Speech Emotion Analysis: Exploring the Role of Context

Ashish Tawari and Mohan Manubhai Trivedi, *Fellow, IEEE*

***Abstract*—Automated analysis of human affective behavior has attracted increasing attention in recent years. With the research shift toward spontaneous behavior, many challenges have come to surface ranging from database collection strategies to the use of new feature sets (e.g., lexical cues apart from prosodic features). Use of contextual information, however, is rarely addressed in the field of affect expression recognition, yet it is evident that affect recognition by human is largely influenced by the context information. Our contribution in this paper is threefold. First, we introduce a novel set of features based on cepstrum analysis of pitch and intensity contours. We evaluate the usefulness of these features on two different databases: Berlin Database of emotional speech (EMO-DB) and locally collected audiovisual database in car settings (CVRRCar-AVDB). The overall recognition accuracy achieved for seven emotions in the EMO-DB database is over 84% and over 87% for three emotion classes in CVRRCar-AVDB. This is based on tenfold stratified cross validation. Second, we introduce the collection of a new audiovisual database in an automobile setting (CVRRCar-AVDB). In this current study, we only use the audio channel of the database. Third, we systematically analyze the effects of different contexts on two different databases. We present context analysis of subject and text based on speaker/text-dependent/-independent analysis on EMO-DB. Furthermore, we perform context analysis based on gender information on EMO-DB and CVRRCar-AVDB. The results based on these analyses are promising.**

***Index Terms*—Affect analysis, affective computing, context analysis, emotional speech, emotion intelligence, emotion recognition, vocal expression.**

I. Introduction

SPEECH signals convey not only words and meanings but also emotions. Besides human facial expressions, speech has been proven to be another promising modality for the recognition of human emotions. Spoken communication between humans is intricately linked with linguistic information (verbal content) and paralinguistic information such as tone, emotional states, and gestures. In such interaction, one's affective state plays a fundamental role in enriching communication. Current human–computer interaction (HCI) systems, however, ignore the user's affective states, losing a significant portion of information available in the interaction process. The human–computer paradigm suggests that user interfaces of the future need to detect subtleties and changes in the user's behavior, especially his/her affective behavior, and to initiate interactions based on this information rather than simply responding to the user's commands. The future human-centered multimodal HCI will change the ways in which we interact with computer systems. For example, an intelligent automobile system with a fatigue detector could monitor the vigilance of the driver and apply appropriate action to avoid accidents. Another important application of automated systems for human affect recognition is in affect-related research (e.g., in psychology, psychiatry, behavior science, and neuroscience), where such systems can eliminate the tedious manual task of processing data. Research areas like social and emotional development studies [1], mother–infant interaction [2], and psychiatric disorders [3] would be substantially benefited. Automatic detection of fatigue, depression, and anxiety could also form an important step towards personal wellness and assistive technologies [4]. Some initial efforts toward such advanced system include [5]–[8].

II. Related Research and Motivation

The auditory signal in conversation carries various kinds of information. If we disregard the manner in which it was spoken by considering only the verbal part, we might miss important aspects of the pertinent utterance and even misunderstand the spoken message. Humans are capable of recognizing subtle differences implied in an utterance. It is currently hard to imagine an artificial system reaching such a high degree of discrimination. A technical approach for classification would rely on the kind and number of emotions allowed. Research on vocal affect recognition is largely influenced by the basic emotion theory. Most of the existing efforts in this direction aim at the recognition of a subset of basic emotions from speech signals. In recent years, however, few studies have been made focusing on interpreting speech signals in terms of certain application-dependent affective states [9]–[12].

Another aspect of the vocal affect analysis is to specify the auditory features to be estimated from the input audio signal. The research in psychology and psycholinguistics provides several results on acoustic features which are correlated with emotion expressions. The popular features are prosodic features (e.g., pitch-related features, energy-related features, and speech rate) and spectral features (e.g., mel frequency cepstral coefficients (MFCCs) and cepstral features). Most of the existing approaches are trained and tested on speech data that were collected by asking actors to speak prescribed utterances with certain emotions [10], [13]. However, the fact

Manuscript received December 15, 2009; revised April 01, 2010; accepted June 02, 2010. Date of current version September 15, 2010. This work was supported in part by a grant from the UC Discovery Project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hamid K. Aghajan.

The authors are with the Computer Vision and Robotics Research Laboratory, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: atawari@ucsd.edu; mtrivedi@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2058095

that deliberate behavior differs in audio profile and timing from spontaneous behavior has led research to shift towards the analysis of spontaneous human behavior in naturalistic audio recordings. In [14], a good overview of recent advancement towards spontaneous behavior analysis in audio, visual, and audiovisual domains is presented. In spontaneous human behavior, identifying subtle changes in vocal affect expression based on only acoustic information is not sufficient [15]. In [16], the use of lexical cues has been reported to improve the performance. However, extracting these features automatically is a challenging task. In contrast to spoken language processing, which has witnessed significant advances, the processing of "emotional speech" is still a difficult problem. Studies show that the accuracy of automated speech recognition system tends to drop to 50%–60% for emotional speech from 89%–90% for neutrally spoken words [17], [18]. The same has been shown for speaker verification systems [19]. In such scenarios, the use of context information can help us improve the performance.

