Koç University

Masters Thesis

Affect Burst Detection and Recognition

Author:
Shabbir Marzban

Supervisor:
Dr. Yucel Yemez

A thesis submitted in fulfilment of the requirements
for the degree of Master of Science
in the

Computer Engineering
Graduate School of Science and Engineering

August 2015

Koc University

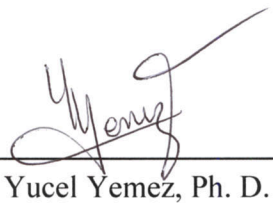
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Shabbir Marzban

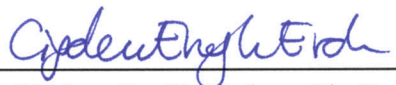
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:



Yucel Yemez, Ph. D. (Advisor)

Engin Erzin, Ph. D.



Çiğdem Eroğlu Erdem, Ph. D.

Date: _____

To my mom, dad and wife ...

KOÇ UNIVERSITY

Abstract

Graduate School of Science and Engineering

Master of Science

Affect Burst Detection and Recognition

by Shabbir Marzban

Affect bursts play a critical role in recognizing underlying emotions and hence their detection is of particular interest to researchers. In the literature, there are a number of approaches for detecting affect bursts, particularly laughters, from audio-only inputs, but multimodal approaches are limited while they can provide more effective systems for the affect burst detection. This thesis presents the current state of the art as well as our approaches to detect affect bursts along with their type. This includes a two layer hierarchical classifier, first to detect affect burst events and then to classify these events into affect burst types. In the hierarchical classifier we combine cues from audio and facial landmark points. In experimental evaluations we use the *Interactive emotional dyadic motion capture database* (IEMOCAP), which contains realistic and natural dyadic conversations with high quality audio and normalized motion capture of facial landmark points. We firstly annotate affect bursts events in the IEMOCAP database and test our proposed approach over it. We observe significant performance improvements with the multimodal approach over audio-only and visual-only unimodal schemes in detecting affect bursts along with their types. We also show comparison among different types of classifiers which serve as baseline.

Acknowledgements

I would like to express my sincere gratitude and profound respect to my advisors and mentors, Dr. Yucel Yemez, Dr. Engin Erzin and Dr. Metin Sezgin for giving me an opportunity to research at a world-class institution in Turkey and for their timely guidance, inspiration and support through out the two years of Masters program. I would also like to thank Dr. Cigdem Eroglu Erdem for serving on my thesis defense committee.

I would also like to thank my parents, Joozer and Mehrunnisa without whom none of this would have been possible and also to my wife, Tasneem, for constant support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Talking to Computers

As the field of computers science emerged, the only way to communicate with computers was through passing messages which was greatly limiting. This triggered a need of developing a communication system which would enable natural communication with them. To lay the foundations, understanding human communication is the first step, which is quite complex even though babies seem to have cracked it as they grow older. Scientists and researchers have been trying for years to make computers understand and respond to spoken language and have natural human level conversations. Current state of the art allows computers to at least recognize and respond, up to a certain accuracy, to what is being said. Good examples of such technology are Siri from Apple, Google voice, and Cortana from Microsoft. There is still a long way to go when it comes fully interpreting the spoken language. Posible solutions are due with recent advancements in artificial intelligence and neural nets like technology

One often overlooked aspect of language understanding is its emotional content. Humans, while communicating, not only use linguistic capabilities but also, among other things, use emotional cues such as laughters or sighs. Computers systems need to evolve to capture these cues so as to have complete spectrum human

communication covered. This thesis presents work this very direction, particularly on affect bursts, defined and described in Section 1.2, detection and recognition. We also present state of the art as well as our approaches in detail.

1.2 Affects bursts

1.2.1 Definition

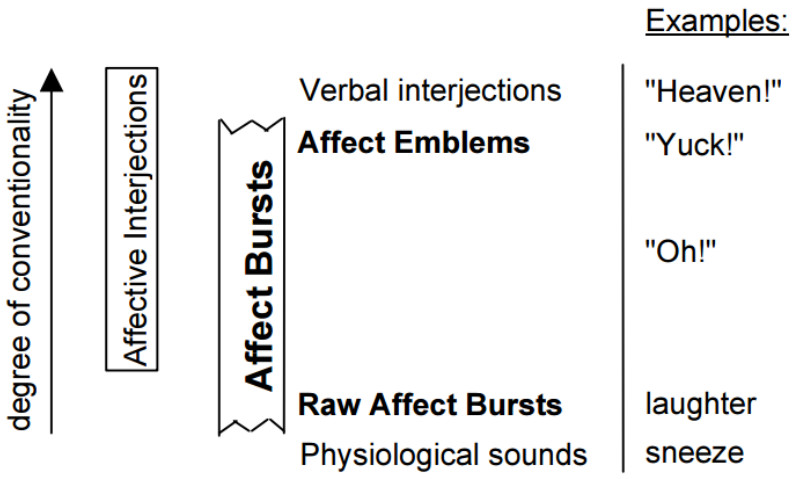


FIGURE 1.1: Illustration of domain of affect bursts by Schroder[1]

'Affect bursts' term was first coined by Scherer[2] who described it as brief, discrete, non verbal expression on both face and voice triggered by events. He proceeds by with giving example of facial/vocal disgust human exhibit upon sudden appearance of a worm. Schroder[1] in the later works tried to put this definition in comparison to other related words and phrases used in the literature and defined its domain. Affect bursts, in his opinion, is general term with a broad spectrum that has 'raw affect bursts' on one end while 'affect emblems' on other. 'Raw affect bursts' result from psychological factors such as pain. These are universal in nature but have relatively higher variability among expressions from different subjects. 'Affect emblems' on the other hand, are culturally expected expressions for the certain events, like showing remorse when found guilty. These are less universal in nature

and show less variability in expression from subjects coming from relative similar background. Figure 1.1 from Schroder's research gives graphical illustration defining the domain of 'affect bursts'.

Furthermore, Schroder[1] connects the affect bursts with underlying emotions. Table 1.1 outlines basic emotional states and their related affect bursts. He conducted a study in German language over these categories of emotions with conclusion that, with more than 80% accuracy, emotions mentioned in the Table 1.1 can be recognized by affect bursts except for that of anger.

Following his work, Belin et al. [3] curated a set of 90 affect bursts from 10 actors corresponding to 8 basic emotions, namely anger, fear disgust, fear, pain, sadness, surprise, happiness and pleasure. Over this database, they conducted a study on recognizing underlying emotions on 30 participants which revealed fairly high recognition rates with mean of 68%. These studies show that emotional content can be effectively gauged by recognizing affect bursts.

TABLE 1.1: Emotional states and the related affect bursts as outlined by Schroder[1]

Emotion State	**Related Affect Burst**
Admiration	Wow, Boah
Threat	Hey, Growl
Disgust	Buah, Igitt, Ih
Elation	Ja, Yippie, Hurra
Boredom	Yawn, Sigh, Hmm
Startle	Int. breath, Ah
Worry	Oje, Oh-oh, Oweh, Hmm
Relief	Sigh, Uff, Puh

1.2.2 Patterns and multi-modality

The synchronized multimodal patterns of affect bursts, especially facial landmark points and voice, make it distinguishable from patterns observed in speech and silence segments in a given dyadic conversation. Out of the types of affect bursts, breathing (which includes growl, sigh, hmm and puh) and laughter are of particular

interest as they convey an emotional response to affective events in day to day conversations.

1.3 Organization

The remainder of this thesis is structured as follows. In Chapter 2, a detailed review of the literature related to affect bursts and the field of affect recognition is presented. In Chapter 3, we present our novel method in detail, a two layer hierarchical classifier, which first detect affect burst events and classifies these events into affect burst types. In Chapter 4, we introduce the IEMOCAP[4] dataset and present our experiments and evaluations.

