# Exploring Facial Expressions Using Cross And Multi Model Techniques For Unveiling Audio Visual Emotions

[I]Ranjidha.P, [II]Bharathalakshmi.R, [III]Kasthuri.S

[I,III]M.E CSE, Srinivasan Engineering College, Perambalur, Tamil Nadu, India.

[II]Assistant Professor, Dept. of CSE, Srinivasan Engineering College, Perambalur, Tamil Nadu, India.

E-mail : [I]ranjidha.it@gmail.com, [II]lakshmibharathi88@gmail.com, [III]kasthuri255@gmail.com

## Abstract

*Feelings of human are captured by the system by recognising the facial reaction and voice tone. To adapt with the situation, system gives motivation command to the user. Thus user can easily tolerate them. Affective computing is human-computer interaction in which a device has the ability to detect and appropriately respond to its user's emotions and other stimuli. Research on human emotional behaviour, and the development of automatic emotion recognition and animation systems, relies heavily on appropriate audio-visual databases of expressive human speech, language, gestures and postures. A computing device with this capacity could gather cues to user emotion from a variety of sources. Facial expressions, speech and body gesture the force or rhythm of key strokes and the temperature changes of the hand on a mouse can all signify changes in the user's emotional state, and these can all be detected and interpreted by a computer. A built-in camera captures images of the user and algorithms are used to process the data to yield meaningful information. Speech recognition and Face recognition are among the other technologies being explored for affective computing applications. Human computer intelligent interaction is an emerging field aimed at providing natural ways for humans to use computers as aids. These agents are sometimes just as an animated talking face, may be displaying simple facial expressions and, when using speech synthesis, with some kind of lip synchronization, and sometimes, with complex body movements and facial expressions.*

**Keywords**

*Facial expressions, Expressive human speech, Synchronization*

## I. Introduction

The interaction is enriched by speech recognition and generation system that allows a minimum instructional conversation with the agent or by an automatic emotion recognizer that transmits the user’s emotion to the agent which reacts accordingly. A webcam takes pictures of the user’s face. The aim of these pictures is to obtain additional information on the user and, in particular, on his or her emotional state. The features extraction program captures each facial frame and extracts the feature points which are sent to the emotion classifier. When an emotional change is detected, the output of the emotion classifier constitutes an emotion code. Emotional facial and tone of voice expressions combined with empathetic verbal behavior when displayed as feedback to students’ fear, sad, and happy emotions in the context of a self-assessment test. Extensive analysis and evaluation has been conducted using the newly released SEMAINE database of human-to-agent communication. . Our experimental results show that the proposed string-based approach is the best performing approach for automatic prediction of Valence and Expectation dimensions, and improves prediction performance for the other dimensions when combined with at least acoustic signal-based features.

The first goal of this study is to understand the emotional entrainment effect during spontaneous interactions present a thorough analysis using the interactive emotional dyadic motion capture (IEMOCAP) database. First, we study the co occurrence of the emotional states of the speakers and listeners. The result show that in 72 percent of the conversation turns the two subjects presented similar emotions. As a result, the expressive behaviours from one subject should be correlated with the behaviours of his/her conversation partner. To address this hypothesis, this study analyses cross-subject emotional entrainment using mutual information (MI).

## II. Related Work

An affect sensor is a device that takes an input signal and processes it for some evidence of emotions. There are a number of different techniques and modalities used to detect affect. These include: physiological signals, facial expression recognition, speech prosody recognition and pressure sensors.More specific subjects were called to assign in each image an emotional state among angry, neutral, sad, happy, disgusted, surprised and scared. Two kinds of problems have to be solved: facial expression feature extraction and facial expression classification. Our work focuses on the second problem: classification. This implies the definition of the set of categories and the implementation of the categorization mechanisms. The best matching decides the classification of the expression. Most of these methods first apply PCA algorithms to reduce dimensionality. The rule-based methods for classify the face expression into basic categories of emotions, according to a set of face actions previously codified.

### A. Detecting and Recognizing Emotional Information

During spontaneous conversation, individuals tend to externalize similar verbal and nonverbal behaviours to promote effective communication (i.e., synchronization). In communication sciences, this effect is referred to as the reciprocity pattern reported that during dyadic interactions the participants usually express similar nonverbal behaviours. They observed that users interacting with the robot made eye contact and imitated its gestures. They use these results to demonstrate the communication capabilities of the robot. These studies suggest that understanding the entrainment effect is important to improve the performance and efficiency of human machine interface.