It is evident that the ability of emotion perception in human beings is greatly influenced by the contextual information. The work in [20] discusses how contextual information influences emotion annotation as well as machine-learned classification. In particular, nine nonexpert annotators when provided with the context information by giving them utterances along with the dialogue where these were produced, marked 3.4% more nonneutral emotions. Similarly, machine-learned classfication of negative emotions (*angry*, *doubtful*, and *bored*) is enhanced by incorporating automatically generated context information in two steps. The first step is to classify emotions into *angry* and *doubtfulORbored* by utilizing users' neutral speaking style, and the second step uses the dialogue context such as the total number of dialogue turns (referred to as *depth*) and the number of additional user turns necessary to obtain a particular piece of information (referred as to *width*) to distinguish between the *bored* and the *doubtful* categories. Authors claim that the classification process is substantially improved by adding both sources of contextual information.

Few studies [21]–[23] have investigated the role of context information like subject and gender. The work in [23] shows improved classification performance by incorporating a gender-detection module as a front end to overall emotion recognition system. Such contextual information is relatively easy to extract and simple to incorporate in existing classification system. We believe that enough attention is not been paid to evaluate the effectiveness of such contextual information for emotion recognition. Towards this end, in this paper, we present a systematic study on the use of context information like user (the speaker), text (the spoken content), and gender (male–female) on a publicly available Berlin Database of Emotional Speech (EMO-DB) and locally collected car database (CVRRCar-AVDB).

Despite the relatively high recognition accuracy reported on deliberate (acted) vocal expression, performance evaluation based on speaker- or text-independent analysis is still not often addressed. The well-documented EMO-DB database gives us the opportunity to perform such analyses. One of the best reported results on EMO-DB is in [24]. This presents the performance results of a series of machine-learning algorithms for emotion classification. It uses overlapping frames of 25 ms with a shift of 10 ms to extract a feature vector consisting of 15 coefficients: the log-energy, 12 MFCCs, the pitch period, and the voicing class. On the basis of the time series of these parameters, over 3800 statistical parameters were extracted to characterize semantic unit of varying lengths. Furthermore, feature selection, normalization, and discretization techniques are utilized to improve the base performance. Best classification performance is achieved using the SVM classifier. However, speaker- and text-independent performance is not analyzed, which is important for practical application of such system. In this paper, we present these studies on EMO-DB emotional speech corpus. An important related issue that should also be examined is how one can utilize information about the context (environment, observed subjects or the current task), in which the observed affective behavior was displayed. Towards this end, we also present the study of a context model based on gender information on the EMO-DB and CVRRCar-AVDB.

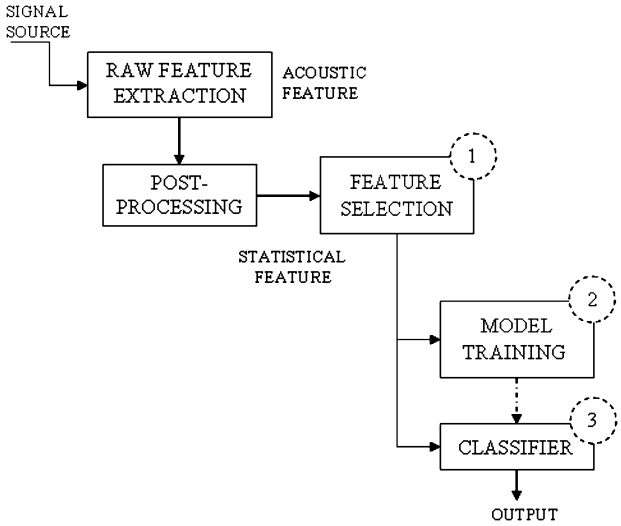


Fig. 1. Block diagram of classification system.

Fig. 1 shows a generic block diagram showing the three fundamental steps of information acquisition, extraction, and processing of parameters and classification of semantic units. The classification system has three different phases: feature extraction and selection (phase 1), to identify features to be used during classification; the model training phase (phase 2); and, finally, the testing phase (phase 3) to evaluate the performance of the system in terms of classification accuracy. The remainder of this paper is organized based on these phases. In Section III, we provide a brief overview of speech corpora used in our experiments. Section IV describes the statistical feature extracted from the speech signal. In Section V, details about the model training and evaluation are provided. Finally, in Section VI, we provide some concluding remarks and discuss future work.

III. Databases for Emotion Recognition Studies

Collecting emotional data is certainly useful for researchers interested in human affective expression recognition. Authentic affective expressions are difficult to collect because they are relatively rare and filled with subtle context-based changes. More-