Chapter 2

Literature Review

In the literature, there are very few works broadly covering affect burst recognition or detection. However, there is plenty work done on detection specific types of affect bursts. There is also a recent push in using multi-modal cues for building such systems. For example, there are some works which are mainly geared towards detection of laughters using audio only [5, 6] or using multi-modal cue [7–9]. Other works are usually motivated by reasons different from the ones we are pursuing. For example, since laughter, which we categorize as one of the affect burst types, is perceived as noise while recognizing speech. Hence, its detection and segmentation, for purpose of discarding them, has been throughly studied[10]. There is also a plethora work done on affect recognition which, although is an entirely different domain, it helps us to identify the history of social signal processing and provides us with valuable insights for affect bursts recognition and detection.

In the following sections, we have grouped together related works in the field of affect recognition, followed by seminal works on affect burst detection in the literature.

2.1 Affect Recognition

To describe affect recognition in historical context Ekman [11, 12] conducted studies concluding that basic emotions such as happiness, sadness, fear, anger, disgust and surprise with respect to facial expressions are universal and hence regardless of culture. This influenced lot of work geared towards recognition among certain categories. Among these, Cohen et al. [13], Kaliouby & Robinson [14] and Lee & Elgammal [15] presented interesting works as they used facial features to classify affects. The features they used were action units, facial landmark points and pixel intensity of face region respectively with decent results.

Although such categorization of emotions is advantageous as they describe very common human experiences, they fail to completely describe range of emotions displayed naturally. To solve this problem, an alternate dimensional description of affect was proposed over which many affect recognition approaches are present in the literature. Most common of dimensional description of affect is Arousal-Valence space also know as Russells two dimensional model [16]. Figure 2.1 shows illustration of affect on evaluation activation space.

Affect recognition works which involve multi-modality are of particular interest to us as they give us insights in features and methods for recognition commonly used and also in strategies employed to synchronize and fuse different modalities. One fo the interesting works is done by Busso et al. [17] where they used Support Vector Machines (SVMs) with prosody and marker positions as features to classify sentences into four affect categories. Similarly, Karpouzis et al. [18] used Recurrent Neural Networks (RNNs) with facial points and prosody as features to classify tunes into four affect categories.

The most common data fusion techniques used in such works are namely feature-level and decision-level. Since human display audiovisual emotions in redundant manner in both modalities, decision-level fusion results in loss of co-relational information because it inherently assumes condition independence among the modalities. There are some works leveraging model-fusion to combined multiple

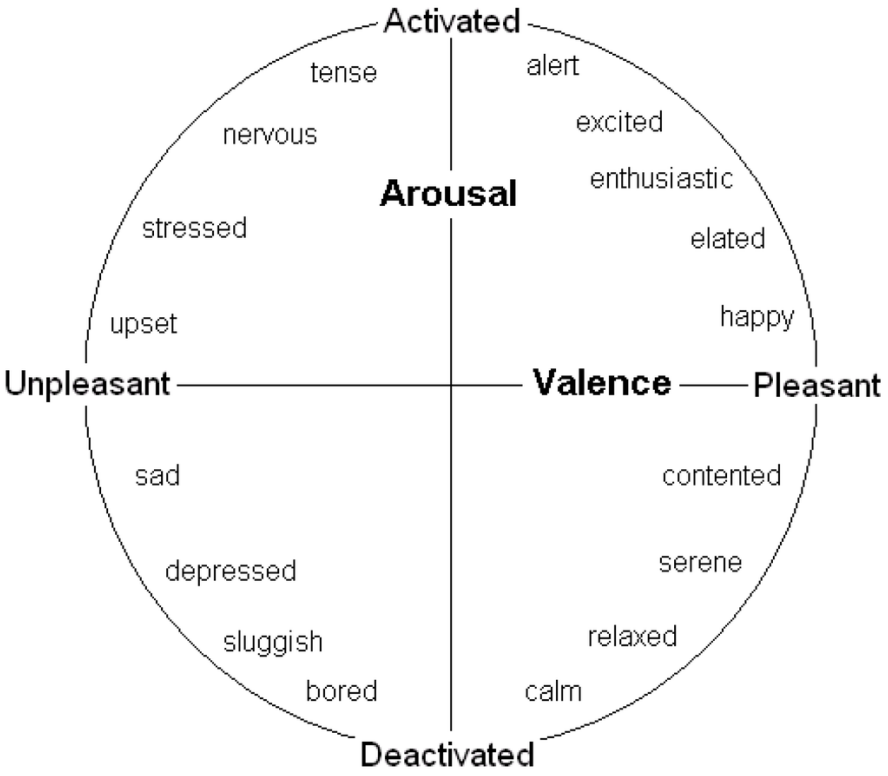


FIGURE 2.1: Illustration of Arousal-Valence space by Russel[16]

streams of data removing the need of synchronization among streams of data. Zeng et al. [19, 20] proposed multiple approaches with complex Hidden Markov Models (HMM) employing model level fusion to combine audio and video channels.

We would like to highlight here that our work is very different from affect recognition which is an entirely separate domain. The survey by Zeng et al. [21] highlights most of such works. Our work proposes a method for detecting 'affect bursts' which are identifiable events as opposed to affect recognition which can be described at all times indicating the emotional state of the person of interest.

2.2 Seminal Works

One preliminary work in comparison to our problem is by Petridis and Pantic [7]. They have proposed method for discrimination between pre-segmented laughters and speech using both audio and video streams. Extracted features are, facial

expressions and head pose from video, and cepstral and prosodic features from audio. They have also showed that the decision-level fusion of the modalities outperformed audio only and video only classifications. Recently Petardis et al. [9] have proposed a method for detection of laughters using time delay neural networks (TDNNs). They have explained their relatively lower values for precision, recall and F_1 score by the presence of unbalanced data, where a typical stream would have way more non-laughter frames than laughter ones. In another recent work, Turker et al. proposed a method for recognition between pre-segmented types of affect bursts, namely, laughter and breathing using HMMs [22]. Although this method is promising, it could not be directly used for detecting affect bursts over streams of audio and video. To distinguish pre-segmented non-verbal vocalizations using audio only features, solutions employing different learning methods have been proposed. Schuller et al. [23] used different classifiers on mel-frequency components to differentiate non-verbal vocalizations (laughter, breathing, hesitation, and consent). Among various classifiers they used, hidden Markov models (HMMs) outperformed other classifiers such as hidden conditional random fields (HCRFs) and support vector machines (SVM).

Relating to multimodal works in affect burst recognition, Cosker and Edge [24] present analysis of correlation between voice and facial marker points using HMMs in four non-speech articulations, namely laughing, crying, sneezing and yawning. Although their work is geared towards synthesis of facial movements using sound, it provides useful insights into affect burst recognition as well. A recent study done by Krumhuber and Scherer [25] shows that the facial action units, coded using Facial Action Coding System (FACS), exhibit significant variations for different affect bursts and hence can serve as cues in detecting and recognizing affect bursts. Recently Griffin et al. [26] highlighted the importance of body gestures as one of the important cues in recognizing affect bursts. This work, however, discriminates across only five pre-defined types of pre-segmented laughter.

Pammi et al. [27] used Automatic Language Independent Speech Processing (ALISP) models pretrained on huge database to detect laughters. ALISP models consist of sound units which are trained on grouped similar sounding segments

in an unsupervised manner. After adapting these models to laughter and non-laughter regions, they used Viterbi decoding to find likelihood of a given unseen segment being a laughter or a non-laughter. They also showed that ALISP method performed better than traditional methods such as Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM).

