INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

COMPARISON RESULTS IN HMM AND SVM FOR AGE AND GENDER RECOGNITION

MS. BHAVANA R. JAWALE, MRS. SWATI PATIL

G.H.R.I.E.M., Jalgaon, Jalgaon.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future people's life[8].

Keywords: HMM, SVM, Age, Gender

PAPER-QR CODE

Corresponding Author: MS. BHAVANA R. JAWALE

Access Online On:

www.ijpret.com

How to Cite This Article:

Bhavana R. Jawale, IJPRET, 2016; Volume 4(9): 1655-1669

INTRODUCTION

Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future peoples life[8].

Speech emotion recognition aims at recognizing the underlying emotional state of the speaker from his or her speech signal [8]. This is mainly motivated by intelligent Human Machine Interaction required for different kinds of applications.

In the field of speech emotion recognition, a number of classification approaches have recognition models can be classified into two types: 1) for supra segmental prosodic features, such as the mean, median, standard deviation, range, or percentile of short time pitch (energy), estimated over the whole utterance, global models such as Gaussian mixture model(GMM), support vector machine(SVM), artificial neutral networks (ANN) and k-NN have been adopted. 2) for frame based dynamic spectral features like Mel Filter bank Cestrum Coefficient (MFCC), dynamical models such as Hidden Markov Model (HMM) are considered [14], [15]. Compared to the global models, dynamic modeling approaches provide a better consideration of the temporal dynamics of emotions[8].

Understanding human emotional states is indispensable for human-human interaction and social contact. Human emotional states affect perception and rational decision making during human-human interactions. Hence, automatic emotion recognition is important in harmonious interactions or communication between computers and human beings. The challenging research field, “affective computing,” introduced by Picard [13] aims at enabling computers to recognize, express, and have emotions.

The emotion recognition focused only on considering single facial or vocal modality. Based on psychological analysis [3], [13] it was found that human emotional states were mainly transferred through multiple channels such as face, voice, body gesture, and speech content. For this reason, exploring data fusion strategies can achieve better recognition performance[11]. Many data fusion approaches have been developed in recent years. Fusion operation can be conducted at the feature level, decision level, and model level for audio-visual emotion recognition [9]. In feature-level fusion [21][11], facial and vocal features are concatenated to construct joint feature

vectors and then modeled by a single classifier for emotion recognition. However, fusion at the feature level will increase the dimensionality and may suffer from the problem of data sparseness. In terms of decision-level fusion [2], multiple signals can be modeled by the corresponding classifier first and then the recognitions from each classifier are fused in the end.

Although fusion at the decision level enables us to interpret the performance of different classifiers and to gain insights into the role of multiple modalities during emotional expression, the assumption of conditional independence does not consider mutual correlation among multiple modalities. Contrary to the decision level, model-level fusion focuses on the mutual correlation among the multiple signal streams, but it is difficult to explore the contributions of multiple modalities during emotional expression.

A peculiar and very important developing area concerns the remote monitoring of elderly or ill people. Indeed, due to the increasing aged population, HCII systems able to help live independently are regarded as useful tools. Despite the sign cant advances aimed at supporting elderly citizens, many issues have to be addressed in order to help aged ill people to live independently. In this context recognizing people emotional state and giving a suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot research area in both industry and academically. There is much research in this area and there have been some successful products [1].

Usually, emotion recognition systems are based on facial or voice features. This is a solution, designed to be employed in a Smart Environment, able to capture the emotional state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smart phones.

Traditional recommender systems then use these profiles, together with meta-data and ratings from other users in the network, to provide personalization. One of the issues however, in the context of broadcast TV, is the lack of an uplink channel, through which information such as ratings can be exchanged with the remaining users. It is therefore highly desirable that feedback from users be collected locally, in the set-top box or smart TV if possible, and as unobtrusively as possible, e.g. such as through unobtrusive relevance feedback [5].

By means of local recommendation and implicit user feedback, these systems can work quite effectively, but it is important to consider the preferences of a group of users as well as a single user. This is a particular issue when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [18]. In the Socially Aware TV Program

Recommender for example [19], groups of users who want simultaneous access to the TV are taken into account, where individual profiles that have a common interest are combined.

HMM's have been very successful in automatic speech recognition, mainly because there is an efficient of fitting an HMM to data: the forward-backward algorithm and the Baum-Welch re-estimation formulas. Despite this success, HMM's have several major limitations as models of sequential data. They represent the recent history of the sequence using a single, discrete K-state multinomial.