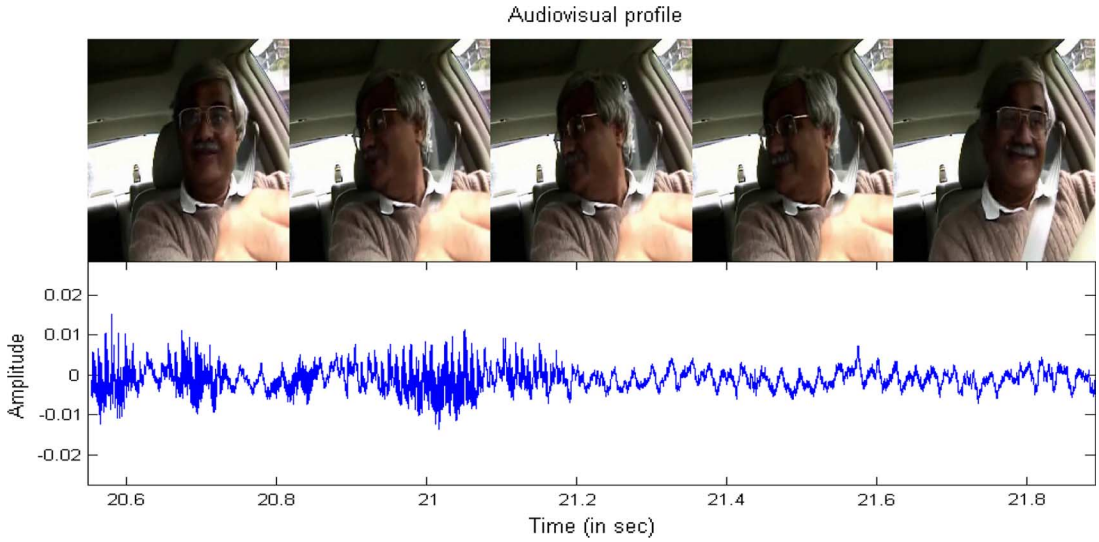


Fig. 2. Example of a "positive" utterance obtained during natural conversation between driver and passenger in a moving-car environment. Film strip shows samples of five images equally spaced in the utterance. The first half of the utterance contains the speech and later half the road noise. Notice, however, that facial features are more expressive after speech content while head dynamics are concomitant with the speech.

over, manual labeling of spontaneous emotions for the ground truth is very time-consuming and error-prone [25]. Due to these difficulties, a majority of databases encompass five to six emotions based on deliberate (acted) affect expression. In this research, two databases are used: EMO-DB [26] and CVRRCar-AVDB).

A. Berlin Database of Emotional Speech

The EMO-DB Berlin was recorded at the Technical University, Berlin. It comprises of six basic emotions (anger, boredom, disgust, anxiety, happiness, and sadness) as well as neutral speech. Ten professional German actors (five female and five male) spoke ten sentences with emotionally neutral content in the seven different emotions. In our study, we used 535 sentences available in the database. These sentences were not equally distributed between the various emotional states: 69 frightened (fea); 46 disgusted (dis); 71 happy(joy); 81 bored (bor); 79 neutral (neu); 62 sad (sad); 127 angry (ang).

B. Audio-Visual Affect Database (CVRRCar-AVDB)

In our ongoing research on driver-assistance systems [27], [28] and audiovisual scene understanding [29], [30], emotion recognition using multimodality will certainly help us improve the interface. With this motivation, we have put in significant effort on collection of audiovisual affect database in a challenging ambient of car settings [31]. The user at the driver's seat was prompted by a computer program specifying the emotion to be expressed. It also provides example of an utterance that can be used by the driver. The database is collected in both stationary and moving-car environments. We also have been collecting natural conversation between driver and passenger. Fig. 2 shows one such example of a happy utterance obtained in a moving-car environment. The cockpit of the automobile does not provide the comfort of noiseless anechoic environment. In fact, moving automobile with a lot of road noise has a drastic effect on signal-to-noise ratio (SNR) for audio channel [32] as well as challenging illumination condition for the video channel. In this study, we analyze emotional speech data from stationary car setting which gives the effect of the cockpit of the car with relatively high SNR value.

The database is collected with the use of an analog video camera facing the driver and a directional microphone beneath steering wheel. Fig. 3 shows the settings of the camera and microphone. Video frames were acquired approximately 30 frames per second, and the audio signal, captured, is resampled to a 16-kHz sampling rate. A version of the software for synchronizing as well as labeling the data is developed. Fig. 4 shows a snippet of the tool. The emotional speech has been labeled into three groups "pos," "neg," and "neu" for positive, negative, and neutral expressions. The data were acquired with four different subjects: two male and two female. Distribution of data for different categories is: 82 pos, 82 neg, and 60 neu. Fig. 5 presents a visualization of typical utterances for three emotions: pos/happy (top row), neg/sad (bottom row) and neutral (middle row) for CVRRCar-AVDB and EMO-DB databases. It can be noticed that the strength (amplitude) and quality (SNR) of the speech signal of the EMO-DB database obtained in a controlled setting is far superior than those of CVRRCar-AVDB obtained in the car setting.

IV. Feature Extraction, Selection and Transformation

A. Feature Extraction

So far, a large number of different features have been proposed to recognize emotional states from speech signal. These features can be categorized as acoustic features and linguistic features. Linguistic features analyzes the spoken content and often requires recognition of various words in the utterance. We avoid using these features since these demand for robust recognition of speech at first place and also are a drawback for multilanguage emotion recognition. Hence, we only consider the acoustic features. Within the acoustic category, we focus on prosodic and spectral features to model emotional states.

