Robust Speech Based Emotion Analysis for Driver Assistance System

Ashish Tawari and Mohan Trivedi
LISA: Laboratory for Intelligent and Safe Automobiles
University of California San Diego, Dept. of ECE
atawari@ucsd.edu, mtrivedi@ucsd.edu

Abstract—Automated analysis of human affective behavior has attracted increasing attention in recent years. In context of driving task, particularly, driver’s emotion often influences driving performance which can be improved if the car actively responds to the emotional state of the driver. It is important for an intelligent driver support system to accurately monitor the driver’s state in an unobtrusive and robust manner. Ever changing environment while driving poses a serious challenge to existing techniques for speech emotion recognition. In our research, we propose to utilize contextual information of the outside environment as well as inside car user to improve the emotion recognition accuracy. In particular, a noise cancellation technique is proposed to suppress the noise adaptively based on the driving context and use gender based context information for developing the classifier.

I. Research Objective and Motivation

The automotive industry is integrating more and more technologies into modern cars. With potentially more complex devices for the driver to control, risk of distracting driver’s attention increases. Current research and attention theory have suggested that speech-based interactions are less distracting than interaction with visual display [1]. Therefore, to improve both comfort and safety in the car, the driver assistance technologies need effective speech interface between the driver and the infotainment system. The introduction of speech-based interactions and conversation into the car brings out potential importance of linguistic cues (like word choice and sentence structure) and paralinguistic cues (like pitch, intensity, and speech rate etc.). Such cues play a fundamental role in enriching human-human interaction and incorporate among other things, personality and emotion [2]. Driving in particular presents a context in which a user’s emotional state plays a significant role. Emotions have been found to affect cognitive style and performance. The road rage phenomenon [3] is an undeniable fact of the impact that emotion can have on the safety of the roadways. Such phenomenon can be mitigated and driving performance can be improved if the car actively responds to the emotional state of the driver. Research studies show that matching in-car voice and driver’s emotional state has great impact on driving performance [4]. While number of studies have shown the potential of emotion monitoring, automatic recognition of emotion is still a very challenging task, specially when it comes to real world driving scenario. The car setting is far from the ideal environment for speech acquisition and processing need to deal with reverberation and noisy conditions inside car cockpit.

Fig. 1 shows the block diagram representing the three fundamental steps of information acquisition and preprocessing, extraction and processing of parameters and classification of semantic units. The pre-processing stage consists of speech enhancement module which adopts to the context of the outside environment. 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, testing phase (phase 3) to evaluate the performance of the system in terms of classification accuracy.
III. COMPUTATIONAL MODULES

A. Adaptive speech enhancement

During last decades, several speech enhancement techniques have been proposed ranging from beamforming through microphone arrays to adaptive noise filtering approaches. In this work, we utilize a speech enhancement technique based on the adaptive thresholding in wavelet domain [7]. Objective criterion for parameter selection, however, used in our experiments was the classification performance as opposed to SNR.

B. Feature extraction and selection

A variety of features have been proposed to recognize emotional states from speech signal. These features can be categorized as acoustic features and linguistic features. We avoid using the latter since these demand for robust recognition of speech in first place and also are a drawback for multi-language emotion recognition. Hence we only consider the acoustic features. Within the acoustic category, we focus on prosodic like speech intensity, pitch and speaking rate, and spectral features like mel frequency cepstral coefficients (MFCC) to model emotional states.

In order to capture the characteristics of the pitch and energy contours, we perform cepstrum analysis over the contour [6]. For all these sequences following statistical information is calculated: mean, standard deviation, relative maximum/minimum, position of relative maximum/minimum, 1st quartile, 2nd quartile (median) and 3rd quartile. Speaking rate is modeled as fraction of the voiced segments. Thus, the total feature vector per segment contains \(3 \cdot (13 + 13 + 13) \cdot 9 + 1 = 1054\) attributes.

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. One such method to eliminate redundant and insignificant features is to identify features with high correlation with the class but low correlation among themselves [8]. We also show that including gender information during feature selection significantly improves the performance.

IV. CLASSIFIER DESIGN

Support Vector Machines (SVMs) have shown great performance in practice due to their various attractive properties including generalization performance. For the task at our hand, we use SVM with linear kernel and one-vs-one multiclass discrimination scheme. Classification is done by a max-wins voting strategy.

V. IN-VEHICLE DATA COLLECTION

Collecting real world driving data is certainly useful for researchers interested in design and development of intelligent vehicular system capable of dealing with different situations in traffic. Real world data, as opposed to driving simulator recordings, are much more closely related to the day to day situations and can provide valuable information on drivers’ behavior.

A. LISA audio-visual affect database

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. We also recorded free conversion between passenger and driver. The database is collected in stationary and moving car environment. The cockpit of 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 as well as challenging illumination condition for video channel. In this study, we analyze emotional speech data in the 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 2 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 16 kHz sampling rate. The emotional speech has been labeled into 3 groups ‘pos’, ‘neg’ and ‘neu’ for positive, negative and neutral expressions. The data has been acquired with 4 different subjects: 2 male and 2 female. Distribution of data for different categories is: 82 pos, 82 neg, and 60 neu.

REFERENCES