The efficiency of the Baum-Welch re-estimation algorithm depends on this fact, but it severely limits the representational power of the model. The hidden state of a single HMM can only convey log2 K bits of information about the recent history. If the generative model had a distributed hidden state representation [15] bits of information, so the information bottleneck scales linearly with the number of variables and only Logarithmically with the number of alternative states of each variable. This suggests that it would be much better to use generative models composed of many small HMM's whose outputs are somehow combined to produce a sequence.

2. LITERATURE REVIEW

The several works have been dedicated to DNNHMMs based large vocabulary continuous speech recognition. However, to knowledge only few works on the application of DNN-HMMs in emotion recognition, have been reported. In [8], a Generalized Discriminant Analysis (GerDA) based on DNNs, is to learn the discriminative features for classifying high or low of arousal and positive or negative valence.

Recently, most researchers have seen increased attention being given to decision level and model-level fusion in data fusion approaches. Accordingly, two popular data fusion approaches at decision and model levels: error weighted classifier combination and the coupled hidden Markov model (C-HMM). The former used an empirical weighting scheme for recognition decision, and the latter modeled the asynchronous (e.g., audio and visual) nature of the multi-stream features for different applications. These models were successfully used in different fields such as emotion recognition, interest detection, human identification, hand gesture recognition, 3-D surface inspection, speech prosody recognition, audio-visual speech recognition, and speech animation.

Visual information has been shown to be useful for improving the accuracy of speech recognition in both humans and machines [16]. These improvements are the complementary nature of the audio and visual modalities. For example, many sounds that are confusable by ear are easily distinguishable by eye. The improvements from adding the visual modality are often more pronounced in noisy conditions where the audio signal-to-noise ratio (SNR) is reduced.

When developing a speech recognition system that incorporates both the audio and visual modalities, a principled method for integrating the two streams of information must be designed. Because of the success of hidden Markov model (HMMs) in audio speech recognition, most audio-visual speech recognition (AVSR) systems extend HMM techniques to incorporate both modalities. This is describe efforts in developing an AVSR system which is built upon existing segment-based speech recognizer [20]. This AVSR system incorporates information collected from visual measurements of the speaker's lip region using an audio-visual integration mechanism that we call a segment-constrained HMM [19].They are a new unified training algorithm for both the feature extractor and HMM classifiers [21]. They interpret the feature extractor as a multilayer perceptron (MLP) with four layers, i.e., one for the filter banks, one for the feature transformation, and two for the delta and acceleration calculations. It enables us to derive efficient expressions of weight update formulas systematically by back propagation for all of the feature extractor modules. The back propagation starts with the output of HMM classifiers through an efficient inversion algorithm.

Determining both the age and gender of speakers is a complicated task and has received considerable attention in recent years. The achieved are encouraging and are beginning to make it feasible to use this technology as a viable alternative to existing methods of providing user demographics. Age and gender classification systems are generally implemented as a fusion of several subsystems [14], with each subsystem operating using a form of Gaussian mixture model, multilayer perceptron, hidden Markov models and/or support vector machines [18].

If the phone is aware of its owner mood can offer more personal interaction and services. Mobile sensing, in recent years, has gone beyond the mere measure of physically observable events. Scientist studying affective computing [3],[13], have published techniques able to detect the emotional state of the user , allowing the development of emotion-aware mobile applications [9]. Existing work focused on detecting emotions rely on the use of invasive means such as microphones and cameras , and body sensors worn by the user [15]. There is method based on the employment of audio signals represents an efficient alternative to the mentioned approaches.

The general influence of speaker age on voice characteristics is being studied since the late 1950s [1] and sustained continuous attention since then (see e.g. [10]), the first actual systems estimating the age and the gender of the speaker were developed only recently [16],[13]. The quality of these systems is difficult to compare, as they vary considerably regarding the number and distribution of speaker age as well as the types of speech material.

The variability of IVR system use patterns across age and gender is investigated in [13], indicating that dialog strategies tailored to specific age and gender groups can be very useful in improving overall service quality. In this context recognizing people emotional state and giving a suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot research area in both industry and academic field. There is much research in this area and there have been some successful products [8].