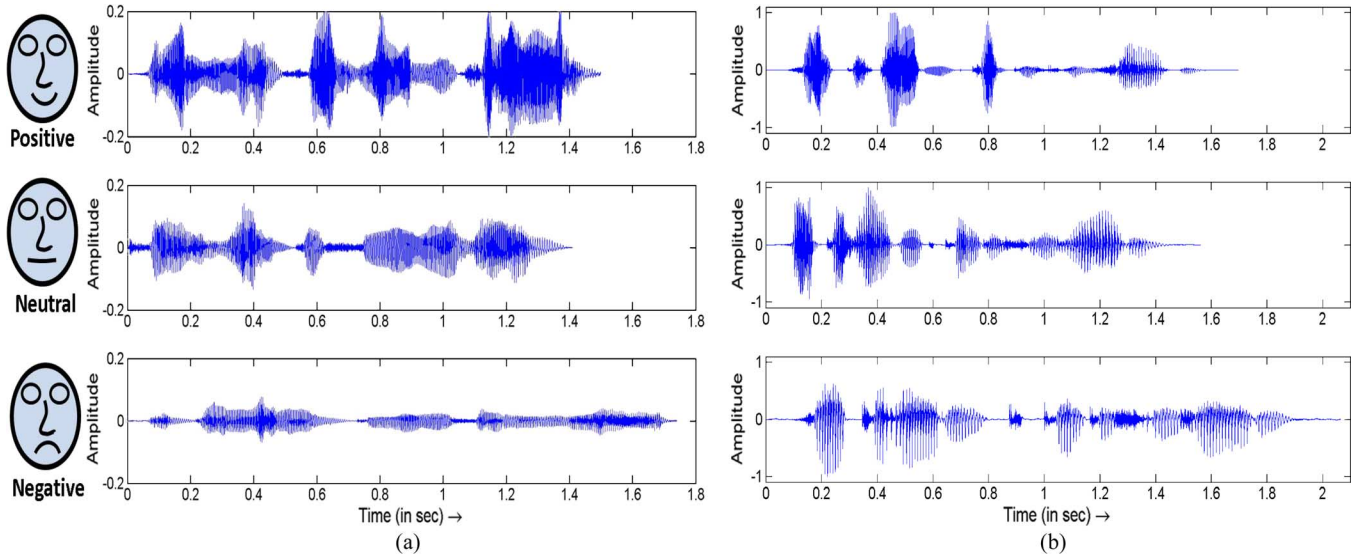


Fig. 3. Visualization of typical utterances in (a) CVRRCar-AVDB database and (b) EMO-DB database for three emotions: positive/happy (top row), neutral (middle row), and negative/sad (bottom row) and. Notice that the signal strength in the realistic car setting for CVRRCar-AVDB is very poor as compared with the controlled setting in EMO-DB.

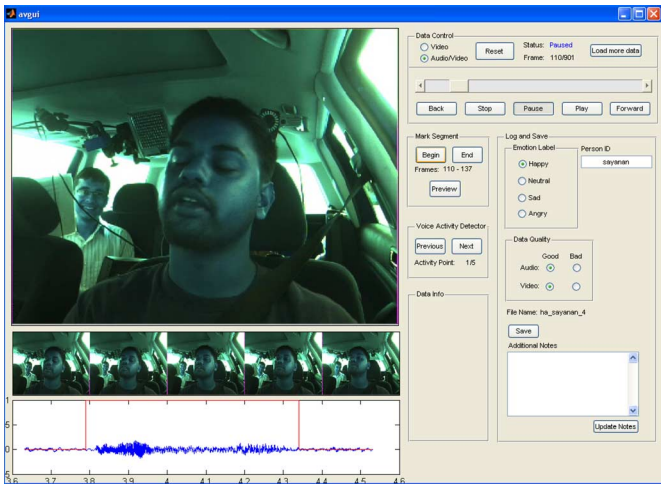


Fig. 4. Tool for audiovisual scene synchronization, cropping, and labeling. It has capability of voice activity detection (VAD) to decide precise onset of speech signal.

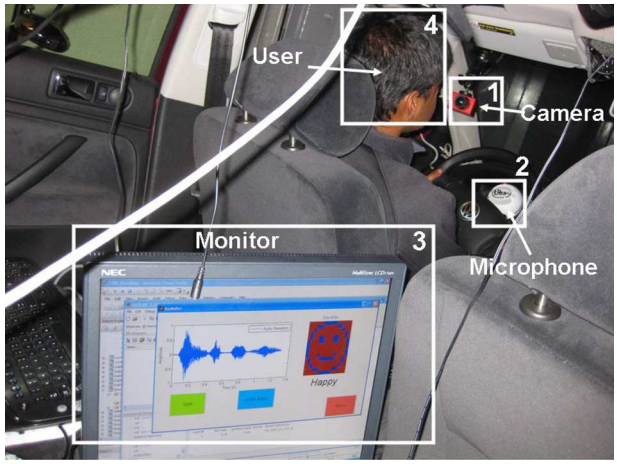


Fig. 5. Data acquisition and online recognition setting.

It is well known that different emotional states carry different prosodic patterns. Hence, prosodic feature like speech intensity, pitch and speaking rate can model prosodic patterns in different emotions. Similarly, spectral feature like MFCCs have been used successfully in emotion recognition.

For pitch calculation, we used the auto-correlation algorithm similar to [33]. The input signal is divided into overlapping frames with shift intervals (difference between the starting point of consecutive frames) of 10 ms. Each frame is of 60 ms long to be able to span three periods of minimum pitch value (in our case, 50 Hz). Pitch candidate over each frame is calculated and a dynamic programming technique is used to get the final pitch contour. Log-energy coefficients are calculated using 30-ms frames with shift interval of 10 ms.

