EXTRACTION OF VISUAL AND ACOUSTIC FEATURES OF THE DRIVER FOR REAL-TIME DRIVER MONITORING SYSTEM

Sandeep Kotte
Department of Computer Science & Engineering,
Dhanekula Institute of Engineering & Technology, Vijayawada, India.

ABSTRACT
Driving is one of the most dangerous tasks in our everyday lives. Statistics show that over the past couple of decades the majority of the accidents are not only due to the poor vehicle technical conditions but also due to the driver's inattentiveness. The major cause for the inattentiveness includes aspects such as drowsiness (sleepy), fatigue (lack of energy) and emotions/stress (for example sad, angry, joy, pleasure, despair and irritation). In order to improve attentiveness of the driver, the Indian Government has introduced regulations amongst others concerning driving time and rest periods for the drivers. Even though these regulations helped partially in improving the driver attentiveness, they largely ignore the real-time dynamic ergonomics which is influenced by diverse factors such as the traffic control strategies, road geometry, vehicle characteristics, changing traffic scenarios, weather, etc.
Different approaches have been proposed for monitoring the driver states, especially drowsiness and fatigue. Fatigue is traditionally measured by observing the eyelid movements. The drowsiness is generally measured by analyzing either head movements patterns or eyelid movements or face expressions or all the lasts together. Concerning emotion/stress recognition visual sensing of face expressions is helpful but generally not always sufficient. Therefore, one needs additional information that can be collected in a non-intrusive manner in order to increase the robustness of the emotion/stress measurement in the frame of a non-intrusive monitoring policy. We choose and find acoustic information emanating from the driver to be appropriate in the analysis of the emotions of driver, provided the driver generates some vocal signals by speaking, shouting, crying, etc., which is not un-common during the driving process. From these acoustic signals, this work extracts the spatial and temporal acoustic features and correlates it to the emotions of the driver.
In this paper, a demonstration on how one can distinguish the emotion based on acoustic features (or combination of features) by testing them over Berlin emotion database.

KEYWORDS: *Feature Extraction, Fatigue, Classifier, LDA, Berlin Database*

I. INTRODUCTION

Driver monitoring plays a major role in order to assess, control and predict the driver behavior. The research concerning driver monitoring systems was started nearly from the 1980's. In the first stages of this research, researchers developed driver monitoring systems based on inferring both driver behavior and state from the observed/measured vehicle performance. In a following generation, driver state and behavior has been directly assessed by intrusive systems measuring the physiological characteristics. But these techniques require driver cooperation since they are intrusive. Further, they may also disturb the driver behavior. And more recently, a significant research has been focusing on developing non-intrusive techniques, generally based on machine vision, directly measure in a non-intrusive manner the driver state.

1.1. Theoretical background

Due to increased number of speech driven applications in the recent years the automatic assessment of emotions from the drivers speech signal has become a research interest [1][2]. Such assessment

provides information about the driver's satisfaction with the cars infotainment system and in particular, it improves efficiency and friendliness of human-machine interfaces. Moreover it reflects the driver's perception of the traffic situation and thus reveals his/her stress level. Therefore the emotional state also affects the driving capability which is of utmost importance for the safety of all occupants [4][3]. It allows for monitoring of physiological state of individuals in several demanding work environments, can be used to augment automated medical or forensic data analysis systems [5].

The performance of such a system is studied in the acoustically demanding environment of vehicular noise while driving. These systems mainly focus on (a) electro physiological data (e.g. EEG: [7][6]), and (b) behavioural expression data (gross body movement, head movement, mannerism, and facial expression; [8]) in order to characterize the user state. But these electrode-(EOG/EEG reaching 15% error rate for fatigue detection; [9]) or video based instruments still do not fulfill the demands of an everyday life measurement system. They have some shortcomings like (a) lack of robustness against environmental and individual-specific variations (e.g. bright light, wearing correction glasses, and angle of face or being of Asian race) and (b) lack of comfort and longevity due to electrode sensor application.

In contrast to these electrode or video-based instruments, the utilization of voice communication as an indicator for emotional states matches the demands of everyday life measurement. Contact free measurements as voice analysis are non-obtrusive (not interfering with the primary driving task) and favorable for emotion detection, since an application of sensors would cause annoyance, additional stress and often impairs working capabilities and mobility demands.

In addition, speech is easy to record even under extreme environmental conditions (temperature, high humidity and bright light), requires merely cheap, durable, and maintenance free sensors and most importantly, it utilizes already existing communication system hardware. Furthermore, speech data is omnipresent in many professional driver settings. Given these obvious advantages, the renewed interest in computational demanding analyses of vocal expressions has been enabled just recently by the advances in computer processing speed [12][13][10][11]. The first investigations to emotion detection from speech were conducted around the mid of the 1980s using statistical properties of certain acoustic features [15][14]. Later, the evolution of computer architectures introduced the detection of more complicated emotions from the speech.

The research towards detecting human emotions is increasingly attracting the attention of the research community [15]. Nowadays, the research is focused on finding powerful combinations of classifiers that increase the classification efficiency in real-life speech emotion detection applications. Some of these techniques are used to recognize the frustration of a user and change their response automatically. Speech based emotion detection has lots of useful applications. Some of them are human-robotic interfaces [16], smart call-centers [18][19][17], intelligent spoken tutoring systems [20], spoken dialog research.

