrally [5, 8, 4, and 9].

Linear Predictive Coding (LPC)

For a period of 10ms, it is believed that the vocal tract's properties stay constant. The speech that we hear is a combination of input from the source and vocal tract contraction, as described in [10, [11].

$$S(n) = E(n)\, V(n) \tag{2.2}$$

Resonant frequency, which is the frequency at which the vocal tract vibrates, allows the vocal tract to be seen as a filter in the frequency domain.

$$S(Z) = \frac{G}{1 + \sum_{i=1}^{q} a_i z \quad 1} \tag{2.3}$$

Because prior q samples may anticipate the current sample, they are termed linear predictive coefficients.

$$S(n) = \sum_{i=1}^{q} a_i S(n\ \ i) + G\ \ u(n) \tag{2.4}$$

where, G is the gain and S(n) is the corresponding speech samples.

Periodogram

Periodograms were employed by J Fisher and T Darell [1] to parametrize speech. The square magnitude of the FFT bins is referred to as the pe-riodogram. Power spectral density serves as the foundation for periodograms.

$$P_{a;N}(w) = \frac{1}{N}\ jDT\ FT(a_w)j^2$$

$$= {}_N{}^{1} \sum a_w(n) e^{jwn}{}^{2}$$

$$N\ \ 1$$

$$n=0$$

Windowed segments of samples are represented by aw (n), window functions by w (n), and samples are represented by N.

$$a_w(n) = a(n) w(n) \tag{2.6}$$

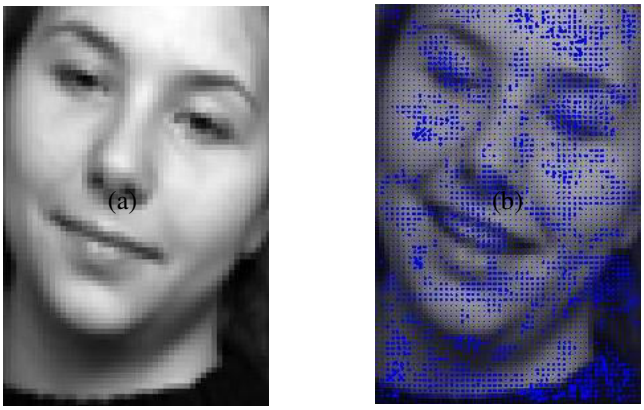
IV. OPTICAL FLOW

Optical Flow, which is brought on by relative motion between consecutive visual frames, may swiftly estimate the apparent mobility between any visual scene's objects, surfaces, and edges. Another way to put it is relative motion between an observer (such an eye or a camera) and the scene. The Horn and Shrunk method described in[12] may be used to calculate the speed of that pixel's movement.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |



An image from a video series is shown in Figure 2.1 (a). A horizontal and vertical arrow displays the relative direction and amplitude of optical flow between two successive frames.

Pre-Whitened image

The cross-correlated word must be eliminated from the image sequence by performing per-whitening in accordance with the entropy maximisation theory. This may be accomplished by multiplying the data by the inverse of the square root of the mean power spectrum in the Fourier domain. Each white image is regarded to be a single sample with a size equal to the number of pixels. Fisher and Darrel used pre-whitened images for information theoretic fusion in [1], [3].

Correspondence Measure of Audio and Video

Pearson Correlation Coefficient

According to [2] Hershey and Movellan, the average acoustic energy of an audio sequence correlates with the acoustic energy of each individual pixel. In the beginning,