The value of framewise parameters extracted from a few milliseconds of audio is of little significance to determine an emotional state. On the contrary, it is of interest to capture the time trend of these features. In order to capture the characteristics of the contours, we perform cepstrum analysis over the contour. For this, we first interpolate the contour to obtain samples separated by sampling period of speech signal, which is then used to calculate the cepstrum coefficients as follows:

$$c(m) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)|\, e^{2\pi k m j/N}, \qquad m = 0, 1, \cdots, N-1$$

where $X(k)$ denotes the N-point discrete Fourier transform of the windowed signal $(x(n))$. Cesptrum analysis is a source-filter separation process commonly used in speech processing. Cepstrum coefficients $c(0)$ to $c(13)$ and their time derivative (first and second order), calculated from 480 samples, are utilized to obtain the spectral characteristic of the contours. For pitch contour analysis, only the voiced portion is utilized. Fig. 6 shows the interpolated pitch contour and voiced segment used for the cepstrum analysis along with the spectrogram of the speech. Other

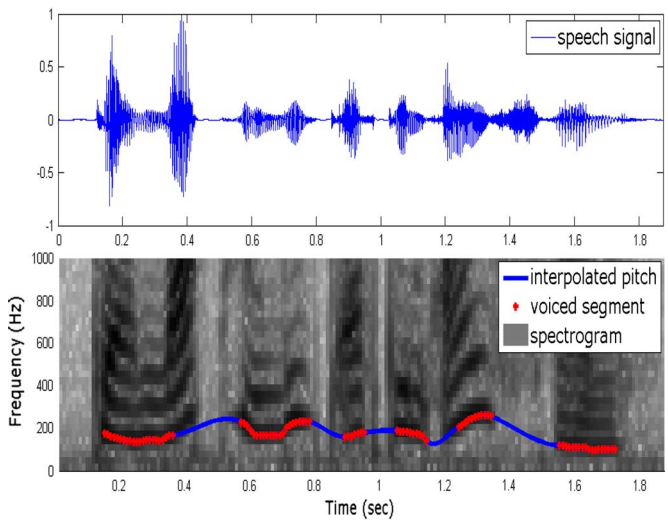


Fig. 6. Pitch contour analysis of an utterance. Further, the contour is upsampled at a sampling frequency of the speech signal and used to calculate the cepstrum coefficient. We only utilize the voiced region (the red crossed marked portion) for final feature extraction.

features that we utilized are 13 MFCCs $C_1 - C_{13}$ with their delta and acceleration components. Input signal is processed using a 30-ms Hamming window with a frame shift interval of 10 ms.

For all of these sequences, the following statistical information is calculated: mean, standard deviation, relative maximum/minimum, position of relative maximum/minimum, first quartile, second quartile (median), and third quartile. The speaking rate is modeled as a fraction of the voiced segments. Thus, the total feature vector per segment contains $3{\cdot}(13{+}13{+}13){\cdot}9{+}1 = 1054$ attributes.

B. Feature Selection

Intuitively, a large number of features would improve the classification performance, however, in practice, a large feature space suffers from the phenomenon of "curse of dimensionality." Therefore, in order to improve the classification performance, a feature selection technique is utilized. This also helps to increase the speed of the system. One such method to eliminate redundant and insignificant features is to identify features with high correlation with the class but low correlation among themselves. In this paper, we used *CFSSubsetEval* feature selection technique provided by WEKA. To determine the best subset, we used a *best-first* search strategy and a *stratified tenfold cross-validation* procedure. Thus, we have ten different sets of selected attributes. An attribute may get selected in $n - number$ of times in different sets. We group the attributes which are selected at least $n - number$ of times and call them "$n - 10$" aggregate. Fig. 7 shows the performance results for different aggregates with a sequential minimal optimization (SMO) classifier provided by WEKA. The "2–10" aggregate provided the best results on EMO-DB. Similarly, for the CVRRCar-AVDB database, the best aggregate set of features were selected for further analysis.

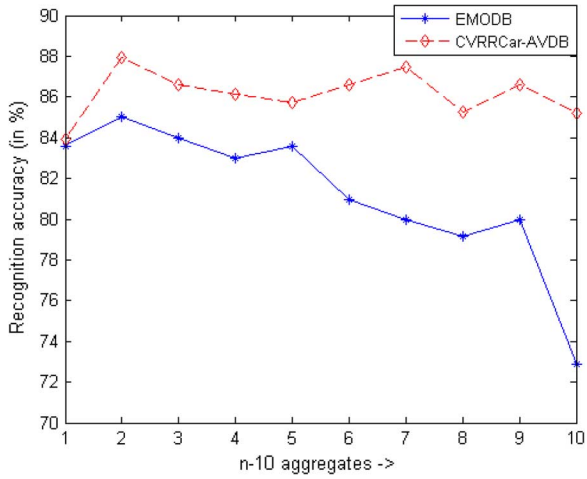


Fig. 7. Correct classification accuracy over seven emotional states in EMO-DB using different aggregate of features and SMO classifier. "$n - 10$" aggregate represents the group of attributes which are selected at least$n - number$ of times in 10 fold feature selection process.

C. Feature Normalization

Feature normalization is a common technique to provide more appropriate attributes to the learning scheme used. In this work, we used the z-score technique. This transforms the original attributes v to new attributes $\hat{v}$ as

$$\hat{v} = \frac{v - v_{\text{mean}}}{v_{\text{std}}}$$

where v_{mean} and v_{std} are mean and standard deviation of v, respectively.