The emotions can be observed from information about the language, what we say and how we say it. How we say something is more important than what we say. In the literature [21], the main focus is on the phonetics and acoustic properties of the affective spoken language. The emotions like anger, joy and sadness affect the driver attentiveness and are therefore relevant for the driving process. Thus we will devote particular attention in this research to those emotions that are badly/negatively affecting the driver behaviour [22]. The important voice features to consider for emotion classification are: Fundamental frequency (F0) or Pitch, Intensity (Energy), Speaking rate, Voice quality and many other features that may be extracted/calculated from the voice information are the formants, the vocal tract cross-section areas, the MFCC (Mel Frequency Cepstral Coefficient), Linear frequency cepstrum coefficients (LFCC), Linear Predictive Coding (LPC) and the teager energy operator-based features [15]. Certain features in the voice of a person can be used to infer the emotional state of the particular speaker. The real-time extracting the voice characteristics conveys emotion and attitude in a systematic manner and it is different from male and female.

Table 1: Some variations of acoustic variables observed in relation to emotions [21]

Emotion	Pitch	Intensity	Speaking rate	Voice quality
Anger	High mean	Wide increased range	Increased	Breathy; blaring timbre
Joy	Increased mean	Increased range	Increased	Sometimes breathy; moderately Blaring timbre
Sadness	Normal or lower Decreased than normal mean	Narrow range	Slow	Resonant timbre

Feature extraction from visual and acoustic information of the driver is an important and basic task to know the driver behavior/emotions. Generally, a feature is a set of measurements. Each measurement contains a piece of information, and specifies the property or characteristics of the object present in the given input. In the daily life, humans are able to guess and understand the state of the other persons by observing multiple features such as body action, voice information and the interpreted knowledge of understanding what he is saying and how he says it. Observing and extracting the multiple features is less complex task (in some cases it is obvious) for humans, for machines it is much more complex. This work mainly focuses on (a) identifying the useful features in the acoustic information (b) extracting the features in real time without missing the frames and (c) correlate the features to parameters/dimensions of the "extended ergonomics status" vector.

II. NON-INTRUSIVE FEATURE EXTRACTION APPROACHES

Feature extraction is used to reduce the large input data into smaller data and it converts the data into small features sets or feature vectors (n-dimensional vector to store numerical features which represents an object). Feature extraction is defined as the process of extracting the feature from a source data, where the data can be embedded from high dimensional data set [23]. We calculate different feature sets for different applications. For computer vision applications edges, corners are calculated as features for images. Features like data and noise ratio, length of sound and relative power are calculated for pattern recognition applications.

2.1. Feature Extraction from the Visual Information

In 1990's, researchers introduced appearance based linear subspace techniques, statistics related techniques, to reduce the dimensionality and to extract the useful visual features. The introduction of the linear subspace techniques is a milestone in the visual feature extraction concept. The performance of appearance based techniques heavily depends upon the quality of the extracted features from image [23]. The appearance based linear subspace techniques extract the global features, as these techniques use the statistical properties like the mean and variance of the image [24]. The major difficulty in applying these techniques over large databases is that the computational load and memory requirements for calculating features increase dramatically for large databases [24]. In order to increase the performance of the feature extraction techniques, the nonlinear feature extraction techniques are introduced. In order to improve the performance of the emotion recognition systems, we have to extract both linear and nonlinear features. We have many nonlinear feature extraction techniques, such as radon transform and wavelet transform. The radon transform based nonlinear feature extraction gives the direction of the local features. This process extracts the spatial frequency components in the direction of radon projection is computed [25]. When features are extracted using radon transform, the variations in this facial frequency are also boosted [25]. The wavelet transform gives the spacial and frequency components present in an image. The performance of these feature extraction approaches are systematically evaluated in our previous work over FERET database for face recognition application as shown in Figure. 1.

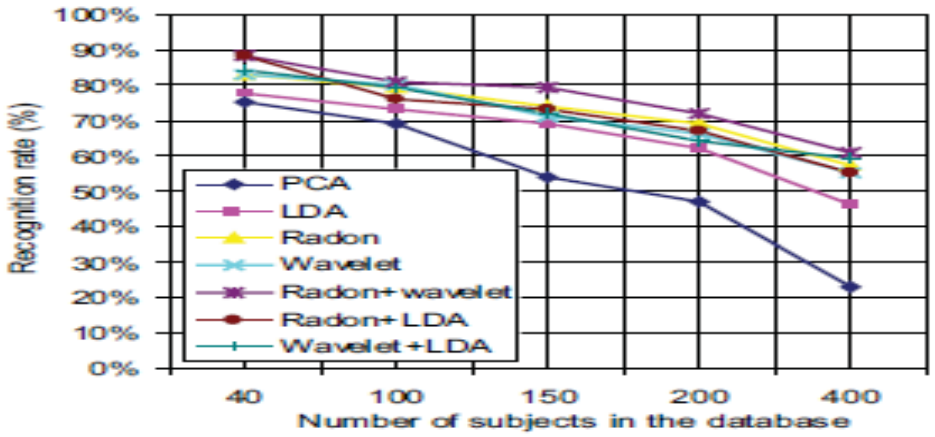

Figure 1: On Performance comparison of different face recognition approaches with profile right images [25].