TABLE 2.1: Submissions in INTERSPEECH challenge [28]

Paper	**Classification Method**	**Dataset**	**Features**	**Comments**
Krikke and Truong [29]	GMM	InterSpeech Challenge SSPNet	MFCCS, Pitch, Position of first and second forments.	Comparable performance, not better than baseline
An et al. [30]	Probabilistic rescoring and thresholding	InterSpeech Challenge SSPNet	Syllable based features.	UAAUC: 84.85%
Oh et al. [31]	SVM	InterSpeech Challenge SSPNet	Segmented clip into median energy envelope followed by feature generation.	UAAUC: 85.3%
Wagner et al. [32]	SVM	InterSpeech Challenge SSPNet	Phonetic Sequence features.	UAAUC: 87.7%
Kaya et al. [33]	Decision Forest	InterSpeech Challenge SSPNet	MFCCs.	UAAUC: 88.4%
Gupta et al. [34]	DNN	InterSpeech Challenge SSPNet	MFCCs.	UAAUC: 91.5 %

Recently in InterSpeech 2013 conference, a challenge to detect laughters fillers and garbage was announced over a provided audio only database, SSPNet Visualization Corpus [28]. This database was labeled into laughters fillers and garbage for each frame. This is quite interesting to us since laughter is one the affect bursts and upon inspection, fillers contained other types of affect bursts we are interested in. Table 2.1 shows some of the interesting submissions to this challenge. Schuller et al. [28] presented a simplistic approach of training using Support Vector Machines (SVMs) with Area Under the Receiver Operator Characteristic Curve (AUC-ROC) of 82.9% for laughters and 83.6% for fillers with Unweighted Average Area Under the ROC Curve (UAAUC) of 83.3% setting benchmark for the challenge.

Interesting submissions to this challenge are presented in the Table 2.1. Kirkke and Truong [29] used GMM to train over MFCC and other basic acoustic features. Oh et al. [31] formulated features on syllable level. Syllables were segmented using

local minimums of energy over stream of audio. Acoustic features were computer over these segments and fed to a trained SVM classifier. They managed to produce better results than baseline with UAAUC of 85.3%, Gupta et al. [34] used same features as baseline but instead of SVM they employed Deep Neural Net resulting in highest performance in the challenge with UAAUC of 91.5%. Wagner et al. [32] used Sphinx Toolbox [28] to the produce phonetic sequence over audio stream. Using distribution of these phonetic sequence over a window as features there were able to get decent of results with UAAUC of 87.7%.

2.3 Contributions of the Thesis

This thesis builds upon our preliminary work presented in Turker et al. [22] where we explore the basic differentiability an=mong two types of affect bursts namely laughter and breathing. This thesis also contains comparison with baseline method of affect burst detection method on raw multi-modal features presented in Turker et al. [35]. The key contributions of this thesis is as follows:

- In order to build a robust affect bursts detection system, a high quality data is needed. For this purpose, we have used IEMOCAP [4] dataset which contains rich dyadic conversations containing improvised and scripted scenarios. This dataet has high quality sound and facial motion capture points making it ideal for multi-modal use. Our main contribution is the annotation of affect burst events along with their type in this dataset of dataset.

- We present a novel solution to detect affect bursts along with their types over audio and visual streams using statistical features. We also draw intuitions from previous works while extracting the best features to represent audio and visual data.

Chapter 3

Affect Bursts Detection and Recognition

3.1 General Method

The objective of our method is to detect affect bursts along with their types over audio and visual streams. This is illustrated in Figure 3.1. First, frame-based audiovisual features are extracted; from audio stream, features are computed in terms of mel-frequency cepstral coefficients (MFCCs) whereas the visual stream comprises of 3D facial points from which visual features, in the form of PCA coefficients, are extracted. Then, sliding windows over these audio and visual features are formed, with window of size 750 milliseconds and shift of 250 milliseconds. Illustration of sliding window over stream of audio or video can be seen in Figure 3.2. Statistical features, over the set of previously computed audio and visual features belonging to a given window, are calculated and fed to a classification engine. Statistical feature extraction is explained in detail in Section 3.1.1. The classification engine then produces a stream of labels indicating occurrences of affect bursts along with their types. The same scheme is used both to train and test the system. The training is performed over pre-segmented affect

bursts annotated on a dataset of audiovisual streams. In the following sections, the feature extraction procedure and the classification engine are described in detail.

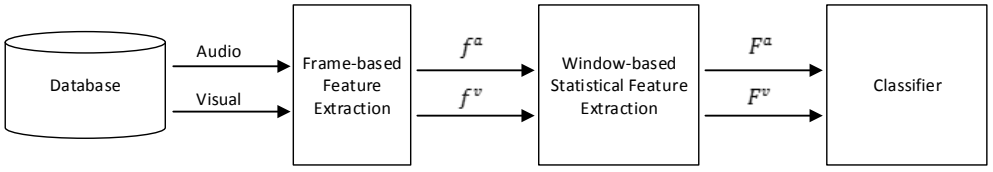


FIGURE 3.1: The block diagram of the proposed method for detection and classification of affect bursts. Frame based features are extracted from audio-visual streams over which windowing is performed, and then statistical features, as explained in Section 3.1.1, are calculated and fed to a classification engine. This scheme is employed both to train and test the classifier.

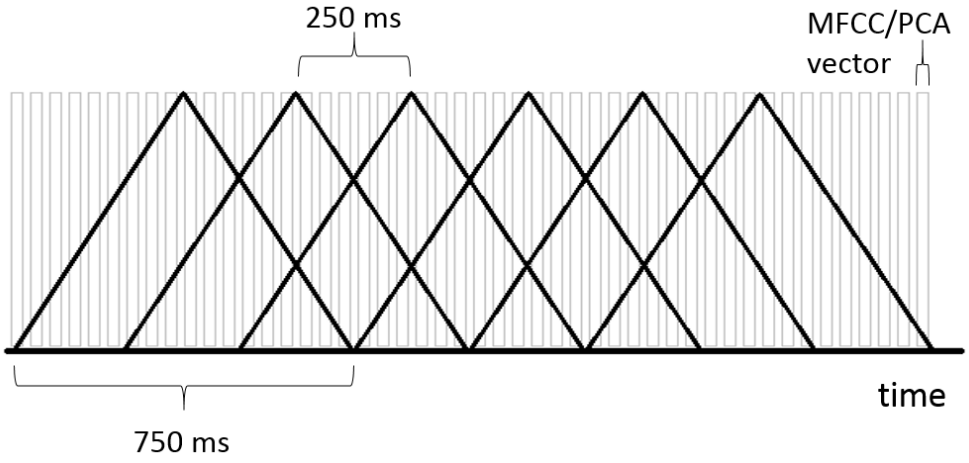


FIGURE 3.2: Illustration of sliding windows over audio and visual features, with window of size 750 milliseconds and shift of 250 milliseconds. Statistical features, over the set of previously computed audio and visual features belonging to a given window, are calculated and fed to a classification engine.

3.1.1 Feature extraction

3.1.1.1 Audio features

MFCCs are well known and successful descriptors for speech and audio processing, and have been shown to have good performance in characterizing affect bursts [23]. We compute MFCCs at every 10 milliseconds with 25 milliseconds window on the audio stream. The concatenation of MFCCs with their first order derivatives gives us the audio feature vector for every frame, that we represent by f^a (see Fig. 3.1).

We then create sliding windows of duration 750 milliseconds, with shift of 250 milliseconds, on these features over which statistical features are calculated. This windowing approach also allows us to implicitly handle segmentation of affect bursts, hence enabling us to perform detection on a stream of input.

Given the observations of f^a over a specific window, statistical features comprising of functionals, which are specifically mean, standard deviation, skewness, kurtosis, range, min, max, first quantile, third quantile, median quantile, inter-quantile range, of the constituent elements of these feature observations are calculated. These statistical features describe short term distribution of the feature space. We represent the statistical features computed from observations of f^a by F^a. Usage of such functionals has been inspired by the work of Metallinou et al. [36], where they present promising results on tracking continuous levels of participants' activation, dominance and valence.

3.1.1.2 Visual features

Our visual data comprises of sequences of 3D facial landmark points captured by a MOCAP system. We first employ PCA (principal component analysis) to reduce the dimensionality of the initial data vector.

Let S represent the raw data vector input to the system for a given frame,

$$S = \begin{bmatrix} x_1 & y_1 & z_1 & ... & x_P & y_P & z_P \end{bmatrix}^T \tag{3.1}$$

where P is the number of facial landmark points. Hence S is formed simply by concatenating x, y and z coordinates of the P facial landmark points for each frame. Using PCA, we determine the best 12 principal directions of the data. We then form the visual feature vector f^v, comprising of 12 PCA coefficients, by computing the projections of each data vector along the principal directions.