Detecting emotional information begins with passive sensors which capture data about the user’s physical state or behaviour without interpreting the input. The data gathered is analogous to the cues humans use to perceive emotions in others. For example, a video

camera might capture facial expressions, body posture and gestures, while a microphone might capture speech. Other sensors detect emotional cues by directly measuring physiological data, such as skin temperature and galvanic résistance. Recognizing emotional information requires the extraction of meaningful patterns from the gathered data. This is done using machine learning techniques that process different modalities speech recognition, natural language processing, or facial expression detection, and produce either labels (i.e. 'confused') or coordinates in a valence-arousal space.
Affective computing is an emerging area of research and practice broadly defined as "computing that relates to, arises from, or deliberately influences emotion". This includes computational models of emotion generation and cognitive-affective interactions, affective user models and cognitive-affective architectures, methods for emotion sensing and recognition, affect-adaptive user interfaces, and virtual affective agents. Of particular interest to cognitive scientists is the area of affective modeling, focusing on the development of computational models of affective processes and cognitive-affective interactions.

- The nature of motivation, emotions and feeling.
- The detection of emotional and other affective states and processes.
- The nature of intelligence and the relationships between intelligence and emotions.
- The physiology of the brain and other aspects of human physiology relevant to affective states .
- Requirements for effective human-computer interfaces in a wide range of situations.
- Wearable devices with a range of sensing and communicative functions.
- Philosophical and ethical issues relating to computers of the future.

It will increasingly be feasible to install sensors and computing devices in furniture, in walls, in car seats, in driving controls, in clothing, in jewelry and even in implants. So it will be possible to have a wide range of sensors, processors and transmitters, constantly monitoring, analysing, recording, and transmitting information about one's blood pressure, temperature, blood sugar level, muscular tension, and many other physiological states. Some of these devices, suitably hidden, could also monitor various aspects of the environment, including other people. Thus even your friends and colleagues will easily be able to record your conversation, your facial expressions, and perhaps with remote sensors your muscular tension, temperature, sweating, etc.
Humans recognize emotional states in other people by a number of visible and audible cues. Facial expression is a valuable means in the communication of emotion. Moreover, there is evidence of the existence of a number of universally recognized facial expressions for emotion such as happiness, surprise, fear, sadness, anger, and disgust. In addition, the body (gesture and posture) and tone of voice are the other core channels for the communication of emotion. There are also a number of psycho-physiological correlates of emotion, such as pulse or respiration rate, most of which cannot easily be detected by human observers, but which could be made accessible to computers given appropriate sensing equipment. From all of these channels, researchers of Artificial Intelligence in education are attempting to infer the student's affective state. Preferably, evidence from many modes of interaction should be combined by a computer system so that it can generate as valid estimations as possible about.

## III.Our System And Assumptions

Facial features extracted from images or video clips can be broadly categorized as geometrical features and appearance based features using hear wavelet transform. Geometrical features consist of shapes of facial components (eyes, lips, smiling etc.) and salient points on the face (nose tip etc.).This provides us the component related to the variations resulting from facial expressions, which are then classified using Support Vector Classifiers (SVC). It is expected that methods that use both geometrical and appearance based features give more accurate results.

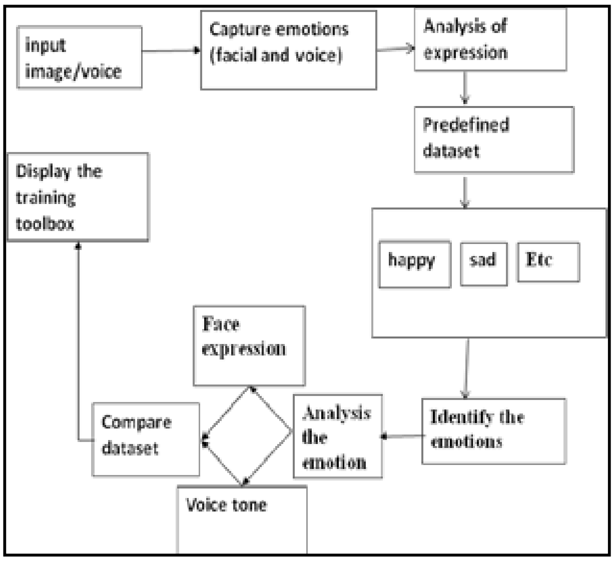


Fig.1: Overall system architecture

The originality of the work consists on the one hand, in the supposition that all the necessary information for the recognition of expressions is contained in the deformation of certain characteristics of the eyes, mouth and eyebrows and, on the other hand, in the use of the Belief Theory to make the classification. Nevertheless, their method has important restrictions. The features extraction program captures each facial frame and extracts the feature points which are sent to the emotion classifier. When an emotional change is detected, the output of the 7-emotion classifier constitutes an emotion code which is sent to Maxine's character. As far as the voice interface is concerned, we have endeavoured to reduce to a minimum the time that elapses between the point at which the user finishes speaking and the point at which the answer begins.