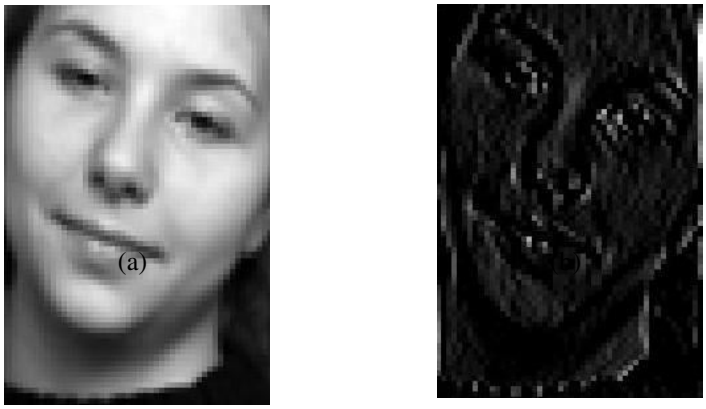


Figure 2.2: (a) frame from a video clip. The lips, eyes, and nose are highlighted in this pre-whitened photograph (b)Initially assuming that the density is Gaussian, they expanded their first assumption and provided a measure for determining the correlation between each pixel and the acoustic energy over a number of successive frames.

$$r = p \frac{\operatorname{cov}(A;V)}{\operatorname{var}(A)\operatorname{var}(V)} \tag{2.7}$$

where r is the pearson correlation coefficient.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

V. EQUIVALENT MARKOV CHAIN MODEL

Figure 3.4 shows the necessary model for which the information theoretic approach will be employed (equivalent Markov model obtained by presence of separation function). The following inequality is true if there is decomposition other than through data processing inequality [17]:

$$\begin{matrix} I(A_Y ; V_Y) & I(A_Y ; Y) \\ I(A_Y ; V_Y) & I(V_Y ; Y) \end{matrix} \tag{3.8}$$

Additionally, this inequality holds for any function of AY and VY (for example, fA = f (AY ; ha), fV = f (VY ; hV)).

$$I(f_A; f_V) \; I(f_A; Y)$$

$$I(f_A; f_V) \; I(f_V ; Y) \tag{3.9}$$

finally, it can be shown using [13]

$$I(f_A; f_V) \; I(A_V ; V_Y) = I(A;V) \tag{3.10}$$

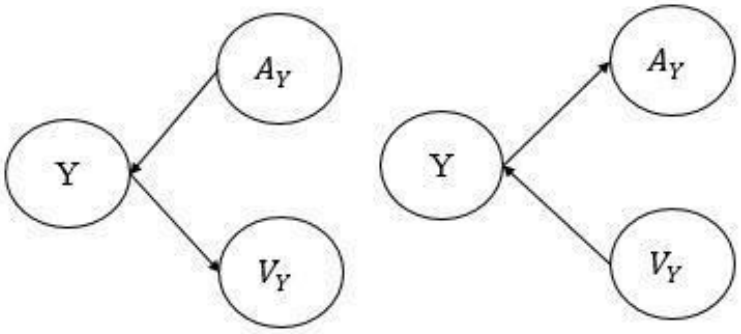


Figure 3.4: similar Markov model was discovered by the presence of a separating function.

It demonstrates that increasing the mutual information between fV and fA would undoubtedly enhance the mutual information between FA and Y, as well as between FV and Y. These functions will also be utilised for non-parametric combined audio and video density estimation in an audio visual sequence.

VI. NONPARAMETRIC PDF ESTIMATION

Due to the fact that mutual information is an integral function of probability, employing it as a criteria for adaptation poses difficulties and calls for a complete grasp of the PDF. Furthermore, density must be inferred from the samples since we did not explicitly offer it. which is difficult to directly measure and for which an assumption must be made about the form of the density function.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

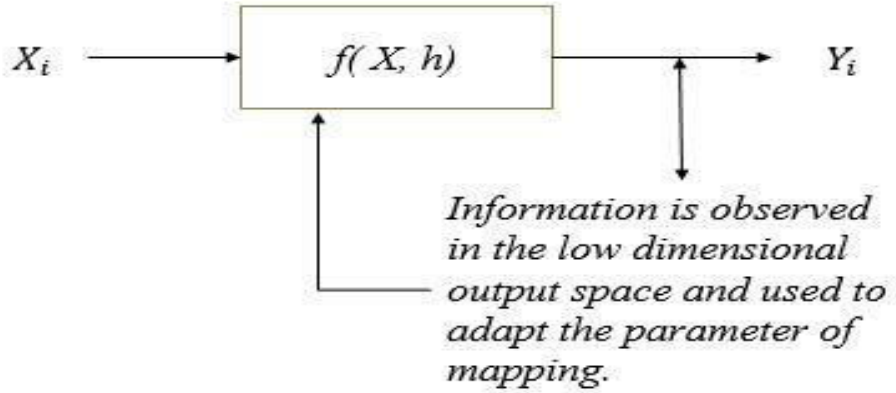
| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