After the frame-based feature extraction, we extract statistical features comprising of functionals computed in the same fashion as that for audio features over the

same sliding windows. We represent the statistical feature vector for a given window by F^v (see Fig. 3.1).

3.1.2 Classification engine

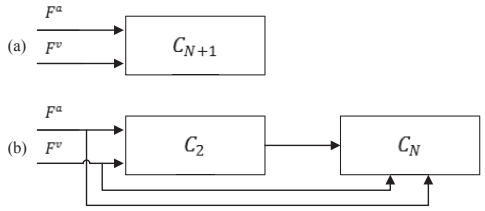


FIGURE 3.3: Classifier configurations. (a) Single layer classifier C_{N+1} which can distinguish between N types of affect bursts and a reject class. (b) Hierarchical classifier, where the first classifier C_2 is a binary classifier, followed by N-class classifier, C_N, to distinguish N affect burst types.

In order to describe classifiers with different configurations, we use the following notation throughout the paper. C_c^m denotes a generic classifier. The superscript m denotes feature type on which classifier is trained and expects as input; $m \epsilon \{a, v, av\}$ where a, v and av denote audio features, video features and feature fusion (of audio and visual features) respectively. The subscript c denotes the number of classes for which the classifier is trained.

Figure 3.3 illustrates two different configurations of classifiers that can be used to perform detection/classification. The first one is a single layer classifier, C_{N+1}, which is used to distinguish between N different types of affect bursts along with a reject class. The second one is an hierarchical classifier denoted by $C_2 \cdot C_N$. This is a cascaded approach where a binary classifier, C_2, to distinguish between affect bursts and non affect-bursts (reject class), is used. The features of the data samples belonging to the affect bursts class are then fed to an N-class classifier, C_N, to discriminate between the N types of affect bursts. After the binary classification, a median filter can smooth out the detection boundaries. The filtered labels are then fed to an N-class classifier to discriminate between the types of affect bursts.

In order to achieve multi-modality, two approaches, namely feature fusion and decision fusion, can be employed. Feature fusion is relatively straightforward where

features from audio and visual streams are concatenated to form a feature set for a given data segment and are then fed to a classification engine. In our classifier notation, this is indicated by C^{av}. For the single layer classification scheme, this feature fusion is denoted by C^{av}_{N+1}. Similarly, for the hierarchical classifier scheme, it is denoted by $C^{av}_{2} \cdot C^{av}_{N}$.

In decision fusion, the probabilistic outputs of unimodal classifiers of audio and visual features are fused together to predict the label using a weighted average of likelihoods. Such a process yields likelihood estimates for each of the classes, denoted by L^{av}_{n}, and the highest one is selected as the label for a given data segment. This is described in the following equation:

$$L^{av}_{n} = \alpha L^{a}_{n} + (1 - \alpha) L^{v}_{n} \tag{3.2}$$

where α is the weighting coefficient optimized by cross validation while training and, L^{a}_{n} and L^{v}_{n} are the likelihoods of the nth class (in the [0,1] interval) for audio and video classifiers, respectively.

We represent the decision fusion operation by $\oplus$ symbol. A single layer classifier with decision fusion can be written as $(C^{a}_{N+1} \oplus C^{v}_{N+1})$, where outputs of audio and video classifiers are fused together. Similarly, the hierarchical approach can be written as $(C^{a}_{2} \oplus C^{v}_{2}) \cdot (C^{a}_{N} \oplus C^{v}_{N})$, where decision fusion of audio and video classifiers takes place in both stages. Different combinations between the above mentioned classifiers and approaches are thoroughly tested and compared. These are discussed in the following section.

Chapter 4

Results and Discussion

4.1 Dataset

In our experimental evaluations, we used IEMOCAP database [4] to train and test our system. It is a multi-speaker and multimodal database of number of conversations. It has audio-visual data which includes video, speech, motion capture of face and text transcriptions. Configuration of motion capture points is illustrated in Figure 4.1. It has rich emotional reactions which are comprised of sporadic affect bursts. Apart from actors being diverse in the database, in half of the recorded sessions, they are required to improvise or only the premise of the scripts are provided and hence the affect bursts occurrence is mostly natural. This gives us an edge compared to previous methods in terms of naturalness.

To segment the affect bursts for training purposes, three annotators segmented and labeled two types of affect bursts, namely laughters and breathing, on four sessions of IEMOCAP database which contain 105 clips of conversations between pair of four people. Only the agreed upon labels were then filtered and used as samples for training and testing. The segmented affect bursts were further fine-tuned by manipulating the starting ending times based on audio/video visualizations. In total, our training set contains 301 affect bursts with 182 segments of 'breathing' type and 119 segments of 'laughter' type. The average length of affect bursts is

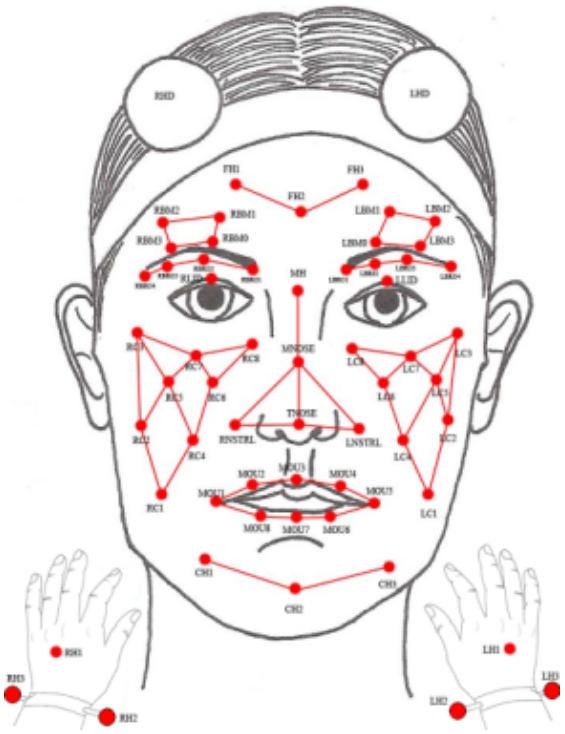


FIGURE 4.1: Configuration of motion capture points in IEMOCAP[4]

1.3 seconds. This distributions of each type is shown in Figures 4.2 and 4.3. Two sample segments of each of the types of the affect burst segments are illustrated in Figures 4.4 and 4.5. These annotations will be made available in the form of a database for researchers' use.

To train the classification system, statistical audio and video features of the sliding windows are extracted from the pre-segmented affect bursts segments. A single affect burst segment may yield multiple data samples depending on the number of sliding windows that span it. Since number of annotated affect bursts are usually limited in the database, sliding windows allow us to have varying data sample versions representing the same affect burst segment. The labels corresponding to these data samples are also produced. In order to introduce reject class, similar operation is done on randomly segmented regions that do not belong to affect burst segments. In total, 872 data samples are extracted from the affect burst segments. For the purposes of introducing reject class for the training, our training set contains 143 such segments of 2 second duration which yield 715 data samples.