### A. Emotional Recognition from Facial Expressions

Knowing how facial expressions relate to the underlying emotional experiences is an important factor in using facial expression measurements as an input signal in affective computing. Therefore, the assessment of emotional experiences from objectively measured facial expressions becomes an important research topic. In the field of facial expression recognition, several efforts have been made in trying to recognize expressions of discrete emotions, especially the ones suggested. Although there is evidence for universal facial expressions of certain emotions, it is important to realize that there are also differences in the facial behaviour of different people. With regard to this issue supported that the most accurate interpretation of facial expression benefits from the knowledge of what is normative for each individual. Hence, the findings that there are considerable differences in facial behaviour between individuals recommend that the best results in emotion estimation could be obtained using a person adaptive system. This system would form an individual model of facial behaviour for each individual user. An important issue is that many of the existing facial recognition systems rely on analysing single facial images instead of tracking the changes in facial expressions continuously. It would be more meaningful if the computerized learning environments could analyse the student's facial expressions continuously to be able to react to changes in the student's emotional state at the right time. The point that the lack of temporal information is a significant limitation in many facial expression recognition systems. Consequently, methods for analysing facial expressions in human-computer interaction, especially those concerning computer-aided learning systems.

## IV. System Preliminaries

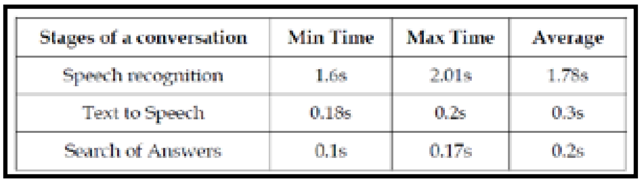
### A. IEMOCAP Database

The IEMOCAP corpus is an audiovisual database, emotions are expressed through a combination of verbal and non-verbal channels, a joint analysis of speech and gestures is required to understand expressive human communication. To facilitate such investigations, this paper describes a new corpus named the "interactive emotional dyadic motion capture database" (IEMOCAP), collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This database was recorded from ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expressions and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). The corpus contains approximately twelve hours of data. The detailed motion capture information, the interactive setting to elicit authentic emotions, and the size of the database make this corpus a valuable addition to the existing databases in the community for the study and modelling of multimodal and expressive human communication.

### B. Segmentation and Emotional Annotation

The data is transcribed and manually segmented into dialog turns. Six annotators were asked to assess the emotional contents of the actors during their speaking turns. The selected labels include happiness, anger, sadness, neutral, frustration, surprise, excited, fear and other. The subjective evaluation was conducted such that each turn was separately annotated by three evaluators. For a given turn, we are interested in studying the emotional states of both the speaker and the listener. We consider all the emotional labels assigned by the evaluators to the surrounding turns not just the consensus labels associated with these turns. Then, we assign the majority vote among these two sets as the emotional state of the listening segments. Consider the first listening turn of subject is a His/her previous turn received the labels happiness (2) and neutral, and his/her following turn received the labels happiness, neutral and surprise.

Table 1
Distribution of the Emotional Labels Assigned to the Actors' Listening and Speaking Turns



| Stages of a conversation | Min Time | Max Time | Average |
|---|---|---|---|
| Speech recognition | 1.6s | 2.01s | 1.78s |
| Text to Speech | 0.18s | 0.2s | 0.3s |
| Search of Answers | 0.1s | 0.17s | 0.2s |

The variability of facial features in the mouth and jaw areas To avoid capturing the anticipatory effect of articulation, the initial and ending 300 ms of the listener's facial expressions are discarded. Also, the experiments do not consider the segments shorter than 500 ms, these constraints limit the number of turns considered in this study (1,252 turns). Table 1 shows the number of samples in each emotional class for both speakers and listeners using the aforementioned portion.

### C. Facial and Acoustic Features

Facial features are extracted from the markers' information. First, the markers are translated and rotated using an approach based on singular value decomposition (SVD) after compensating for rotation and translation; the remaining movements of the facial markers correspond to facial expressions. The study uses as features the 3D location of the 53 facial markers and the head rotation parameters (i.e., pitch, roll, and yaw)., For each turn, seven high-level statistics are extracted from the facial features: minimum, maximum, standard deviation, mean, median, lower quartile, and upper quartile. Altogether, we create a 1,134 dimension feature vector. Due to the high dimension of this feature set, we used correlation feature selection (CFS) criterion to reduce its dimension for analysis section (Section 5). This technique extracts a set of features having high correlation with the emotional labels, but low correlation between themselves. The spectral feature comprises of RASTA-style filtered auditory spectrum, Mel frequency cepstral coefficients (MFCCs), and a set of statistics extracted at frame level, across spectral components. The statistics include energy in low band (25-650 Hz) and high band (1-4 kHz), multiple roll-off points, flux, entropy, variance, skewness, kurtosis, and slope. Spectral components are estimated with the short-time discrete Fourier transform (DFT) amplitudes.