V. Experimental Studies and Analysis

Here, we describe a series of experiments for user- and text-independent analysis for emotion recognition. In all of our experiments, we have used an SVM trained with the SMO algorithm with "2–10" aggregate features. Before these analyses, we provide performance results for randomized tenfold cross validation, that is, the database is divided into ten folds in stratified manner so that they contain approximately the same proportions of labels as the original database, and the system is trained on nine folds and tested on the left out fold. This is repeated ten times, each time leaving out a different fold. The confusion matrix obtained through the cross validation is presented in Table I. The overall recognition rate (weighted accuracy) on this seven-way classification task was 84.5%, whereas average per class accuracy (unweighted accuracy) was 83.1%. This will compare with some of the best performances reported on EMO-DB database. One of the best performances reported so far on EMO-DB is given in [24]. The work in [24] achieved close to 85% weighted accuracy using a similar feature selection technique and SMO classifier. In their experiments, however, they used 494 sentences based on a listening test of 20–30 judges as opposed to 535 used in our experiments. That work further used a discretization technique to improve the results. Our results are comparable to those obtained prior to disctretization. We did not use the discretization step to tailor the system for the database, a route that we wish to avoid for the sake of generality. We believe that, if the database size is sufficiently

TABLE I
CONFUSION TABLE FOR RANDOMIZED TENFOLD STRATIFIED CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	53	0	8	0	2	2	4
dis	2	37	1	5	1	0	0
joy	11	0	41	0	1	0	18
bor	0	0	0	75	3	3	0
neu	1	1	1	6	70	0	0
sad	1	0	0	2	2	57	0
ang	1	0	6	0	1	0	119

Unweighted Accuracy $= 83.1\%$
Weighted Accuracy $= 84.5\%$

TABLE II
CONFUSION TABLE FOR LEAVE-ONE-SUBJECT-OUT CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	45	1	11	0	6	2	4
dis	1	34	3	5	1	2	0
joy	12	3	30	0	1	0	25
bor	0	11	0	58	5	7	0
neu	6	3	1	5	63	1	0
sad	2	1	0	2	3	54	0
ang	4	0	10	0	1	0	112

Unweighted Accuracy $= 72.6\%$
Weighted Accuracy $= 74.8\%$

large, such a step will have little effect on recognition accuracy [34]. Also of note is that our feature set based on cepstrum analysis of pitch and energy contour is quite novel and performed well, despite this being our first attempt to employ these features. Thus, we believe this approach shows great potential for improvement as we begin exploring the parameters of the technique in greater detail. Other reported results on EMO-DB is given in [35], which achieved 83.8% recognition rate using fusion of GMM and HMM based classifiers on fivefold cross validation. However, a conclusion cannot be made about superiority of one method over other for the reason that training and test set while cross validation can be quite different. Either the database should provide a seperate training and test set or a deterministic method of performance evaluation such as *leave-one-out* cross validation should be performed.

Next, we present a series of experiments for user- and text-independent analysis to understand the practical utility of these system in a real-world scenario. For user-independent analysis, we used *leave-one-subject-out* cross validation, where each time system is trained leaving one speaker out of the training set and tested performance on the speaker left out. Table II provides the confusion matrix for this seven-way classification task. Using the same SMO classifier and 2–10 aggregate feature set, the performance measure (unweighted accuracy) drops down to 72.6%. For text-independent analysis, *leave-one-text-out* cross validation is employed where each time the system is trained leaving all the utterances with same spoken content across speaker out of the training set and testing on the left out utterances. Table III shows the confusion matrix obtained for the analysis. Unweighted accuracy, in this case, is still over 80%. These analyses suggest that the classifier

TABLE III
CONFUSION TABLE FOR LEAVE-ONE-TEXT-OUT CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	51	0	9	0	3	2	4
dis	2	39	0	3	1	0	1
joy	8	0	39	0	1	0	23
bor	0	0	0	74	3	4	0
neu	6	2	0	7	63	1	0
sad	2	1	0	3	1	55	0
ang	3	1	8	0	0	0	115

Unweighted Accuracy $= 80.6\%$
Weighted Accuracy $= 81.4\%$

TABLE IV
GENDER-DEPENDENT CONFUSION TABLE FOR RANDOMIZED TENFOLD CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	53	1	9	0	2	1	3
dis	2	37	0	4	2	0	1
joy	8	0	43	0	1	1	18
bor	1	2	0	75	2	1	0
neu	1	1	0	4	73	0	0
sad	0	0	0	2	1	59	0
ang	3	0	10	0	0	0	114

Unweighted Accuracy $= 84.0\%$
Weighted Accuracy $= 84.9\%$

TABLE V
GENDER=DEPENDENT CONFUSION TABLE FOR LEAVE-ONE-SUBJECT-OUT CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	43	2	9	0	4	2	9
dis	2	35	0	4	3	1	1
joy	16	1	29	0	2	1	22
bor	0	12	0	58	6	5	0
neu	4	2	1	9	63	0	0
sad	1	0	0	4	1	56	0
ang	4	1	15	0	0	0	107

Unweighted Accuracy $= 72.2\%$
Weighted Accuracy $= 73.1\%$

is learning a particular manner that a speaker might express his/her emotions while ignoring the verbal content to some degree. This is a very important observation for a successful design of a practical system, which suggests a need for speaker adaptation in speech emotion recognition task.