2.2. Feature Extraction from the Acoustic Information

Speech is easy to record even under extreme environmental conditions (temperature, high humidity and bright light), requires merely cheap, durable and maintenance free sensors and most importantly, it utilizes already existing communication system hardware. Furthermore, speech data is omnipresent in many professional driver settings. Given these obvious advantages, the renewed interest in computational demanding analyses of vocal expressions has been enabled just recently by the advances in computer processing speed. The first investigations to emotion recognition from speech were conducted around the mid of the 1980s. Due to increased number of speech driven applications in the recent years the automatic assessment of emotions from the drivers speech signal has become a research interest [26][27]. Such assessment provides information about the driver's satisfaction with the cars infotainment system and in particular, it improves efficiency and friendliness of human-machine interfaces.

For the machine based state estimation, we will not focus on the voice content but rather on voice-signal features that are relevant for an emotional state inference. In this regard, to make the system more robust in predicting the driver state we will analyze the acoustic information such as pitch, intensity, speaking rate, voice quality, etc. in order to extract appropriate features [27]. Acoustic features extraction is challenging in many aspects as it highly depends on the age and gender of the person. The acoustic features are quite varying for different age groups and different gender. Angry males show higher levels of energy than angry females. It is found that males express anger with a slow speech rate as opposed to females who employ a fast speech rate under similar circumstances [27]. Acoustic information will however be precious, whenever available, to better assess and understand the effect of driving process ergonomics on the driver state and mood.

The real-time extraction of acoustic characteristics from the voice signal conveys emotion and attitude in a systematic manner and it is different from male and female. Acoustic features are used to recognize the frustration of the driver (i.e. recognize vocal signals by shouting, crying, etc which are not un-common during the driving process). The performance also depends upon the given (input) acoustic information. In this work, input data is taken from audio sensors like microphone and the feature extraction algorithms are executed to extract the features in real time. Features are extracted from the real time data by performing time and frequency domains algorithms [23]. These algorithms extract temporal, spectral features and cepstral coefficients. These features are extracted based on the amplitude and spectrum analyzer of the audio data. The basic approach to the extraction of acoustic features is frame blocking such that a stream of given input audio signal is converted to a set of frames. And the time duration of each frame is about 10~30ms. If the frame size is shorter than 10ms we may miss some important information and sometimes we cannot extract valid acoustic features. If the frame size is longer than 30ms redundancy may occur and we cannot catch the time-varying characteristics of the audio signals.

The important voice features to consider for emotion classification are: Fundamental frequency (F0) or Pitch, Intensity (Energy), Speaking rate, Voice quality and many other features that may be

extracted/calculated from the voice information are the formants, the vocal tract cross-section areas, the MFCC (Mel Frequency Cepstral Coefficient), Linear frequency cepstrum coefficients (LFCCs), Linear Predictive Coding (LPC) and the teager energy operator-based features [27]. Pitch is the fundamental frequency of audio signals, which is equal to the reciprocal of the fundamental period. It can also be defined as the highness or lowness of a sound.

Generally for pitch estimation wavelet transforms is used. The shape of the vocal track is modified depending up on the emotion [28]. The MFCC is "spectrum of the spectrum" used to find the number of voices in the speech. It has been proven beneficial in speech emotion detection, and speech detection tasks. The teager energy operator is used to find the number of harmonics due to nonlinear air flow in the vocal track [29][30]. The LPC provides an accurate and economical representation of the envelope of the short-time power spectrum. One of the most powerful speech coding analysis techniques providing very accurate estimates of speech parameters and is known as being relatively efficient for computation at the same time. The LFCC is similar to MFCC but without the perceptually oriented transformation into the Mel frequency scale; emphasize changes or periodicity in the spectrum, while being relatively robust against noise. These features are measured from the mean, range, variance and transmission duration between utterances [26][27]. To calculate these voice features, different techniques are used. The variation in the feature set for Happy is shown in Figure. 2 and Figure. 3 respectively.

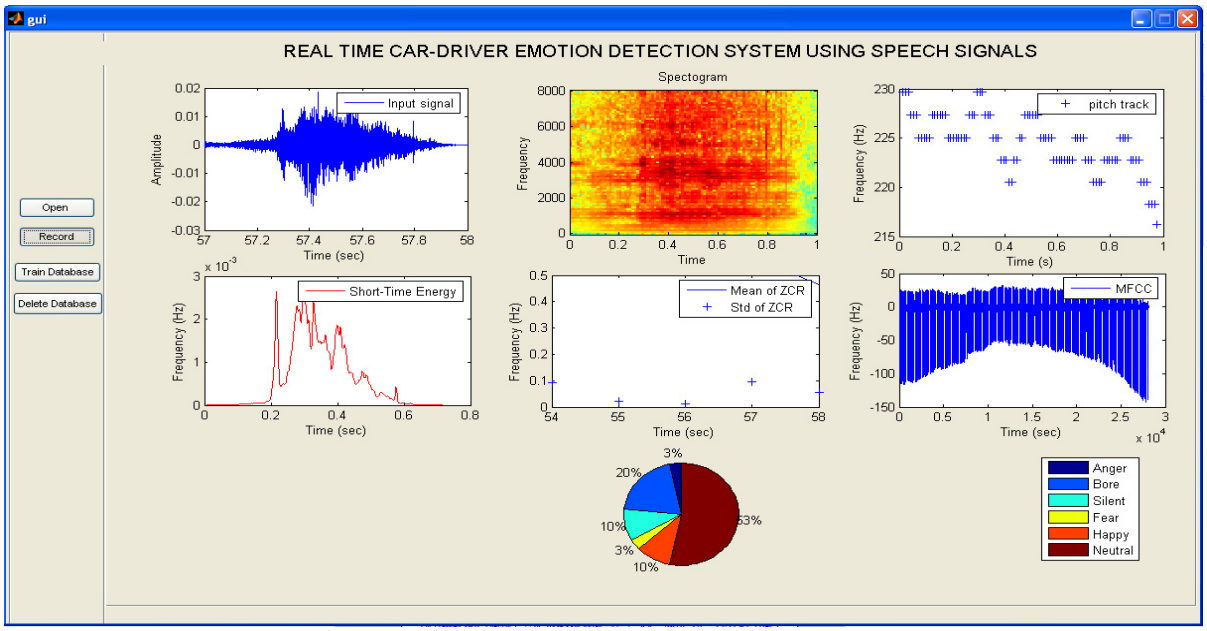

Figure 2: Extracted features from the joy emotional audio file.