The energy related features include sum of auditory components before and after RASTA filters, root mean square (RMS) and zero-crossing rate. The voices LLDs include the fundamental frequency (F0), probability of voicing, jitter and shimmer.

## V. Experimental Evaluation

### A. Emotional Speech Recognition

The modulation of voice intonation is one (of the) main channel(s) of human emotional expression. Certain emotional states, such as anger, fear, or joy, may produce physiologic reactions, such as an increase of cardiac vibrations and more rapid breathing. These in turn have quite mechanical and thus predictable effects on speech, particularly on pitch timing and voice quality .Some researchers have investigated the existence of reliable acoustic correlates of emotion in the acoustic characteristics of the signal. Their results agree on the speech correlates that are derived from physiological constraints and correspond with broad classes of basic emotions, but disagree and are unclear concerning the differences between the acoustic correlates of fear and surprise or boredom and sadness. This is perhaps explained by the fact that fear produces similar physiologic reactions to surprise, and boredom produces similar physiologic reactions to sadness, and consequently very similar physiological correlates result in very similar acoustic correlates . This also provides an explanation for the results of Tickle' experiments, demonstrating that the best emotional speech recognition score for humans was only. Additionally, Tickle's experiments indicated that there is only little difference between the performance in detecting the emotions conveyed by someone speaking the same language or another language. This could be attributed to the fact that physiological effects of emotional states are rather universal, meaning that there are common tendencies in the acoustical correlates of basic emotions across different cultures .Research dealing with speech modality, both for emotional automated production and recognition by technology, has only been active for a few years and has gained much. However, it is uncertain whether research results would effectively generalize to naturally produced, rather than an "acted" emotional expression. The task of machine recognition of basic emotions in non-formal everyday speech is extremely challenging and will greatly contribute toward the evolution of computerized learning systems.

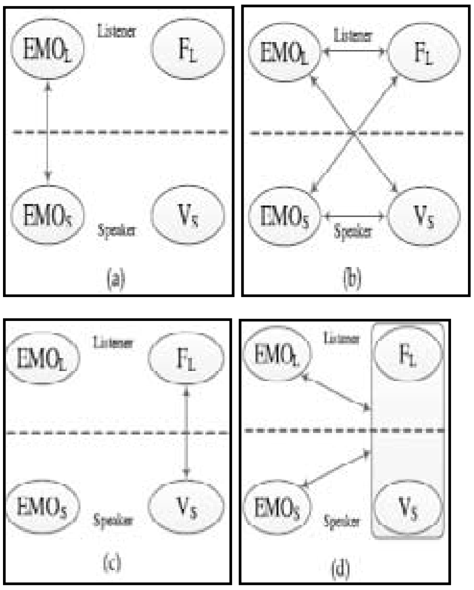


Fig.2: The four aspects in dyadic interactions considered in the analysis:

(a) Dependence between the emotional states of the dialog partners,
(b) Dependence between the emotion of one subject and the expressive behaviours,
(c) Dependence between heterogeneous behaviours from the dialog partners and
(d) Effect of cross-subject multimodal information for emotion discrimination.

EMOL: listener's emotion,
EMOS: speaker's emotion,
FL: listener's facial features, and
VS: speaker's voice

The Similarities with metrics such as distance or correlation. Instead, we use mutual information to quantify the dependences rather than similarities between modalities.

### Feature selection

The process of feature selection was divided into two steps in which an optimal set of features was selected from highly dimensional data set. In the first step, one of the features X and Y was filtered if the absolute value of Pearson's correlation coeficient rPCC(X; Y ) was higher than 0.95.

$$r_{PCC}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}} \qquad (1)$$

In the second step, a genetic algorithm was applied as the heuristic to find the optimal set of features. Each individual in population of solutions was represented as a set of features and was assessed by 10-fold cross validation whose outcome in form of F □ score served as the fitness function.

### Evaluation and Results

Precision and recall were used to evaluate the performance of the proposed system. The precision P is the percentage of correctly classified items among all items that were assigned to the category by the classifier. It is defined by following formula:

$$P = \frac{N_{TP}}{N_{TP}+N_{FP}} \cdot 100\ \% \qquad (2)$$

Where NTP represents number of true positives and NFP number of false positives. The recall R is the percentage of correctly classified items among all items belonging to that category and is computed according to formula:

$$R = \frac{N_{TP}}{N_{TP}+N_{FN}} \cdot 100\ \% \qquad (3)$$

where NTP represents number of true positives and NFN number of false negatives. The F-score is combination of formula 2 and formula 3:

$$F_{\text{score}} = 2 \cdot \frac{P \cdot R}{P+R} \qquad (4)$$

Equation (2) gives the mutual information for discrete variables X and Y, given their marginal and joint probability mass functions (PMFs). Facial and acoustic features provide continuous values. Therefore, we discredited the features using the K-means algorithm. Given the differences in the range across features, we apply z-normalization before estimating the clusters. The proposed to compare the similarity in behaviours between individuals during their interactions (paired condition), with the similarity in behaviours between individuals engaged in different conversations (unpaired condition).The analysis follows a similar approach by comparing the mutual information in paired and unpaired conditions. The nodes EMOL and EMOS are the emotional states of the listener and speaker, respectively. The node FL describes the facial features of the listener. The node VS represents the features from the speaker's voice.