2.1. Classification Of Feature Extraction Any audio signal can be classified under a given class, the features in that audio signal are to be extracted. These features will decide the class of the signal. Feature extraction involves the analysis of the input of the audio signal. The feature extraction techniques can be classified as temporal analysis and spectral analysis technique. Temporal analysis uses the waveform of the audio signal itself for analysis. Spectral analysis utilizes spectral representation of the audio signal for analysis [21].

All audio features are extracted by breaking the input signal into a succession of analysis windows or frames, each of around 10-40-ms length, and computing one feature value for each of the windows. One approach is to take the values of all features for a given analysis window to form the feature vector for the classification decision, so that class assignments can be obtained almost in real time, thus realizing a real-time classifier. Another approach is to use the texture window, in which the long-term characteristics of the signal are extracted and the variation in time of each feature is measured, that often provides a better description of the signal than the feature itself. A texture window is a long-term segment in the range of seconds containing a number of analysis windows. In the texture based approach only one feature vector for each texture window is generated. The features are not directly obtained in each analysis window, but statistical measures of the values are obtained for all analysis windows within the current texture window. Therefore in this case real-time classification is not possible, since at least one whole texture window has to be processed to obtain a class decision.

Since the analyzed audio files are supposed to contain only one type of audio, a single class decision is made for each type of audio, which can be derived following one of two possible approaches. The first approach is the single vector mode, which consists of taking the whole file

length as the texture window. In this way, each file is represented by a single feature vector, which in turn is subjected only once to classification. The second approach is the texture window mode, which consists of defining shorter texture windows and making several class decisions along each file, one for each texture window. At the end of the file the decisions are averaged to obtain a final class decision. This average computation is weighted by the certainty of each class decision. As discussed previously feature extraction plays an important role in classification of an audio signal. Hence it becomes all the more important to select those features that help the classification process more efficient. There are different types of features, such as the pitch, timbral features, rhythm features etc that are explained below.

2.2.1 Pitch:

The sound that comes through vocal tract starts from the larynx where vocal cords are situated and ends at mouth. The vibration of the vocal cords and the shape of the vocal tract are controlled by nerves from brain. The sound, which we produce, could be categorized into voiced and unvoiced sounds. During the production of unvoiced sounds the vocal cords do not vibrate and stay open whereas during voiced sounds they vibrate and produce known as glottal pulse. A pulse is a summation of a sinusoidal wave of fundamental frequency and its harmonics (Amplitude decreases as frequency increases). The fundamental frequency of glottal pulse is known as the pitch [21].

In music, the position of a tone in the musical scale is designated by a letter name and determined by the frequency of vibration of the source of the tone. Pitch is an attribute of every musical tone. The fundamental or first harmonic of any tone is perceived as its pitch. Absolute pitch is the position of a tone in the musical scale determined according to its number of vibrations per second, irrespective of other tones. The term also denotes the capacity to identify any tone upon hearing it sounded alone or to sing any specified tone.

For example pitch helps the human ear to distinguish between string instruments, wind instruments and percussion instruments such as the drums, tabla etc.

After the voiced parts of the sound are selected the pitch has to be determined. There are several algorithms currently in use for accomplishing this task. These could be categorized into Time-domain and Frequency-domain analysis. In time domain analysis the pitch could be estimated by using the peaks, but due to the presence of formant frequencies (harmonics) this method could give a wrong estimation. So the formant frequencies are filtered out using a low pass filter and then zero crossing methods or any other suitable method is used to determine the pitch. The

speech signal is also passed through a low pass filter in the frequency domain analysis and then the pitch is determined by analyzing the spectrum.

2.2.2. Timbral features:

Sound "quality" or "timbre" describes those characteristics of sound, which allow the ear to distinguish sounds that have the same pitch and loudness. Timbre is then a general term for the distinguishable characteristics of a tone. Timbre is mainly determined by the harmonic content of a sound and the dynamic characteristics of the sound such as vibrato and tremolo.

In music timbre is the quality of a musical note that distinguishes different types of musical instrument. Each note produced by a musical instrument is made of a number of distinct frequencies, measured in hertz (Hz). The lowest frequency is called the fundamental and the pitch produced by this frequency is used to name the note. However, the richness of the sound is produced by the combination of this fundamental with a series of harmonics and/or partials (also collectively called overtones). Most western instruments produce harmonic sounds, and these can be calculated by multiplying the fundamental by an increasing series of numbers - x2, x3, x4, etc (whole number multiples). However many instruments produce inharmonic tones, and may contain overtones which are not whole number multiples, these being the partials. Therefore, when the orchestral tuning note is played, the sound is a combination of 440 Hz, 880 Hz, 1320 Hz, 1760 Hz and so on. The balance of the amplitudes of the different frequencies is responsible for giving each instrument its characteristic sound.