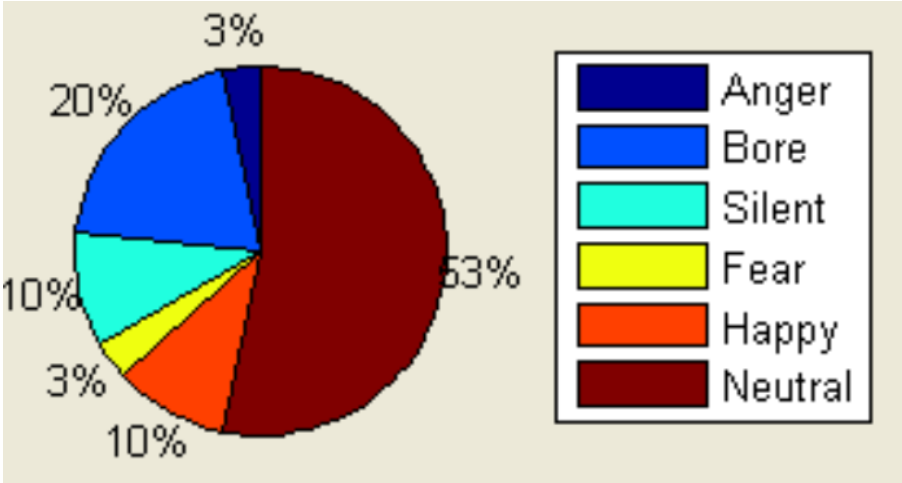

Figure 3: Detected emotions in 58 seconds representing in a pie.

III. EMOTION CLASSIFICATION

Emotion detection has been implemented on a variety of classifiers including Fisher's linear discriminant analysis (FLDA), maximum likelihood classifier (MLC), neural network (NN), k-nearest neighbor (k-NN), and Gaussian mixture model (GMM) [27]. The main criterion in evaluating the effectiveness of the classification algorithm in this work is the scalability of the classifier. By considering the importance of scalability of the classification algorithm, the algorithms are evaluated over the open source Berlin database. For validation purposes, we create the test and training set databases using Berlin Database [31]. A training set database contains

- 535 utterances speech recordings
- 10 actors: 5 males, 5 females
- 10 utterances (in German) by each voice
- Seven different emotional speeches (anger, joy, neutral, boredom, fear, sadness and disgust)[31].

The test database contains different emotions of the persons present in the training database. The backend classifier is working with a fixed length feature vector and is used for classification as shown in Figure. 4.

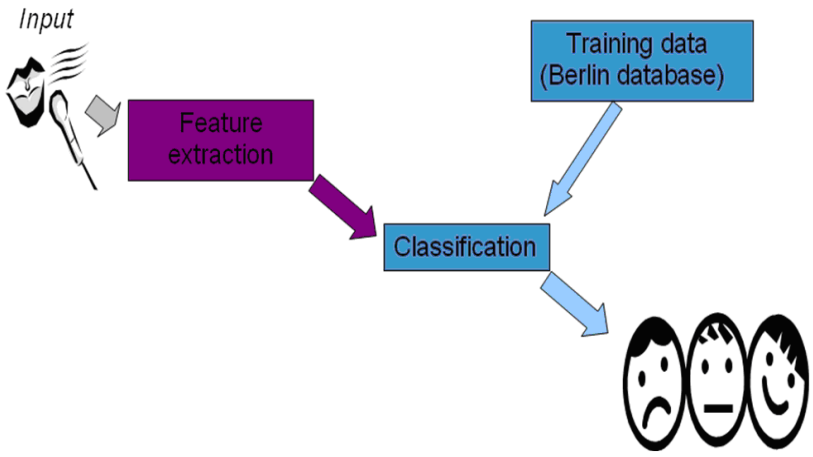

Figure 4: The overall feature extraction approach from acoustic information.

Linear discriminant analysis (LDA) classifier is used for feature selection (i.e. converts the variable large dimensional feature vectors into the fixed small dimensional feature vector). LDA uses information about the class [23]. A class contains one person with different emotions. LDA tries to maximize the between class variance and minimize the within class variance. In other words, it decreases the distance between same class files and increases the distance between different class files [23]. Because of that LDA easily recognizes the emotions among large databases. The success rate of the different popular classifiers is compared in Table 2.

Table 2: Major Classification techniques and their Success rates

Classification techniques	Success rate
Linear discriminant analysis (LDA)	67.22%
k-nearest neighbor (k-NN)	63.33%
neural network (NN)	57.78%
maximum likelihood classifier (MLC)	55%
Gaussian mixture model (GMM)	53.33%

3.1. Proposed Classifier

The LDA performs considerably better when compared to above classifiers for Berlin database. But the performance of LDA is also not sufficient for real world applications.