### 1. Emotion Entrainment

Fig.2a the interpersonal adaptation theory, Conversational partners tend to converge in the behaviours showing reciprocal and mirroring patterns. The exception occurs when the subjects decide to diverge in their behaviours to cope with a given situation. Co occurrence between the Emotions Displayed by Speakers and Listeners in the Turns during Spontaneous Dialogs (IEMOCAP Corpus). This result supports the emotional adaptation hypothesis. If we account for the marginal distribution of the speaker's and listener's emotions (see Table 1), and assuming their independence, the expected ratio of observing similar emotions by chance is 30 percent.The co occurrence of emotions between dialog partners decreases when one of them is in neutral state. Notice that neutral state is not always well defined and it is often confused with other emotions.

### B. Cross Subject Relation of Emotion and Modalities

The aforementioned emotional synchronization patterns, we hypothesize that facial gestures of the listeners provide complementary information about the speakers' emotions, and that the acoustic features of the speakers provide information about the emotion of the listeners i.e., diagonal arrows this cross-subject emotional entrainment is studied with mutual information.

### C. Emotional Recognition with the Use of Questionnaire

Many researchers have used static methods such as questionnaires, dialogue boxes, etc., in order to infer a user's emotions. These methods are easy to administer but have been criticized for being static and thus not able to recognize changes in affective states. It recognized that self-reporting of emotions simplifies the recognition problem. It stated that this approach transfers one of the hardest problems in adaptive affective interfaces from the computer to the user. Another advantage of the questionnaire is that it provides feedback from the user's point of view and not an outsider's. Questionnaires can be used to infer users' emotions, either standalone or assisting another affect recognition method. On the other hand, the way questions are framed and demonstrated. The order in which questions are asked, and the terminology employed in questions is all known to affect the subject's responses. Similarly, there is evidence that judgments on rating scales are non-linear, and that subjects hesitate to use the extreme ends of a rating scale. Hence, when using verbal scales, one should make sure that the terminology employed and the context in which it is to be presented, really reflect the subjective significance of the subject population. The student's recognized emotional state should be properly managed from the computer aided affective learning system, based on pedagogical models which integrate our knowledge about emotions and learning. The system would assess whether the learning process is developing at a healthy rate. If there is a positive development, the system should help the learner maintain this emotional.

### D. Complementariness of Cross-Subject Behaviours

Fig. 4d the previous results highlight the connection between nonverbal behaviours of one subject and the emotions displayed by the other subject. An important question is to determine whether the cross-subject behaviours are complementary to or redundant with the own behaviours displayed by the subject. To address this question, we compare the mutual information in single modality setting with the mutual information in cross-subject multimodality setting.

The listener's facial features and speaker's acoustic features are concatenated into a single vector before performing the K-means algorithm. The mutual information between the speaker's emotion and speaker's voice, I(EMOS ; VS ), with the mutual information between speaker's emotion and the cross-subject, cross-modality features, I(EMOS ; ½VS ; FL Š). Both figures show an increase in mutual information in the cross-subject multimodal settings (solid lines). These results indicate that cross subject behaviours provide complementary information about the displayed emotion during dyadic interactions. Section 6 validates these results in emotion recognition experiments.

An emotional state among Angry, Neutral, Sad, Happy, Disgust, Surprise, and Fear. Results indicated that Happy and Sad facial expressions were easily recognized by the participants with high percentages, 93 and 97 percent, respectively. Fear and Neutral facial expressions were recognized with lower percentages by the participants, 73 and 77 percent, respectively. Fear was mostly confused with Surprise and Neutral with Angry. In our final experiment, however, facial expressions were accompanied by speech with a related tone of voice because we believed that this combination could improve recognition of the target emotions. The relevant sentences were uttered by a trained actress to convey the desired emotion.

### E. Results on the IEMOCAP Database

The evaluation assesses the improvement in emotion recognition performance when we consider cross-subject multimodal information. We separately consider both speaker's emotions and listener's emotions recognition tasks. The experiments are conducted using leave-one- speaker-out cross-validation (speaker independent training/testing partitions).The evaluation uses linear kernel support vector machine (SVM) with sequential minimal optimization (SMO). The soft margin parameter c is selected by optimizing the baseline classifiers: SVML (FL) that recognizes the listener's emotions using his/her facial features.