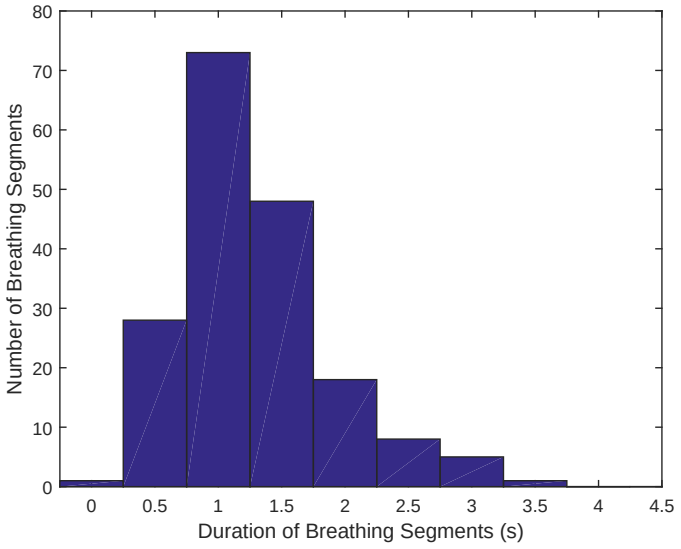


FIGURE 4.2: Histogram of breathing durations annotated in IEMOCAP[4]

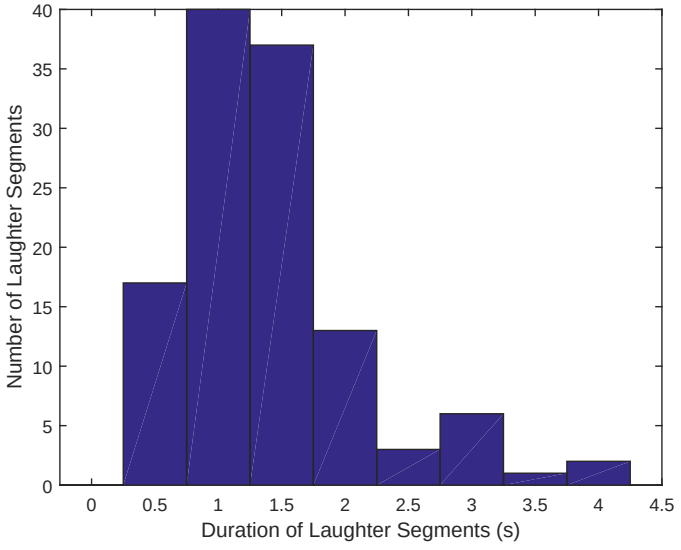


FIGURE 4.3: Histogram of laughter durations annotated in IEMOCAP[4]

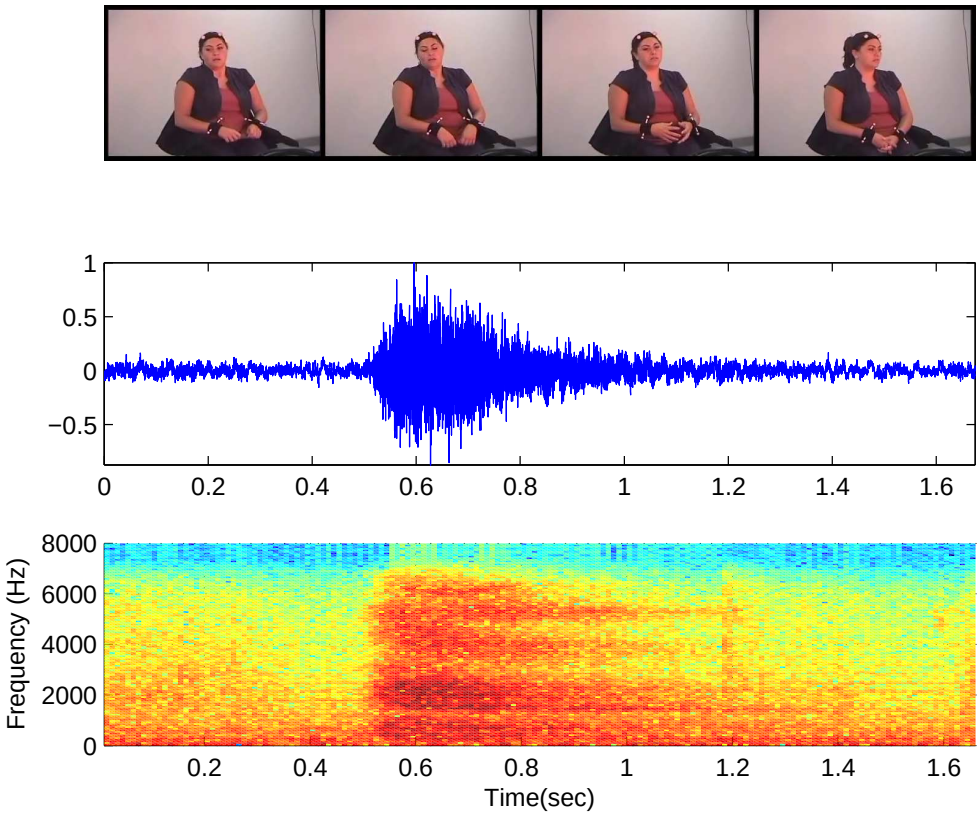


FIGURE 4.4: Sample video images and audio profile including signal plot and spectrogram of a breathing segment.

4.2 Support Vector Machines (SVMs)

In our experiments we employ support vector machines (SVMs) with radial basis function kernel for all the classifiers in different configurations using the implementation of Chang et al. [37]. We employ 10-fold cross validation technique for training and testing these classifiers. The folds are created by selection of segments rather than the data samples in order to avoid presence of data samples corresponding to the same segment in any particular instance of training and test folds. Figure 4.6 presents the confusion matrices of the experiments performed on our pre-segmented regions using various classifiers. In each of these confusion matrices, rows represent ground truth for laughter breathing and reject classes, whereas columns represent the predicted output. The first row of Figure 4.6 presents confusion matrices of single layer classifiers using audio features, visual features,

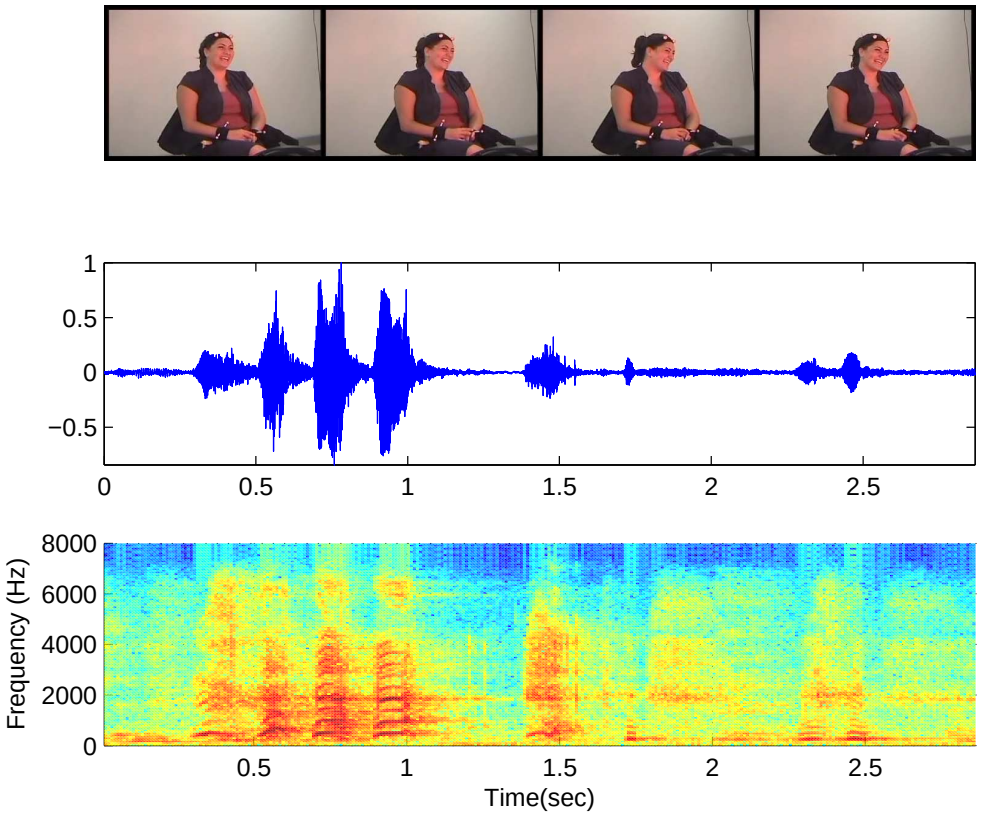


FIGURE 4.5: Sample video images and audio profile including signal plot and spectrogram of a laughter segment.

audio-visual features with feature fusion, and audio-visual features with decision fusion, respectively. Similarly, the second row presents confusion matrices of the hierarchical classifier configurations. We can observe that the unimodal classifiers present an interesting trend. Audio only classifiers perform better in recognizing breathing affect bursts than the visual only classifiers as the former ones show less confusion of the breathing class. This can be attributed to the characteristic of breathing affect burst, which has distinctive noise-like audio compared to other classes. Similarly, visual only classifiers perform better in recognizing laughter as it involves significant movement of facial marker points compared to other classes. The above mentioned factors lead to a significant increase in performance of multimodal classifier compared to unimodal ones.