The ordinary definition of vibrato is periodic changes in the pitch of the tone, and the term tremolo is used to indicate periodic changes in the amplitude or loudness of the tone. So vibrato could be called FM (frequency modulation) and tremolo could be called AM (amplitude modulation) of the tone. Actually, in the voice or the sound of a musical instrument both are usually present to some extent. Vibrato is considered to be a desirable characteristic of the human voice if it is not excessive. It can be used for expression and adds richness to the voice. If the harmonic content of a sustained sound from a voice or wind instrument is reproduced precisely, the ear can readily detect the difference in timbre because of the absence of vibrato.

The following are some of the timbral features,

2.2.2.1. Zero crossings:

The zero crossings feature counts the number of times that the sign of the signal amplitude changes in the time domain in one frame. For single-voiced signals, zero crossings are used to

make a rough estimation of the fundamental-frequency. For complex signals it is a simple measure of noisiness.

2.2.2.2. Centroid:

The centroid is the measure of the spectral shape and higher centroid values correspond to brighter textures with more high frequencies. Centroid models the sound sharpness. Sharpness is related to the high-frequency content of the spectrum. Higher centroid values correspond to spectra in the range of higher frequencies. Due to its effectiveness to describe spectral shape, centroid measures are used in audio classification tasks.

2.2.2.3. Roll off:

The roll off is defined as the frequency below which 85% of the magnitude distribution of the spectrum is concentrated. Like the centroid, it is also a measure of spectral shape and yield higher values for high frequencies. Therefore it can be said that there exists a strong correlation between both the features. If M is the largest value of k for which this equation is satisfied then this frequency M is the roll off.

2.2.2.4. Flux:

The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions that correspond to successive signal frames. Flux is an important feature for the separation of music from speech.

2.2.2.5. Mel frequency Cestrum coefficients (MFCC s):

MFCCs are a compact representation of the spectrum of an audio signal taking into account the nonlinear human perception of pitch, as described by the mel scale. They are one of the most used features in speech recognition and have recently been proposed to analyze and represent musical signals. MFCCs are computed by grouping the Short Time Fourier Transform (STFT) coefficients of each frame into a set of 40 coefficients, using a set of 40 weighting curves that simulate the frequency perception of the human hearing system. Then the logarithm of the coefficients is taken, and a discrete cosine transform (DCT) is applied to de- correlate them. Normally the five first coefficients are taken as features.

3. PROBLEM DEFINATION

A new gender and age group recognition approach based on Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender and age group of unseen test speakers.

Automatic recognition of gender and age from voice has increased the attention in recent years. Existing method identifies gender and age from facial images using GMM based HMM ,buy studying facial features. Also by using Hybrid DNN-HMM model identifies gender and age based on emission recognition. In some small duration of proposed method uses energy features to identifies age and gender.

4. PROPOSED WORK

4.1. Proposed work:

In many criminal cases, evidence might be in the form of recorded conversations, possibly over the telephone. Therefore, law enforcement agencies have been concerned about accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate to identify him/her or at least narrow down the number of suspects. Here they propose a new gender and age group recognition approach based on Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender and age group of unseen test speakers.

Proposed Work:

1. Designing a new HMM-based approach for speaker, gender and age estimation, this improves the accuracy of the state-of-the-art speaker age estimation methods with statistical significance.

2. Analysing the effect of major factors influencing the automatic gender and age estimation systems.

4.2. Age and Gender using Speech Recognition System Architecture:

Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. The Front-End block acquires and samples it with frequency in order to obtain the discrete sequence. After this step, a feature

vector is computed by the Features Extraction block. The automatic speech recognition and speaker identification, content-based multimedia indexing systems, interactive voice response systems, voice synthesis and smart human-computer interaction.