The scalability of the real world emotion recognition system is limited, as the computational load and memory requirements increase dramatically with the large data sets. So scalability is major issue here. Up to now, in LDA we are using linear metric (Euclidean distance). Linear metric cannot compare different dimensions of the acoustic feature vectors accurately.
To improve the performance (success rate and process speed), we propose the nonlinear Hausdorff metric based LDA. A special nonlinear metric Hausdorff distance which is able to compute the distance between different sized matrices having a single common dimension, like the acoustic matrices representing our acoustic feature vectors [32][33].

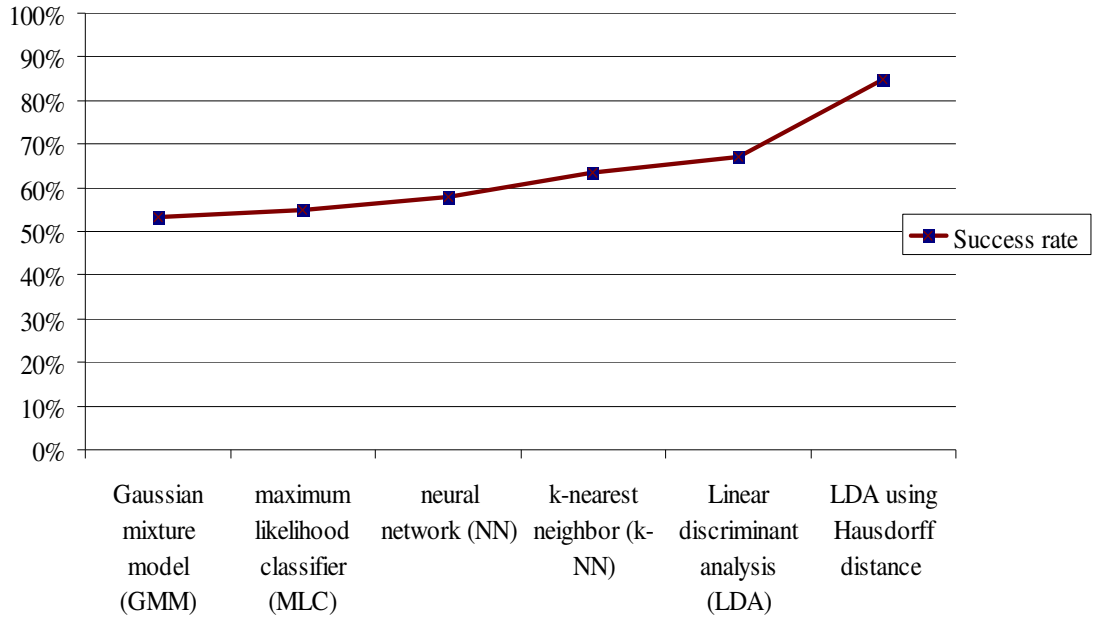

Figure 5: The graphical representation of the success rate for different classifiers.

Hausdorff distance measure is often used in content-based retrieval applications [32][33]. Hausdorff distance is meant as a measure between two point collection A and B in a metric space S (whose distance is d), it can be viewed as dissimilarity measure between two feature vectors A and B. By considering the hausdorff distance measure instead of the linear eculidean distance measure, the success rate of the LDA algorithm is increased by around 20 % as shown in Figure. 5.

3.2. Summary

In this chapter, Linear Discriminant Analysis (LDA) is explained in detailed. LDA only uses a second order statistics. LDA is using any information about the class. It tries to maximize the between class variance and minimize the within class variance.

IV. SYSTEM ARCHITECTURE & IMPLEMENTATION RESULTS

The overall architecture of the system is shown in below Fig: 6. The architecture depicts the process of transforming given input speech signals to driver emotions.

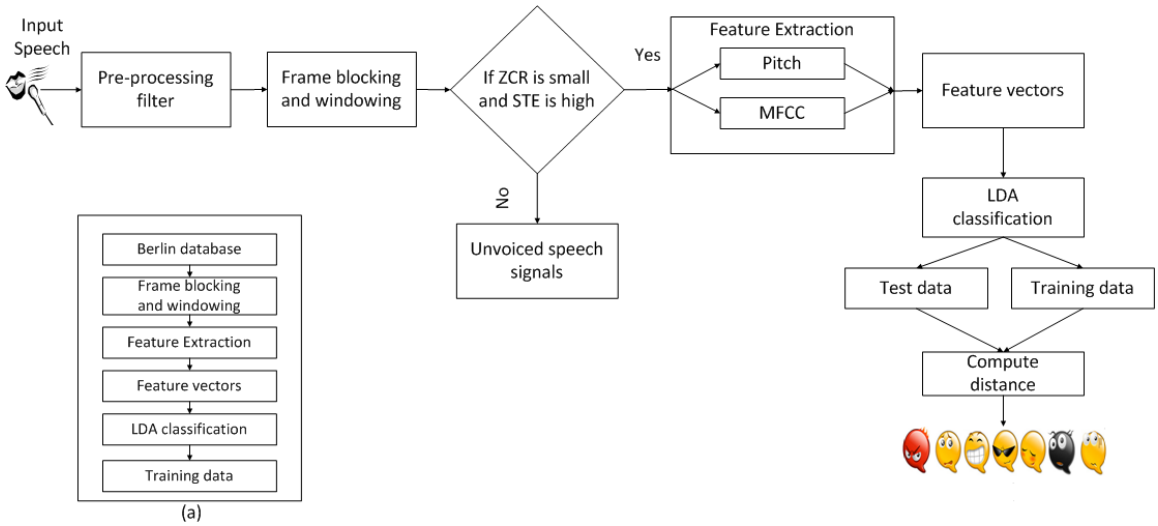

Figure 6: The overall architecture of the emotion detection system. (a)Training

4.1. Signal Preprocessing

As the input data is recorded using audio sensors like microphone, the recorded data may be affected by noise due to the weather conditions or any other disturbances. To reduce the noise affect, we performed filter operations which also optimize the class separability of features. This filter operation is performed with pre-emphasis high pass filter.