Given that the data is not emotionally balanced (see Table 1), we estimate the precision rate for each emotional class (i.e., fraction of retrieved samples for one emotional class that are relevant).

Then, we estimate and report the average precision (P) across classes. Likewise, we estimate the recall rate for each emotional class (i.e., fraction of relevant samples that are correctly classified). We report the average recall (R) across classes. With these values, we calculate the F-score (F).In addition, we report the accuracy (A) of the classifiers.

### 1. Recognition of the Listener's Emotion

The results of the listener's emotion classification task under different conditions. The first row shows the baseline classifier, which is trained with only the facial features extracted from the listeners—SVML (FL),the emotional adaptation effect by recognizing the listeners' emotions using only the speakers' emotions—SVML (EMOS). Table 3 shows that this classifier achieves an accuracy of 70.2 percent. In many real applications, the speakers' emotion is not available and needs to be estimated. Following this direction, we consider both explicit and implicit modelling of the speaker's emotions to recognize the listener's emotions. We propose a cascade SVM in which we explicitly estimate the speaker's emotion using his/her acoustic features (see Fig. 3a).

The output of this classifier and the facial features from the listeners are used as input to recognize the listener's emotion Cascade SVML (FL ; VS ).We also explore the case in which the speaker's Emotion is implicitly incorporated in the classifiers by directly using the speaker's behaviours. First, we evaluate the performance of the classifier when we consider only features extracted from the speaker's voice—SVML (VS). This classifier achieves an accuracy of 55 percent, which is lower than the baseline classifier. However, the performance is significantly higher than chances (25 percent).This result demonstrates the discriminative power of the speaker's voice to distinguish the listener's emotion. The first challenge on Facial Expression Recognition and Analysis (FERA'11) focused on these two kinds of affect description. Meta-analysis of challenge results. These methods generally use discrete systems whether based on static descriptors (geometrical or appearance features) and/or on. For each system, a new correlation-based feature selection is performed using a delay probability estimator. This process is particularly well-adapted to unsure and possibly time-delayed labels. The prediction is then done by a nonparametric regression using representative samples selected via a k-means clustering process.

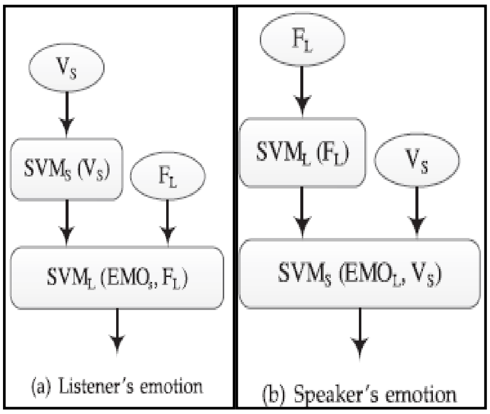


Fig.3: Cross-subject emotion recognition with cascade SVMs. The dialog partner's emotion is used as feature to recognize the target subject's emotion.

## 2. Recognition of the Speaker's Emotion

It follows a similar approach to recognize the speaker's emotion. The trained with features describing the speaker's voice SVMS (VS). Although the portion of the corpus used in the evaluation is different, the average recall of our baseline (50.6 percent) is similar to the one reported in a previous study using only acoustic features. When the listener's emotion is known, the speaker's emotion can be recognized with 72 percent accuracy SVMS (EMOL). When the speaker's voice and the listener' emotions are used, the classification accuracy improves to 74 percent SVMS (VS; EMOL). When the listener's emotion is explicitly estimated using a cascade SVMS (V; FL) we achieve a 62.5 percent accuracy.

## 3. Results on the SEMAINE Database

This section validates the analysis on cross-subject, cross-modality affective entrainment in more natural recordings (i.e., no acted corpus recorded with video cameras). For this purpose, we present emotion classification experiments using the sustained emotionally colored machine-human inter- action using nonverbal expression (SEMAINE) database [40]. This multimodal corpus was collected using the sensitive artificial learner (SAL) technique [41] to engage users in emotional conversations with an operator. The operator can be a virtual character (i.e., semi-automatic SAL and automated SAL) or another human (i.e., solid SAL)

The study relies on the computer expression recognition toolbox (CERT)to extract facial features. CERT automatically extracts action units (AUs), defined in the facial action coding system (FACS). AUs describe the facial movements of individual muscles or groups of muscles. The toolkit processes the video frame-by-frame, providing high accuracy and robustness against different illumination conditions This standard also defines a set of facial animation parameters (FAPs) to modulate the facial appearance by moving the FPs. These FAPs are derived from the definition of the AUs. Therefore, there is a close relationship between the markers' trajectory—features on the IEMOCAP corpus—and the AUs—features on the SEMAINE corpus. The classification experiments consider 20 AUs and three head rotation parameters provided by CERT. Similar to the approach used with the facial markers, we estimate seven statistics from these features at turn level (minimum, maximum, standard deviation, mean, median, lower quartile, and upper quartile). Altogether, a turn is represented with a 161D facial feature vector.