Another observation that we make is that the decision fusion does worse than feature fusion in Figure 4.7. This can be explained as a limitation of SVM classifier

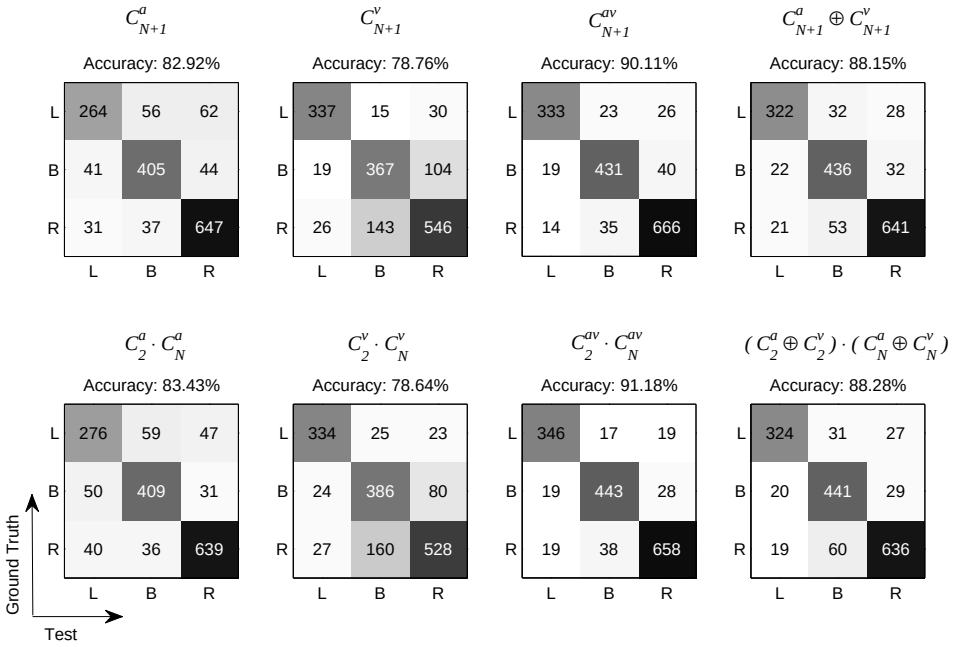


FIGURE 4.6: Confusion matrices of the test experiments performed over our pre-segmented regions using various classifier configurations. The first row presents confusion matrices of the single layer classifier using audio features, visual features, and audio-visual features with feature fusion, and decision fusion, respectively. The second row presents confusion matrices of the corresponding hierarchical classifiers. L, B and R denote laughter, breathing and reject classes, respectively.

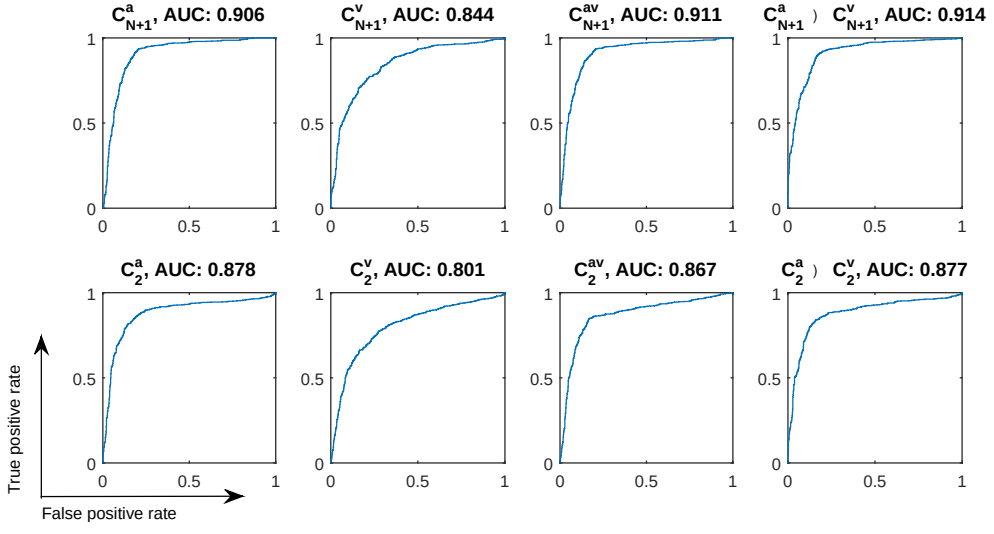


FIGURE 4.7: ROC curves between merged classes of affect bursts and Reject of the test experiments performed over our pre-segmented regions using various classifier configurations. The first row presents ROC curves of the single layer classifier using audio features, visual features, and audio-visual features with feature fusion, and decision fusion, respectively. The second row presents ROC curves of the first detector, C_2, of corresponding hierarchical classifiers.

whose probabilistic output is based on boundaries it learns rather than the actual distribution of data samples. Nonetheless, this warranted a further investigation where we analyzed affect bursts versus reject AUCROC values and ROC curves. For each of the classifier configurations, ROC curves with AUCROC values can be seen in Figure 4.7. Here we see an interesting trend that decision fusion in terms of AUC does better than feature fusion. In terms of performance, however, best performing classifier is $C_2^{av} \cdot C_N^{av}$ with highest accuracy of 91.18%.

TABLE 4.1: Confusion matrix of leave 10% of clips out cross validation experiment.

		Test		
		L	B	R
Ground Truth	L	525	29	115
	B	54	599	272
	R	4319	7850	101292

TABLE 4.2: Performance metrics of leave 10% of clips out cross validation experiment for each of the classes.

Labels	**Recall(%)**	**Precision(%)**	**F_1 score(%)**
Laughter	78.5	10.7	18.9
Breathing	64.8	7.1	12.7
Reject	89.3	99.6	94.2
UA	77.5	39.1	41.9

As an additional experiment, we selected the best performing classifier, $C_2^{av} \cdot C_N^{av}$ and performed a leave 10% of clips out cross validation experiment in order to check performance on stream of data rather than pre-segmented regions, which depicts a real world test case. In each fold, out of 105 clips, 10 clips were selected as test while rest were used for training. In the training fold, we used pre-segmented regions belonging to training clips to train the classifier, tested it with the entire clips of the test fold. In this experiment, the classes are very unbalanced with most of the data samples belonging to the reject class. Table 4.1 provides the confusion matrix of this experiment and Table 4.2 lists class-wise precision, recall and F_1 scores where the last row lists unweighted averages (UA) for each of these metrics. The relatively low scores of precision of affect burst types are due to high number of false positives. The high number of false positives can be explained by the

fact that the ground truth only contains agreed upon segments of the annotators, which results in conservative boundaries of the segments. While annotating, a large number of affect bursts segments, which had disagreement on type of the affect burst between the annotators or audio had interference from other speaker during dyadic conversation, were removed from the training set which also contributes to a higher number of false positives.