Other experiments include the analyses of context like gender information which can be reliably recognized automatically. Performance results based on randomized tenfold cross validation, *leave-one-subject-out* and *leave-one-text-out* cross validation on EMO-DB speech corpus for gender-dependent analysis (male–female combined performance) is summarized in Tables IV, V, and VI, respectively. Tables VII and VIII provide gender-dependent and -independent analysis on CVRRCar-AVDB for randomized tenfold cross validation. From the analysis, it is evident that using gender information in emotion recognition leads to significant improvement in recognition rate. For EMO-DB, text-independent analysis has relative improvement of over 3%. Similarly, for CVRRCar-AVDB relative improvement of over 3% is quite encouraging.

TABLE VI
GENDER-DEPENDENT CONFUSION TABLE FOR LEAVE-ONE-TEXT-OUT CROSS VALIDATION

Reference Emotion	Recognized Emotion						
	fea	dis	joy	bor	neu	sad	ang
fea	54	2	5	0	3	1	4
dis	2	38	0	4	1	0	1
joy	8	0	42	0	1	1	19
bor	2	2	0	73	2	2	0
neu	3	1	0	6	69	0	0
sad	0	0	0	2	0	60	0
ang	4	0	11	0	0	0	112

Unweighted Accuracy $= 83.21\%$
Weighted Accuracy $= 83.7\%$

TABLE VII
GENDER-INDEPENDENT CONFUSION TABLE FOR RANDOMIZED TENFOLD STRATIFIED CROSS VALIDATION

Reference Emotion	Recognized Emotion		
	pos	neu	neg
pos	76	5	1
neu	9	64	9
neg	1	4	55

Unweighted Accuracy $= 87.9\%$
Weighted Accuracy $= 88.5\%$

TABLE VIII
GENDER-DEPENDENT CONFUSION TABLE FOR RANDOMIZED TENFOLD STRATIFIED CROSS VALIDATION

Reference Emotion	Recognized Emotion		
	pos	neu	neg
pos	80	1	1
neu	7	72	3
neg	3	3	54

Unweighted Accuracy $= 91.79\%$
Weighted Accuracy $= 91.96\%$

VI. DISCUSSION AND CONCLUSIONS

We presented a systematic study to understand the importance of the context in emotion recognition from speech. We also introduced a novel feature set based on cepstrum analysis of pitch and speech-intensity contours. Using this feature set, we were able to achieve a high recognition rate of over 84% for the seven emotions on EMO-DB and over 87% for the three groups of emotions on CVRRCar-AVDB using tenfold stratified cross validation. We then performed user- and text-independent analysis on EMO-DB database. The results suggest that the system is robust towards the spoken containt, however, it performs significantly better when speaker information is incorporated into the training set. Furthermore, we analyzed the use of gender-based context information on recognition rate over the two databases. Experimental results suggest that a gender-specific emotion recognizer works more accurately than a gender-independent one. In future, we will extend our basic emotion recognition system by a preceeding stage of automatic gender-detection system that would determine which gender-specific emotion recognition system should be used. However, there is still much room for improvement. Our initial experiments with vision modality and audiovisual analysis have shown promising results. Although these two modalities do not couple strongly in time as also shown in Fig. 2, they seem to complement each other [36]. In some cases, similar facial expressions may have different vocal characteristics, and vocal emotions having similar properties may have different facial behaviors. Our initial observations also suggest that head dynamics carry useful information for emotion classification. We will thoroughly present such multimodal affect recognition analysis in our future efforts.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and editors for their insightful comments and helpful suggestions. The authors would also like to thank their colleagues at the Computer Vision and Robotics Research Laboratory, University of California, San Diego, for useful discussions and assistance.

REFERENCES

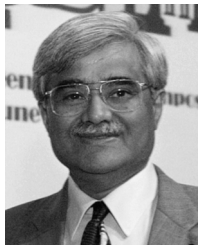
[1] G. Roisman, J. Tsai, and K. Chiang, "The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview," *Devel. Psychol.*, vol. 40, no. 5, pp. 776–789, 2004.
[2] J. F. Cohn and E. Z. Tronick, "Mother-infant face-to-face interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior," *Devel. Psychol.*, vol. 23, pp. 68–77, 1998.
[3] P. Ekman, D. Matsumoto, and W. Friesen, "Facial expression in affective disorders," *What the Face Reveals*, pp. 429–439, 2005.
[4] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," in *Face Recognition*, K. Delac and M. Grgic, Eds. Vienna, Austria: I-Tech Education, 2007, pp. 377–416.
[5] C. L. Lisetti and F. Nasoz, "Maui: A multimodal affective user interface," in *Proc. 10th ACM Int. Conf. Multimedia*, New York, 2002, pp. 161–170.
[6] Z. Duric, W. Gray, R. Heishman, F. Li, A. Rosenfeld, M. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proc. IEEE*, vol. 90, no. 7, pp. 1272–1289, Jul. 2002.
[7] L. Maat and M. Pantic, "Gaze-x: Adaptive affective multimodal interface for single-user office scenarios," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2006, pp. 171–178.
[8] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Human-Comput. Stud.*, vol. 65, no. 8, pp. 724–736, 2007.
[9] T. Zhang, H. M. Johnson, and S. E. Levinson, "Children's emotion recognition in an intelligent tutoring scenario," in *Proc. ICSLP*, 2004.
[10] O.-W. Kwon, K.-L. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, 2004, pp. 172–187.
[11] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human–computer dialog," in *Proc. ICSLP*, 2002, pp. 2037–2040.
[12] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr, "On the necessity and feasibility of detecting a driver's emotional state while driving," in *Proc. ACII*, 2007, pp. 126–138.
[13] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for mexi," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Aug. 2005, pp. 1138–1144.
[14] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[15] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Commun.*, vol. 40, no. 1–2, pp. 117–143, 2003.
[16] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc 9th Int. Conf. Spoken Language Processing (ICSLP)*, 2006, pp. 801–804.
[17] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. ICASSP*, Washington, DC, 1999, pp. 2079–2082.
[18] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J Pers. Soc. Psychol.* vol. 70, no. 3, pp. 614–636, Mar. 1996 [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/8851745