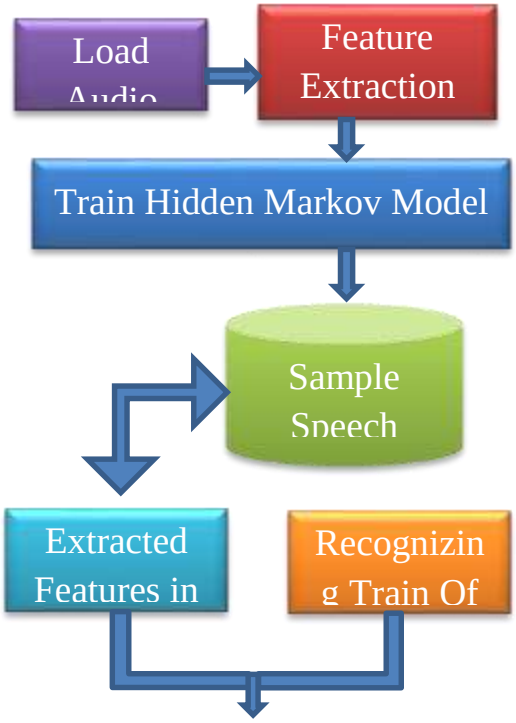

Fig.4.1. Speech recognition scheme overall architecture in age & gender recognition by using HMM.

The Gender-driven speech Recognition block that provides the output of the overall process: the recognized speech. The recognition of the gender is used as input for the age identification.

5. METHODOLOGY

5.1. Hidden Markov Models:

In the Markov model each state corresponds to one observable event. But this model is too restrictive, for a large number of observations the size of the model explodes, and the case where the range of observations is continuous is not covered at all. The Hidden Markov concept extends the model by decoupling the observation sequence and the state sequence. For each state a

probability distribution is defined that specifies how likely every observation symbol is to be generated in that particular state. As each state can now in principle generate each observation symbol it is no longer possible to see which state sequence generated a observation sequence as was the case for Markov models, the states are now hidden, hence the name of the model. A Hidden Markov model can be defined by the following parameters:

- The number of distinct observation symbols M.
- An output alphabet ={ } 1 2 M V v ,v ,...v
- The number of states N.
- A state space Q = {1, 2,...N}

States will usually be indicated by i, j a state that 'the model is in' at a particular point in time t will be indicated by qt.

Thus, qt = i means that the model is in state i at time t.

A probability distribution of transitions between states, where $aij = P(qt + 1 = j|qt = i) \quad 1 \leq i, j \leq N$

Real-world processes generally produce observable outputs which can be characterized as signals. The signals can be discrete in nature or continuous in nature. The signal source can be stationary (i.e., its statistical properties do not vary with time), or non-stationary (i.e., the signal properties vary over time). The signals can be pure or can be corrupted from other signal sources or by transmission distortions.[14]

A problem of fundamental interest is characterizing such signal in terms of signal model, signal model gives us:

•Theoretical description of a signal processing system which can be used to process the signal and so as to provide desired output.

• It helps up to understand great deal about signal source without having to have source available.

6. Comparisons Results in HMM & SVM:

6.1. Table:

Table 6.1.: Comparative Results in HMM & SVM

Age & Gender classification	HMM (Shripi dataset)	SVM (Shripi dataset)	SVM_wav Eustace_wav
Fem20	85.9024	81.4515	81.4587
Fem40	85.878	81.2436	81.1428
Fem60	85.6567	81.5282	81.6272
mal20	85.778	81.5375	81.7271
mal40	85.9024	81.5977	81.0847
mal60	85.5665	81.4078	81.4565

6.2. Graph:

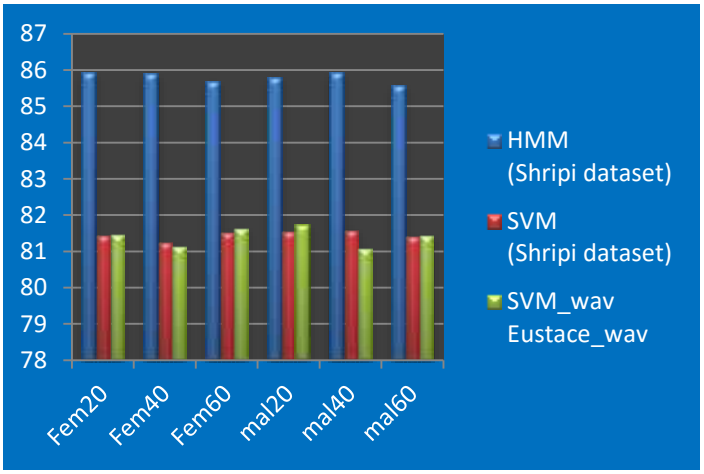

Fig:6.1. Comparative Results in HMM &SVM

HMM (Shripi dataset) compared with SVM (Shripi dataset) and SVM (Eustace_wav) and it is observed that the proposed model showed better recognition rate.