4.3 Gaussian Mixture Models and Hidden Markov Models

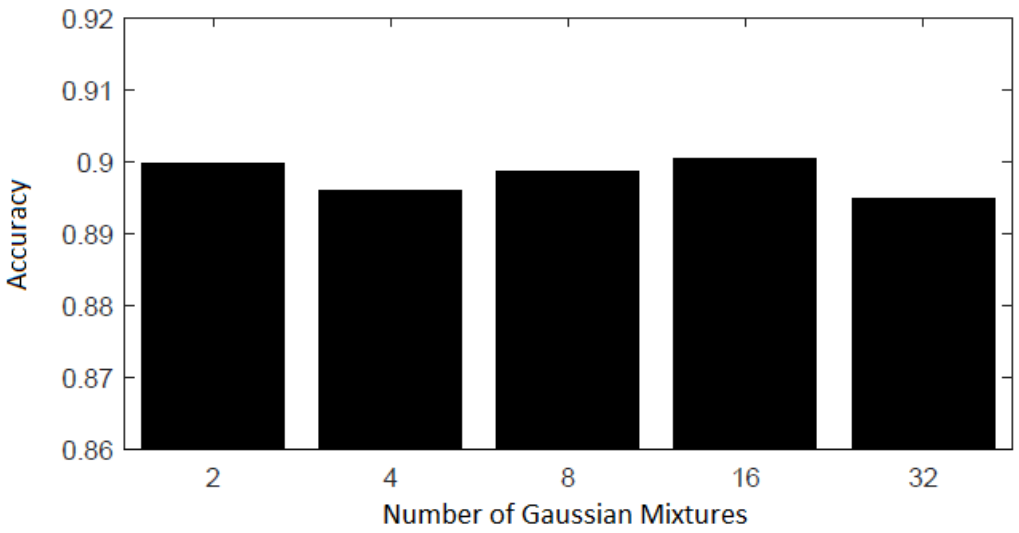


FIGURE 4.8: Performance of GMM classifier with vary number of Gaussian mixtures.

Experiments with GMM and HMM serve as baseline methods to compare SVM classifiers performances. One key difference between these schemes is that in SVM, statistical features are used where as here, raw features are used. Similar scheme of windowing is used as presented in Section 3.1.1. For audio and visual features however, instead of computing functionals, raw MFCCs features and PCA coefficients are used respectively. GMMs are trained on audio and video segments separately and fused together using decision fusion. In training, each frame of MFCCs and PCA constitutes as single sample as opposed to whole window in

TABLE 4.3: Performance of SVM and GMM classifier over pre-segmented regions

Classifier	**Configurations**	**Accuracy (%)**	**Recall (%)**	**Precision (%)**
SVM	C^{a}_{N+1}	82.92	L: 69.11 B: 82.65	L:78.57 B: 81.33
	C^{v}_{N+1}	78.76	L:88.22 B: 74.90	L: 88.22 B: 69.90
	C^{av}_{N+1}	90.11	L: 87.17 B: 87.96	L: 90.98 B: 90.98
	$C^{a}_{N+1} \oplus C^{v}_{N+1}$	88.15	L: 84.29 B: 88.98	L: 88.22 B: 83.69
	$C^{a}_{2} \cdot C^{a}_{N}$	83.43	L: 72.25 B: 83.47	L: 75.41 B: 81.15
	$C^{v}_{2} \cdot C^{v}_{N}$	78.64	L: 87.43 B: 78.78	L: 86.75 B: 67.60
	$C^{av}_{2} \cdot C^{av}_{N}$	91.18	L: 90.58 B: 90.41	L: 90.10 B: 88.96
	$(C^{a}_{2} \oplus C^{v}_{2}) \cdot (C^{a}_{N+1} \oplus C^{v}_{N+1})$	88.28	L: 84.82 B: 90.00	L: 89.26 B: 82.89
GMM	NoM: 2	89.98	L: 83.25 B: 91.43	L: 88.33 B: 86.65
	NoM: 4	89.60	L: 84.55 B: 92.65	L: 88.98 B: 84.54
	NoM: 8	89.86	L: 82.98 B: 92.45	L: 91.62 B: 84.51
	NoM: 16	90.04	L: 85.34 B: 91.22	L: 91.57 B: 86.46
	NoM: 32	89.48	L: 85.08 B: 88.37	L: 91.04 B: 87.47

exp: NoM stands for number of mixtures in GMM configurations.
L and B values in recall and precision calumnious are class wise recall and precision values of laughter and breathing classes respectively.

general method, however while testing, sum of log likelihood from frames is used to produce decision on a window. We employ a 10-fold cross validation technique to train Gaussian mixtures on pre-segmented regions similar to Section 4.2. We used Matlab's inbuilt function to train this classifier.

Figure 4.8 presents accuracies of classifier with varying number of mixtures, Table 4.3 gives a more detailed view, where accuracies, recall and precision values of GMM are compared with SVM counterpart. We see SVM with $C^{av}_{2} \cdot C^{av}_{N}$ is the best performing classifier among all the experiments. Although GMM model is theoretically more complex and can model the distribution of the vectors better, we believe its performance is inferior to SVM due loss of temporal information captured by the statistical functionals in SVMs.

As an additional experiment, we selected the best performing classifier, GMM with 16 mixtures, and performed a leave 10% of clips out cross validation experiment in order to check performance on stream of data rather than pre-segmented regions. In each fold, out of 105 clips, 10 clips were selected as test while rest were used for training. In the training fold, we used pre-segmented regions belonging to training clips to train the classifier, tested it with the entire clips of the test fold. Table 4.4 provides the confusion matrix of this experiment and Table 4.5 lists class-wise precision, recall and F_1 scores where the last row lists unweighted averages (UA) for each of these metrics. In comparison with SVM, it under performs in terms of F_1 scores.

TABLE 4.4: Confusion matrix of leave 10% of clips out cross validation experiment of GMM with 16 Gaussian mixtures.

		Test		
		L	B	R
Ground Truth	L	481	56	132
	B	51	663	211
	R	17050	23264	73147

TABLE 4.5: Performance metrics of leave 10% of clips out cross validation experiment for each of the classes of GMM with 16 Gaussian mixtures

Labels	**Recall(%)**	**Precision(%)**	**F_1 score(%)**
Laughter	71.90	2.74	2.64
Breathing	71.68	2.76	2.66
Reject	64.47	99.53	39.13
UA	69.35	35.01	14.81

In a similar experiment, we employed Hidden Markov Models (HMM) with Gaussian emissions to model the affect bursts and their types. These are more complex than GMMs and can capture the temporal behavior in the hidden states. The GMM can be considered as a specific case of HMM with single hidden states. Similar experiment with fully connected is conducted as in Section 4.2 for comparison purposes. The implementation of HMM used for training and testing purposes is HTK Tool set [38]. Figure 4.9 presents accuracies of HMM with varying number of states against number of mixtures. Since the training strategies are different, GMM uses each frame as separate sample where as HMM considers whole window

to single training sequence, results may not be comparable. Within lower number of Gaussian mixtures, trend of increase in performance, as number of states increase, supports our hypothesis that temporal information matters. However it can be argued that such increase may not be statistically significant. Overall, Using HMM does not yield a significant advantage in terms of performance.