The user's emotional reactions are annotated in terms of activation (i.e., active versus passive) and valence (i.e., positive versus negative) dimensions, using the MATLAB. Instead of turn level assessments, this annotation scheme continuously captures the perceived emotional primitives values, as the annotators move the mouse cursor over a graphical user interface (GUI) displaying the activation/ valence space. The classification experiments in this study consider only the user's emotions. The emotional labels include turns when the user is both speaking and listening. Therefore, this corpus is suitable for the proposed cross-subject, cross-modality evaluation. The emotional ground truth for each of these turns is calculated by averaging the scores across evaluators and across frame.

The emotional ground truth for each of these turns is calculated by averaging the scores across evaluators and across frames The delay is caused by the intrinsic reaction time between the perception of the expressive behaviours and the annotation of the stimuli (i.e., moving the cursor). Nicolle et al. the delay on four emotions attributes (activation, valence, expectation, and power) in the SEMAINE database using correlation analysis. They reported average delays between 3 and 6 seconds. Following a similar approach, we propose to estimate the optimal delay with the mutual information between the frame-level facial features (F).Instead of dealing with continuous emotional attributes, and created K emotional clusters in the activation-valence space by using the K-means algorithm.

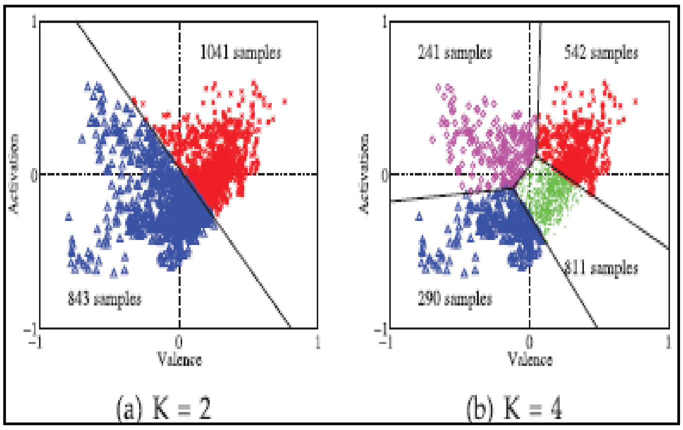


Fig.4. Clusters obtained by the K-means algorithm in the valence-activation space (SEMAINE).

The figure gives the number of turns assigned to each class. The classification experiments follow the settings described in Section 3.1 (i.e., SVM-SMO with c = 0.1). The selected portion of the database contains nine users. We train and test the classifiers using a leave-one-speaker-out cross-validation. The feature sets are reduced using CFS, using the training set of each fold. For K =2, CFS selects an average of 29 facial and 94 acoustic features. For K =4, CFS selects an average of 39 facial and 98 acoustic features. The feature selection is performed with audiovisual features from the users. However, the same feature set is estimated from the operator's behaviours. It reports the user's emotion classification experiments during the listening and speaking segments, considering the two emotional space clustering.

### 4. Recognition of User's Emotion while Listening

For K=2 (50 percent chance level), a classifier trained with only the user's facial expressions achieved an accuracy of 55.4 percent (turns when the user was listening). Incorporating features describing the operator's face, voice or both significantly improves the accuracy by at least 3.8 percent. The best performance is obtained when features describing the operator's voice and facial expression are added to the user's facial features This result represents statistically significance improvement over the performance of the classifier trained with only the facial features of the user (p-value < 0:0001) population proportion test).

### 5. Recognition of User's Emotion while Speaking

During the segments when the user is speaking, the face is the only modality available for the operator. From the user, we extract his/her facial and acoustic features. It provides the performance for different combinations. There are three baseline classifiers for which only features from the user are used (i.e., face, voice or both modalities). The baseline classifiers trained with features describing the user's face achieve 53.6 and 47.2 percent accuracy rates for K=2 and K= 4, respectively. The user's acoustic features do not provide significant discriminated information to recognize his/her emotion. That adding cross-speaker information (i.e., the operator's face) improves the accuracy and F-score rates in all the settings, both for K=2 and K =4. When K=2, the addition of features describing the operator's face yields statistically significant improvements for accuracy and F-score (p-value < 0:0001), across classifiers. The best performance is achieved when only acoustic cues of the user and facial expressions of the operator are employed (A ¼ 65:1%). For K=4, the accuracy improves in the three cross-subject, multimodal settings. The best performance is achieved by incorporating user's voice and operator's facial expressions (A 52:7%). These results validate the benefits of using cross-subject features for multiparty emotion recognition.