However, non-parametric estimators will perform better if the dimension of the RV can be modified by converting it into a new RV, as illustrated in [1]. These estimators rely on such estimates in the output space, which reduce computational cost.Non parametric kernel based estimator such as Parzen window method are

Figure 4.1: Mapping as feature extraction. Information content is measured in the low dimen-sional space of the observed output.

used for estimating the probability density function which is given as [15].



N_x

$$\hat{f}_X(u) = \frac{1}{N} \sum_{j=1}^{x} k(x_j - u) \qquad (4.6)$$

here, x_i 2 ^{N} are observation of RV, k () is the guassian kernel which satisfies

the property of PDF i.e. k (u) > 0 and $\int_R k(u)du = 1$

This kernel estimator can be viewed as the convolution of kernal function

about the obsession. This method's objective is to estimate the PDF locally, hence the kernel has to be locally unimodal and decaying to zero, and k (u) needs to be differentiable everywhere.

VII. CONCLUSION

This study uses a method for assessing the signal-to-noise ratio of audio and video observations on a movie from the CUAVE database. The aforementioned technique may be used to determine whether or not a separately recorded video and audio fragments are from the same speaker, according to experimental findings. This method produces signal level fusion. The use of any auditory or visual model is not restricted. For lip tracking, the whole video and audio frame is utilised, without any filtering or segmenting, and it is not language-specific. When previous models of specific users are available or in fields where such assumptions are practical, this knowledge may be exploited effectively. This approach works as long as there is continuous audio and video signal, which means there shouldn't be any monologues or significant head or body movement in between. This technique guarantees audio-video coherence and may confirm if the audio and the video refer to the same event or not. The underlying joint features of the modalities being fused are not strongly assumed by this procedure (e.g. Gaus-sian statistics). Here, audio-visual data undergoes modification for a brief period of time (about 2–2.5 sec). This approach just needs continuous audio-video data for a period of time; it doesn't need any special recognition.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

REFERENCES

1. J. W. Fisher and T. Darrell, “Speaker association with signal-level audiovi-sual fusion,” IEEE Transactions on Multimedia,, vol. 6, no. 3, pp. 406–413, 2004.
2. J. Hershey and J. Movellan, “Audio-vision: Using audio-visual synchrony to locate sounds,” in Advances in Neural Information Processing Systems 12, Citeseer, 2000.
3. J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola, “Learning joint statistical models for audio-visual fusion and segregation,” in NIPS, 772–778, 2000.
4. M. Slaney and M. Covell, “Facesync: A linear operator for measur-ing synchronization of video facial images and audio tracks,” in NIPS,814–820, 2000.
5. H. J. Nock, G. Iyengar, and C. Neti, “Assessing face and speech consis-tency for monologue detection in video,” in Proceedings of the tenth ACM international conference on Multimedia, pp. 303–306, ACM, 2002.
6. J. P. Barker and F. Berthommier, “Evidence of correlation between acous-tic and visual features of speech,” Ohala et al, pp. 199–202, 1999.
7. H. Izadinia, I. Saleemi, and M. Shah, “Multimodal analysis for identifi-cation and segmentation of moving-sounding objects,” IEEE Transactions on Multimedia,, vol. 15, no. 2, pp. 378–390, 2013.
8. L. Muda, M. Begam, and I. Elamvazuthi, “Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques,” arXiv preprint arXiv:1003.4083, 2010.
9. H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” the Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.
10. E. Gopi, Digital Speech Processing Using Matlab. Springer, 2014.
11. B. K. Horn and B. G. Schunck, “Determining optical flow,” in 1981 Tech-nical Symposium East, pp. 319–331, International Society for Optics and Photonics, 1981.
12. A. T. Ihler, J. W. Fisher, and A. S. Willsky, “Hypothesis testing over fac-torizations for data association,” in Information Processing in Sensor Net-works, pp. 239–253, Springer, 2003.
13. T. Ihler, J. W. Fisher, and A. S. Willsky, “Nonparametric hypothesis tests for statistical dependency,” IEEE Transactions on Signal Processing,, vol. 52, no. 8, pp. 2234–2249, 2004.
14. E. Parzen, “On estimation of a probability density function and mode,” The annals of mathematical statistics, pp. 1065–1076, 1962.
15. J. W. Fisher III and T. Darrell, “Probabalistic models and informative sub-spaces for audiovisual correspondence,” in Computer VisionECCV 2002, pp. 592–603, Springer, 2002.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580|