The main goal of pre-emphasis is to boost the amount of energy in the high-frequencies. Mainly boosting is used to get more information from the higher formants available to the acoustic model and to improve the phone recognition performance. Lower frequencies have more energy in voiced segments compared to higher frequencies and this is called spectral tilt [34]. The Figure.7 is a plot of the frequency response of before & after filtering.

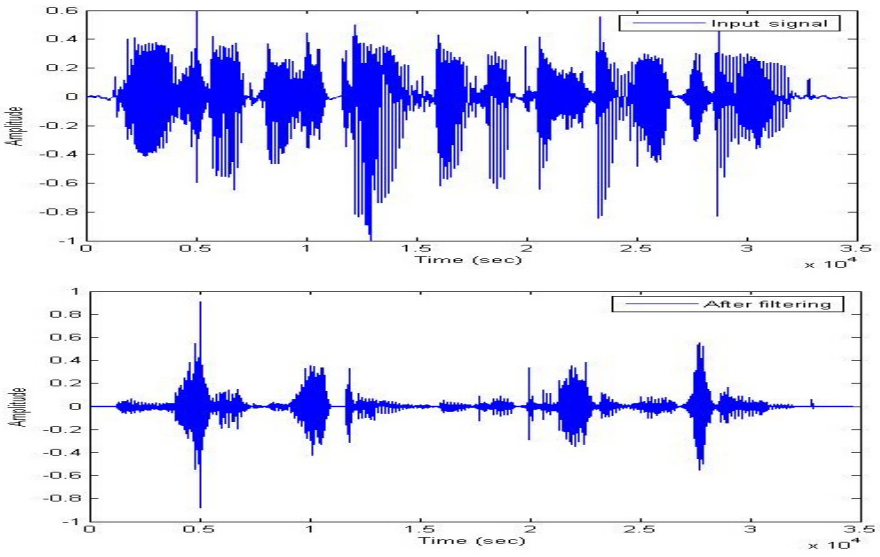

Figure 7: Frequency Response of Signal Preprocessing High Pass Filter (a) original

4.2. Frame Blocking and Windowing

In the window operation, the large input data is divided into small data sets and stored in sequence of frames. While dividing, some of the input data may be discontinuous. So to achieve the continuity the sequences of frames are overlapped. This window operation is performed using hamming window method to reduce the spectral leakage in the input data.

4.3. Zero Crossing Rate and Short Time Energy

In time domain, the input data is recorded for every instance of time. The audio features in time domain are calculated based on the amplitude of the audio data and the variation of amplitude. These audio features are also known as temporal features.
The rate at which zero crossings occur is a simple measure of a signal. Zero-crossing rate is measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero.
The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations [35][36].

4.4. Mel frequency cepstral coefficient (MFCC)

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech.
It turns out that humans perceive sound in a highly nonlinear way. Basic parameters like pitch and loudness highly depend on the frequency, adding weight to components at lower frequencies. In Fig: 8 this behaviour is illustrated, relating the perceived pitch to the physical frequency. The pitch associate with a tone is measured on the so-called mel-scale (By definition 1,000 mels correspond to the perception of a sinusoidal tone at 1000kHz, 40dB above the hearing threshold). The graph clearly shows that the perceived pitch increases all the more slower as we go to higher frequencies. Essentially we observe a logarithmic increase that is illustrated by the almost linear curve (with respect to a logarithmic scale) in Figure: 8 at high frequencies. MFCCs extensively use this property and add weight to lower frequencies, because more discriminative information can be found there.

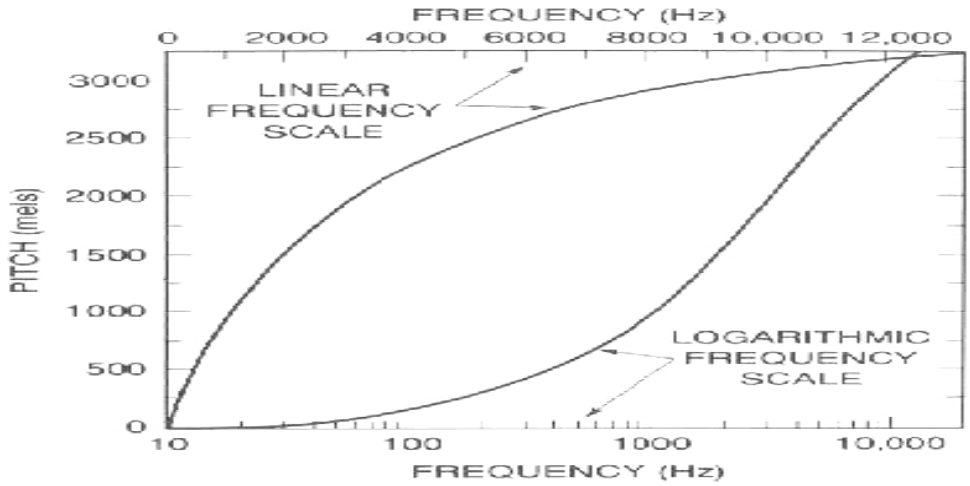

Figure 8: Relation between the perceived pitch and frequency [38]

4.5. Pitch Extraction

Pitch is the fundamental frequency of audio signals, which is equal to the reciprocal of the fundamental period. This is mainly explained in terms of highness or lowness of a sound. Pitch in reality can be defined as the repeat rate of a complex signal, i.e., the rate at which peaks in the autocorrelation function occur. The three main difficulties in pitch extraction arise due to following factors.