[19] T. Scherer, K. R. Johnstone, and T. Bänziger, "Automatic verification of emotionally stressed speakers: The problem of individual differences," in *Proc. Int. Workshop Speech Comput.*, St. Petersburg, Russia, 1998.
[20] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Commun.*, vol. 50, no. 5, pp. 416–433, 2008.
[21] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proc. ACL*, Morristown, NJ, 2004, p. 351.
[22] K. Forbes-Riley and D. Litman, "Predicting emotion in spoken dialogue from multiple knowledge sources," in *Proc. HLT-NAACL*, D. M. S. Dumais and S. Roukos, Eds., Boston, MA, May 2–7, 2004, pp. 201–208.
[23] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. LREC*, Genoa, Italy, 2006, pp. 1123–1126.
[24] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Proc. Int. Conf. Semantic Computing*, Aug. 2008, pp. 158–165.
[25] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Network*, vol. 18, no. 4, pp. 407–422, 2005.
[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, D. M. S. Dumais and S. Roukos, Eds., 2005, pp. 1517–1520.
[27] M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Trans. Intell. Transportation Syst.*, pp. 108–120, Mar. 2007.
[28] M. M. Trivedi and S. Y. Cheng, "Holistic sensing and active displays for intelligent driver support systems," *IEEE Computer,* Special Issue on Human-Centered Computing, vol. 40, no. 5, pp. –68, May 2007.
[29] S. T. Shivappa, B. D. Rao, and M. Trivedi, "Role of head pose estimation in speech acquisition from distant microphones," in *Proc. IEEE ICASSP*, Apr. 2009, no. 1, pp. 3557–3560.
[30] M. M. Trivedi, K. S. Huang, and I. Mikic, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Trans. Syst., Man, Cybern. A*, vol. 35, pp. 145–163, 2005.
[31] A. Tawari and M. M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *Proc. IEEE Intell. Vehicles Symp*, Jun. 2010, pp. 174–178.
[32] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real world environment," in *Proc. 20th ICPR*, 2010.
[33] B. Paul, Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound," 1993 [Online]. Available: citeseer.ist.psu.edu/boersma93accurate.html
[34] T. Vogt and E. André, "Exploring the benefits of discretization of acoustic features for speech emotion recognition," in *Proc. 10th Conf. INTERSPEECH*, Brighton, U.K., Sep. 2009, pp. 328–331.
[35] E. Bozkurt, C. Erdem, E. Erzin, T. Erdem, M. Ozkan, and A. Tekalp, "Speech-driven automatic facial expression synthesis," in *Proc. 3DTV Conference: The True Vision—Capture, Transmission and Display of 3-D Video*, May 2008, pp. 273–276.
[36] S. Shivappa, M. M. Trivedi, and B. Rao, "Audio-visual information fusion in human computer interfaces and intelligent environments: A survey," *Proc. IEEE*, to be published.

Ashish Tawari received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 2006, and is currently working toward the Ph.D. degree at the University of California at San Diego (UCSD), La Jolla.

He served as a DSP Engineer with Qualcomm, India during 2006–2008. He is currently with Qualcomm Inc., San Diego, CA, under a summer cooperative internship program with the Multimedia R&D Speech team. His research interests lie in the areas of multimodal signal processing, machine learning, Speech and audio processing and computer vision.

Mr. Tawari was the recipient of the Powell Fellowship 2008–2011 at UCSD. His thesis proposal, advised by Mohan Trivedi, received an Honorable Mention at the 2010 IEEE Intelligent Vehicles Symposium Ph.D. Forum.



Mohan Manubhai Trivedi (F'08) received the B.E. degree (with honors) from the Birla Institute of Technology and Science, Pilani, India, and the Ph.D. degree from Utah State University, Logan.

He is a Professor of electrical and computer engineering and the Founding Director of the Computer Vision and Robotics Research Laboratory, University of California, San Diego (UCSD), La Jolla. He has established the Laboratory for Intelligent and Safe Automobiles (LISA), UCSD, to pursue a multidisciplinary research agenda. He and his team are currently pursuing research in machine and human perception, active machine learning, distributed video systems, multimodal affect and gesture analysis, human-centered interfaces, intelligent driver assistance systems. He regularly serves as a consultant to industry and government agencies in the U.S. and abroad. He has given over 55 keynote/plenary talks. He served as the Editor-in-Chief of the *Machine Vision and Applications* journal.

Prof. Trivedi is a Fellow of the International Association for Pattern Recognition (IAPR) and the International Society for Optical Engineers (SPIE). He is currently an editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS AND IMAGE AND VISION COMPUTING. He served as the General Chair for IEEE Intelligent Vehicles Symposium IV 2010. He was the recipient of the Distinguished Alumnus Award from Utah State University, the Pioneer Award, the Meritorious Service Award from the IEEE Computer Society, and several Best Paper awards.