7. FUTURE SCOPE AND CONCLUSION

Deep Neural Network Hidden Markov Models, or DNN-HMMs, are recently very promising acoustic models achieving good speech recognition results over Gaussian mixture model based HMMs (GMM-HMMs). HMM-based approach for speaker, gender and age estimation, this improves the accuracy in very long time periods , but the SVM-based approach for speaker, gender and age estimation, improves the accuracy in very short time periods. The statistical analysis should that through SVM is fast, accuracy of HMM is better from SVM.

In future, hybrid approach using HMM & SVM model can improves the results.

BIBLIOGRAPHY

1. I. BISIO, A. DELFINO, F. LAVAGETTO, M. MARCHESE, and A. SCIARRONE, \Gender-driven emotion recognition through speech signals for ambient intelligence applications," VOLUME 1, NO. 2, DECEMBER 2013.
2. I. Levent M. Arslan, Member and I. John H. L. Hansen, Senior Member, \Selective training for hidden markov models with applications to speech classification," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 7, NO. 1, JANUARY 1999.
3. L. Li, D. J. Yong Zhao, I. G. Yanning Zhang, Fengna Wang, E. Valentin, and H. Sahli, \Hybrid deep neural network - hidden markov model (dnn-hmm) based speech emotion recognition," IEEE, 2013.
4. K. Rakesh, S. Dutta, and K. Shama, \Gender recognition using speech processing techniques in lab view," International Journal of Advances in Engineering and Technology, May 2011.
5. D. Zhang, Y. Wang, and B. Bhanu, \Age classification based on gait using hmm," 2010.
6. J.-H. Im and S.-Y. Lee, \Uni_ed training of feature extractor and hmm classi_er for speech recognition," IEEE SIGNAL PROCESSING LETTERS, VOL. 19, NO. 2, FEBRUARY 2012.
7. J.-C. Lin, C.-H. Wu, I. Senior Member, and W.-L. Wei, \Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 1, FEBRUARY 2012.
8. M. I. Wooil Kim and I. John H. L. Hansen Fellow, \Time-frequency correlation-based-missing-feature re- construction for robust speech recognition in band-restricted conditions," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PRO- CESSING, VOL. 17, NO. 7, September 2009.
9. A. I. Garca-Moral, R. Solera-Urea, S. M. IEEE, C. P.-M. M. IEEE, and F. D. de Mara Member IEEE, \Data balancing for e_cient training of hybrid ann /hmm automatic speech recognition systems,"

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 3, MARCH 2011.

10. A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mario, \Speech emotion recognition using hidden markov models," Euro speech 2001.

11. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, \Deep nural network for acoustics modeling in speech recognition," NOVEMBER 2012.

12. I. A. I. G.-M. C. P.-M. M. I. M. M.-R. S. M. I. Rubn Solera-Urea, Member and I. Fer-nando Daz-de Mara, Member, \Real-time robust automatic speech recognition using compact support vector machines," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 4, May 2012.

13. Z.-H. T. Sven Ewan Shepstone, Member IEEE, S. M. IEEE, and S. M. I. Sren Holdt Jensen, \ Audio-based age and gender identification to enhance the recommendation of tv content," IEEE, 2013.

14. H. S. Supervisors, P. P. Rao, and D. S. D. R. M.Tech, \Audio signal classification," Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Submitted, November2004.

15. N. MINEMATSU, M. SEKIGUCHI, and K. HIROSE, \Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," IEEE, 2002.

16. I. Timothy J. Hazen, Member, \Visual model structures and synchrony constraints for audio-visual speech recognition," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 3, may 2006.

17. J. Picone, \Continuous speech recognition using hidden markov models," IEEE ASSP MAGAZINE, JULY 1990.

18. I. T. N.-H. Z.-Z.-H. L.-T. T. M. I. K. T. M. I. S. K. S. M. I. Junichi Yamagishi, Member and I. Steve Renals, Member, \Robust speaker-adaptive hmm-based text-to-speech synthesis," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 6, AUGUST 2009.

19. Bhavana R. Jawale. , **Swati patil.** "Identification of Age And Gender Using HMM "et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1643-1647.