- Vocal cord vibration does not necessarily have complete periodicity especially at the beginning and end of the voiced sounds.
- From speech wave, vocal cord source signal can be extracted but its extraction is difficult if has to be extracted separately from the vocal tract effects.

- The fundamental frequency possesses very large dynamic range.

4.6. Berlin Database of Emotional Speech (BDES)

It is developed by the Technical University, Institute for Speech and Communication, Department of Communication Science, Berlin [37]. It has become one of the most popular databases used by researchers on speech emotion recognition, thus facilitating performance comparisons with other studies. 5 actors and 5 actresses have contributed speech samples for this database it mainly has 10 German utterances, 5 short utterances and 5 longer ones and recorded with 7 kinds of emotions: happiness, neutral, boredom, disgust, fear, sadness and anger. The sentences are chosen to be semantically neutral and hence can be readily interpreted in all of the seven emotions simulated. Speech is recorded with 16 bit precision and 48 kHz sampling rate (later down-sampled to 16 kHz) in an anechoic chamber.

4.7. Summary

In this chapter, pitch, zero-crossing rate, short time energy and MFCC is explained in detailed. Zero-crossing rate and short time energy are calculated for the voiced and unvoiced signals.
After adding energy, delta, and double delta features to the 12 cepstral features, totally 39 MFCC features are extracted. Again, one of the most useful facts about MFCC features is that the cepstral coefficients tend to be uncorrelated, which turns out to make our acoustic model much simpler.

V. CONCLUSION

Acoustic information will be precious whenever available, to better assess and understand the effect of driving process ergonomics on the driver state and mood, since these are drawn in an non-intrusive manner. The acoustic features are quite varying for different age groups and different gender.
For the machine based state estimation, we will not focus on the voice content but rather on voice-signal features that are relevant for an emotional state inference. In this regard, to make the system more robust in predicting the driver state we analyzed the acoustic information such as pitch, short time energy, zero crossing rate and MFCC etc. in order to extract appropriate features.
From the literature emotion recognition based on acoustic information has been implemented on a variety of classifiers including Fisher's linear discriminant analysis (FLDA), maximum likelihood classifier (MLC), neural network (NN), k-nearest neighbor (k-NN), Bayes classifier, support vector classifier, artificial neural network (ANN) classifier and Gaussian mixture model (GMM) etc.
An experimental result shows that the LDA classifier with linear metric produces 67.22% recognition rate. To improve the recognition rate later we used a special non linear metric called Hausdorff distance measure. The recognition rate is improved to 85% approximately. The overall system does determine driver's emotions such as anger, despair, pleasure, sadness, irritation and joy.

VI. FUTURE WORK

The present system already chosen some of the features, but we can try with remaining features like voice quality and speaking rate etc. Although the current MFCC already given very good emotion recognition performance, a further exploitation may contribute to the development of even more powerful features. Moreover, the features will be tested under more complex real-world conditions (e.g. reverberant and noisy speech).
To simulate emotion detection in cars, direct interfacing between MATLAB and the TMS board is required. TMS board is a micro-controller which is embedded with a pipelined digital signal processor for real-time signal processing.

REFERENCES

[1]. Christian Martyn Jones and Ing-Marie Jonsson. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In OZCHI '05: Proceedings of the 17th Australia conference on Computer-Human Interaction, pages 1-10, Narrabundah, Australia, Australia, 2005. Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