| Volume 9, Issue 7, July 2022 |

| DOI: 10.15680/IJMRSETM.2022.0907017 |

16. T. M and J. A. C. Thomas, Elements of information theory. John Wiley & Sons, 2012.
17. J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," Un-supervised adaptive filtering, vol. 1, pp. 265–319, 2000.
18. J. W. Fisher III and J. C. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions," in Aerospace/Defense Sensing and Con-trols, pp. 2–13, International Society for Optics and Photonics, 1996.
19. J. W. Fisher and J. C. Principe, "A methodology for information theo-retic feature extraction," in The 1998 IEEE International Joint Conference on Neural Networks Proceedings, 1998.IEEE World Congress on Computa-tional Intelligence., vol. 3, pp. 1712–1716, IEEE, 1998.
20. R. Linsker, "Self-organization in a perceptual network," Computer, vol. 21, no. 3, pp. 105–117, 1988.
21. J. W. Fisher and J. C. Principe, "Entropy manipulation of arbitrary non-linear mappings," in Proceedings of the 1997 IEEE Workshop Neural Net-works for Signal Processing [1997] VII., pp. 14–23, IEEE, 1997.
22. S. Kullback, Information theory and statistics. Courier Corporation, 1997.
23. M. D. Plumbley, On information theory and unsupervised neural networks. University of Cambridge, Department of Engineering, 1991.
24. R. Linsker, "How to generate ordered maps by maximizing the mutual in-formation between input and output signals," Neural computation, vol. 1, no. 3, pp. 402–411, 1989.
25. B. Logan et al., "Mel frequency cepstral coefficients for music modeling.," in ISMIR, 2000.
26. T. M and J. A. C. Thomas, Elements of information theory. John Wiley & Sons, 2012.
27. J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," Un-supervised adaptive filtering, vol. 1, pp. 265–319, 2000.
28. J. W. Fisher III and J. C. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions," in Aerospace/Defense Sensing and Con-trols, pp. 2–13, International Society for Optics and Photonics, 1996.
29. J. W. Fisher and J. C. Principe, "A methodology for information theo-retic feature extraction," in The 1998 IEEE International Joint Conference on Neural Networks Proceedings, 1998.IEEE World Congress on Computa-tional Intelligence., vol. 3, pp. 1712–1716, IEEE, 1998.
30. R. Linsker, "Self-organization in a perceptual network," Computer, vol. 21, no. 3, pp. 105–117, 1988.
31. J. W. Fisher and J. C. Principe, "Entropy manipulation of arbitrary non-linear mappings," in Proceedings of the 1997 IEEE Workshop Neural Net-works for Signal Processing [1997] VII., pp. 14–23, IEEE, 1997.
32. S. Kullback, Information theory and statistics. Courier Corporation, 1997.
33. M. D. Plumbley, On information theory and unsupervised neural networks. University of Cambridge, Department of Engineering, 1991.
34. R. Linsker, "How to generate ordered maps by maximizing the mutual in-formation between input and output signals," Neural computation, vol. 1, no. 3, pp. 402–411, 1989.
35. B. Logan et al., "Mel frequency cepstral coefficients for music modeling.," in ISMIR, 2000.

Impact Factor: 7.580

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com