## VI. Conclusion

Special care has been taken in making it possible multimodal natural user-agent interaction: communication is accomplished via text, image and voice (natural language). Our embodied agents are equipped with an emotional state which can be modified throughout the conversation with the user, and depends on the emotional state detected from the user's facial expressions. In fact, this nonverbal affective information is interpreted by the agent, which responds in an empathetic way by a comparing its voice intonation, facial expression and answers.The chapter has focused on two main aspects: the capture of the user emotional state from web cam images and the development of a dialog system in natural language that takes also emotional aspects into account. The facial expression recognizer is based on facial features' tracking and on an effective emotional classification method based on emotional classification. From a set of distances and angles extracted from the user images the analysis of a sufficiently broad image database, the classification results are acceptable, and recent developments has enabled us to improve success rates. The utility of this kind of information is clear: the general vision in that is a user's emotion could be recognized by a computer, human computer-interaction would become more natural, enjoyable and productive. The dialog system has been developed so that the user can ask questions, give commands or ask for help to the agent. It is based on the recognition of patterns, to which fixed answers are associated.. Special attention has also been paid in adding an emotional component to the synthesized voice in order to reduce its artificial nature. Voice emotions also follow ones and are modeled by means of modifying volume, speed and pitch.

## VII. Acknowledgement

First and foremost, The authors would like to thank the God Almighty, who guides us always in the path of knowledge and wisdom.We thank the editors and anonymous reviewers for their valuable comments to significantly improve the quality of this paper.We are very much grateful to all the staff members and my friends who helped a lot to complete this work.

## References

[1]. *Gary McKeown, Michel Valstar, Member, IEEE, Roddy Cowie, Member, IEEE, Maja Pantic, Senior Member, IEEE and Marc Schr oder" The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent"*

[2]. *Jérémie Nicolle,Vincent Rapp,Kévin Bailly" Robust Continuous Prediction of Human Emotions using Multiscale Dynamic Cues"*

[3]. *L. Bell, J. Gustafson, and M. Heldner, "Prosodic Adaptation in Human-Computer Interaction," Proc. 15th Int'l Congress of Phonetic Sciences (ICPhS '03), pp. 2453-2456, Aug. 2003.*

[4]. *Yan Tong, Student Member, IEEE, Wenhui Liao, Member, IEEE, and Qiang Ji, Senior Member, IEEE-2007" Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships"*

[5]. *M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," J. Multimedia, vol. 1, pp. 22-35, Sept. 2006*

[6]. *S. Brennan, "Lexical Entrainment in Spontaneous Dialog," Proc. Int'l Sump. Spoken Dialogue (ISSD '96), pp. 41-44, Oct. 1996.*

[7]. *M.E. Babel, "Phonetic and Social Selectivity in Speech Accommodation,"PhD dissertation, Dept. of Linguistics, Univ. of California Berkeley, 2009.*

[8]. *R. Levitan and J. Hirschberg, "Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions," Proc. 12th Ann. Conf. Int'l Speech Comm. Assoc.*

[9]. *M. Natale, "Convergence of Mean Vocal Intensity in Dyadic Communication as a Function of Social Desirability," J. Personality and Social Psychology, vol. 32, no. 5, pp. 790-804, Nov. 1975.*

[10]. *R. Coulston, S. Oviatt, and C. Darves, "Amplitude Convergence in Children's Conversational Speech with Animated Personas," Proc. Int'l Conf. Spoken Language Processing (ICSLP '02), vol. 4, pp. 2689-2692, Sept. 2002.*

*Ranjidha.P received the B.Tech. degree in Information Technology from Roever Engineering College affiliated to Anna University, Tamil Nadu, India, 2012.She joined Srinivasan Engineering College affiliated to Anna University in 2012 for the M.E. degree in Computer Science and Engineering. Her research interests include affective computing, network security, and mobile computing.*

*Bharathalakshmi.R received the B.E. degree in Computer science and engineering from VPMM Engineering college affilated to Anna University, Tamil Nadu, India, 2010 and received M.E. degree in Computer Science and Engineering from Sairam Engineering College affiliated to Anna University, Tamil Nadu, India in 2012 respectively. She is an Assistant Professor and supervisor of M.E. students at Srinivasan Engineering College affiliated to Anna University. Her research interests include networking, cryptography, network security and cloud computing, mobile computing, affective computing.*

*Kasthuri.S received the B.E. degree in Computer Science and Engineering from Periyar Maniammai, University Tamil Nadu, India, 2012. She joined Srinivasan Engineering College affiliated to Anna University in 2013 for the M.E. degree in Computer Science and Engineering. Her research interests include Affective computing, mobile computing, data mining and image processing.*