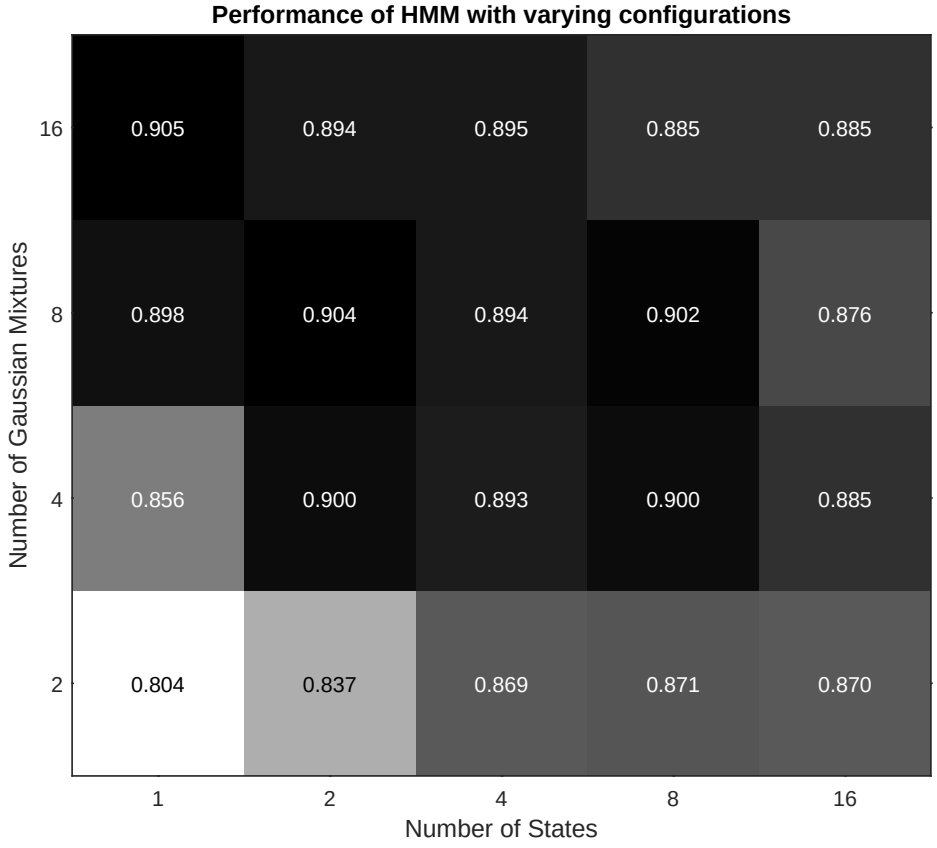


FIGURE 4.9: Performance of HMM classifier with varying number of Gaussian mixtures and number of states.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In this thesis, we firstly define a functional definition of affect bursts and present it as a novel way to detect underlying emotions based on evidences in psychology literature. We also provide an extensive review of the state of the art similar to our problem. These are mostly in the areas of affect detection and specific affect burst, laughter detection.

Building upon these fundamentals, we present a novel scheme of detecting affect bursts along with its type over stream of multi-modal inputs which we term as hierarchical affect burst classifier. We prove the robustness of our method by conducting various experiments. We have demonstrated that combining visual cues with audio produces a more effective affect burst classifier compared to unimodal schemes. We have also provided evidence that hierarchical configurations perform better than single layer configurations.

Furthermore we have provided in depth comparison of SVM classifiers on statistical features with different classification engines on raw features. We justify the use to statistical features over raw features by demonstrating superior performance of SVM classifier with statistical features compared to HMM and GMM methods where raw features are used.

5.2 Future Works

In future work, we plan testing our approach on a more diverse set of affect bursts, which we currently are unable to do, due to sparse appearance of these affect burst types in the database. We also wish to explore another promising modality, upper body gestures, which can lead to an increase in performance of affect burst detection. Also, investigating action units as visual features can provide valuable insights, although accurate automatic tracking of action units still remains a challenging problem.

5.2.1 Joker Dataset

For the purpose of have more diverse and frequent affect bursts samples we collected our own dataset using Kinect (v2) since it can provide stream of video along with depth (RGB-D), audio with direction, skeleton and high definition facial features which can enable us to explore multi-modal detection methods. An experiment was designed and executed based on Taboo game while recording all the modalities using Kinect (v2). This dataset was then further processed for annotation of affect bursts such as laughters and sigh gestures. Further more we recently collected data, in collaboration with European partners under the Joker Project [39], of social interaction dialogs involving humor between a human participant and a robot. In this work, many interaction scenarios have been designed in order to study social markers such as laughter and empathy. This dataset in currently being processed for annotation where three independent Once annotated, this we will test our current system one this dataset and set out work in two directions.

- Evaluation of our method in cross-dataset case where its trained on one dataset while being tested on the other. Evolve it be robust to such cases
- Inclusion of more modalities to improve our system such as body gestures and changes in pixel intensities around face.

Bibliography

[1] Marc Schröder. Experimental study of affect bursts. *Speech communication*, 40(1):99–116, 2003.

[2] K. R. Scherer. Affect bursts. *in Emotions, eds S.H.M.van Goozen, N.E.van de Poll, and J.A.Sergeant*, pages 161–193, 1994.

[3] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, 40(2):531–539, 2008.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[5] Khiet P Truong and David A Van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[6] Kornel Laskowski and Tanja Schultz. Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *Machine Learning for Multimodal Interaction*, pages 149–160. Springer, 2008.

[7] Stavros Petridis and Maja Pantic. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *Multimedia, IEEE Transactions on*, 13(2):216–234, 2011.

[8] Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. Decision-level fusion for audio-visual laughter detection. In *Machine Learning for Multimodal Interaction*, pages 137–148. Springer, 2008.

[9] Stavros Petridis, Maelle Leveque, and Maja Pantic. Audiovisual detection of laughter in human-machine interaction. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 129–134. IEEE, 2013.

[10] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. Noice robust real world spoken dialogue system using gmm based rejection of unintended inputs. 2004.

[11] Paul Ekman. Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284. University of Nebraska Press Lincoln, 1972.

[12] Paul Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. 1994.

[13] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1):160–187, 2003.

[14] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[15] Chan-Su Lee and Ahmed Elgammal. Facial expression analysis using nonlinear decomposable generative models. In *Analysis and Modelling of Faces and Gestures*, pages 17–31. Springer, 2005.

[16] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[17] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[18] Kostas Karpouzis, George Caridakis, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, Lori Malatesta, and Stefanos Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artifical intelligence for human computing*, pages 91–112. Springer, 2007.

[19] Zhihong Zeng, Yuxiao Hu, Ming Liu, Yun Fu, and Thomas S Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 65–68. ACM, 2006.

[20] Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson. Audio-visual affect recognition. *Multimedia, IEEE Transactions on*, 9(2):424–428, 2007.

[21] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[22] Bekir Berker Turker, Shabbir Marzban, Engin Erzin, Yucel Yemez, and Tevfik Metin Sezgin. Affect burst recognition using multi-modal cues. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 1608–1611. IEEE, 2014.

[23] Björn Schuller, Florian Eyben, and Gerhard Rigoll. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer, 2008.

[24] Darren Cosker and James Edge. Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. *Proceedings of Computer Animation and Social Agents, CASA*, 2009.

[25] E Krumhuber and KR Scherer. Affect bursts: Dynamic patterns of facial expression. *Emotion*, 11:825–841, 2011.

[26] Harry J Griffin, Min SH Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. Laughter type recognition from whole body motion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 349–355. IEEE, 2013.

[27] Sathish Pammi, Houssemeddine Khemiri, Dijana Petrovska-Delacrétaz, and Gérard Chollet. Detection of nonlinguistic vocalizations using alisp sequencing. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7557–7561. IEEE, 2013.

[28] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013.

[29] Teun F Krikke and Khiet P Truong. Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. 2013.

[30] Gouzhen An, David-Guy Brizan, and Andrew Rosenberg. Detecting laughter and filled pauses using syllable-based features. In *INTERSPEECH*, pages 178–181, 2013.

[31] Jieun Oh, Eunjoon Cho, and Malcolm Slaney. Characteristic contours of syllabic-level units in laughter. In *INTERSPEECH*, pages 158–162, 2013.

[32] Johannes Wagner, Florian Lingenfelser, and Elisabeth André. Using phonetic patterns for detecting social cues in natural conversations. In *INTERSPEECH*, pages 168–172, 2013.

[33] Heysem Kaya, Ali Mehdi Erçetin, Albert Ali Salah, and F Gürgen. Random forests for laughter detection. In *Proceedings of Workshop on Affective Social Speech Signals-in conjunction with the INTERSPEECH*, 2013.

[34] Rahul Gupta, Kartik Audhkhasi, Sungbok Lee, and Shrikanth Narayanan. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In *INTERSPEECH*, pages 173–177, 2013.

[35] B Berker Turker, Shabbir Marzban, M Tevfik Sezgin, Yucel Yemez, and Engin Erzin. Affect burst detection using multi-modal cues. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, pages 1006–1009. IEEE, 2015.

[36] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, 2013.

[37] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[38] S.J. Young and Sj Young. The htk hidden markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 1994.

[39] Laurence Devillers et al. Multimodal data collection of human-robot humorous interactions in the joker project. In *International Conference on Affective Computing and Intelligent Interaction (ACII2015), 2015 6th*. IEEE, 2015.