[2]. B. Schuller et al. Effects of in-car noise-conditions on the recognition of emotion within speech. In Proc. DAGA, 2007.
[3]. Clifford Nass, Ing-Marie Jonsson, Helen Harris, and Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. Improving automotive safety by pairing driver emotion and car voice emotion. In CHI '05: CHI '05 extended abstracts on Human factors in computing systems, pages 1973-1976, New York, NY, USA, 2005. ACM.
[4]. J. Healey and R. Picard. Smartcar: detecting driver stress. Volume 4, pages 218 -221 vol.4, 2000.
[5]. J. G. Taylor, K. Scherer, and R. Cowie. Introduction: 'emotion and brain:Understanding emotions and modelling their recognition'. Neural Netw., 18(4):313-316, 2005.
[6]. Sommer D. Holzbrecher M. Golz, M. and T. Schnupp. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In RS4C (Eds.), Proceedings 14th International Conference Road Safety on Four Continents, Bangkok, Thailand, 2007.
[7]. Batliner A. Hönig, F. and E. Nöth. Fast recursive data-driven multi-resolution feature extraction for physiological signal classification. In In J. Hornegger, et al. (Eds.): 3rd Russian-Bavarian Conference on Biomedical Engineering, pages 47-52, Erlangen, 2007.
[8]. Robert Horlings, Dragos Datcu, and Leon J. M. Rothkrantz. Emotion recognition using brain activity. In CompSysTech '08: Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, pages II.1-1, New York, NY, USA, 2008. ACM.
[9]. Martin Golz, David Sommer, Mo Chen, Udo Trutschel, and Danilo Mandic. Fusion of state space and frequency domain features for improved microsleep detection. In W. Dutch et al. (Eds.), Proceedings International Conference Artificial Neural Networks (ICANN 2005), pages 753-7592, 2005.
[10]. Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. Combining efforts for improving automatic classification of emotional user states. In Proc. IS-LTC 2006, Ljubliana, pages 240-245, 2006.
[11]. P Juslin and K Scherer. Vocal expression of affect. The New Handbook of Methods in Nonverbal Behavior Research, January 2005.
[12]. M. J. Owren and J.-A Bachorowski. Measuring emotion-related vocal acoustics. In J. Coan and J. Allen (Eds.). Handbook of emotion elicitation and assessment, pages 239-266. New York: Oxford University Press, 2007.
[13]. Seppi D. Steidl S. Vogt T. Wagner J. Devillers L. Vidrascu L. Amir N. Kessous L. Schuller B., Batliner A. and Aharonson V. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In Proceedings of Interspeech, pages 2253-2256, 2007.
[14]. R. Van Bezooyen. Characteristics and Recognizability of Vocal Expressions of Emotion. Foris Pubns, USA, June 1984.
[15]. Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9):1162 -1181, 2006.
[16]. T. Kanda, K. Iwase, M. Shiomi, and H. Ishiguro. A tension-moderating mechanism for promoting speech-based human-robot interaction. Pages 511-516, aug. 2005.
[17]. Englert R. Stegmann J. Burleson W Burkhardt F., Ajmera J. Detecting anger in automated voice portal dialogs. In Proceedings of Interspeech, Pittsburgh, 2006.
[18]. Van Ballegooy M. Englet R. Huber R Burkhart, F. An emotion aware voice portal. In Proc. Electronic Speech Signal Processing ESSP, 2005.
[19]. Laurence Devillers and Laurence Vidrascu. Real-life emotion detection with lexical and paralinguistic cues on Human-Human call center dialogs. Proc. INTERSPEECH' 06. Pittsburgh, 2006.
[20]. Hua Ai, Diane J. Litman, Kate Forbes-riley, Mihai Rotaru, Joel Tetreault, and Amruta Pur. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In Proceedings of Interspeech, pages 797-800, 2006.
[21]. J I. Murray, Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2):1097-1108, 1993.
[22]. Thurid Vogt, Elisabeth Andre, and Johannes Wagner. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. Pages 75-91, 2008.
[23]. O. M. E. Wahlstrom, N. Papanikolopoulos, "Vision-based methods for driver monitoring," in Proceedings of the 6th IEEE International Conference on Intelligent Transportation Systems, Shanghai, China, 2003, pp. 903-908.

[24]. A. E. M. C. Esra Vural, Gwen Littlewort, Marian Bartlett and Javier Movellan, "Drowsy Driver Detection through Facial Movement Analysis" Springer Berlin / Heidelberg, vol. 4796, pp. 6-18, 2007.
[25]. H. D. Vankayalapati and K. Kyamakya, "Nonlinear Feature Extraction Approaches for Scalable Face Recognition Applications," ISAST transactions on computers and intelligent systems, vol. 2, 2009.
[26]. R. V. Bezooijen, "The Characteristics and Recognisability of Vocal Expression of Emotions," Foris, Drodrecht, the Netherlands, 1984.
[27]. J. W. E. A. e. Thurid Vogt, "Automatic Recognition of Emotions from Speech: a Review of the Literature and Recommendations for Practical Realisation," in Affect and Emotion in Human-Computer Interaction: From Theory to Applications Berlin, 2008.
[28]. J. A. I. Murray, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," Journal of the Acoustical Society of America, vol. 93 (2), pp. 1097–1108, 1993.
[29]. C. K. D. Ververidis, "Emotional speech recognition: Resources, features, and methods," presented at the Speech Communication, 2006.
[30]. L. G. Yongjin Wang, "An investigation of speech-based human emotion recognition," pp. 15 - 18, 2004.
[31]. M. R. A. Paeschke, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech," Proc. Interspeech, 2005.
[32]. T. Barbu, "Discrete speech recognition using a hausdorff based metric," in Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks, ICETE, Setubal, Portugal, 2004, pp. 363-368.
[33]. T. Barbu, "Speech-dependent voice recognition system using a nonlinear metric," International Journal of Applied Mathematics, vol. 18, pp. 501-514, 2005.
[34]. James H. Jurafsky, Daniel Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Ser (2ND), May 2008.
[35]. Tobias Andersson. Audio classification and content description. Master's thesis, Lulea University of Technology, Multimedia Technology, Ericsson Research, Corporate unit, Lulea, Sweden, March 2004.
[36]. Abdillahi Hussein Omar. Audio segmentation and classification. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, February 2005.
[37]. M. Rolfes W. Sendlmeier F. Burkhardt, A. Paeschke and B. Weiss. Berlin database of emotional speech on-line. In Interspeech: http://pascal.kgw.tu-berlin.de/emodb/index-1024.html, pages 1517-1520, 2005.

Author

Sandeep Kotte Graduated in Information technology from JNTU in 2007 and M.S in Information technology from the University of Klagenfurt, Austria in 2010 specialized in Intelligent Transportation System, pervasive computing and Business Informatics. Currently working as Assistant professor in Dhanekula Institute of Engineering & Technology, India.