# 1

# Visual and acoustic features based emotion detection for advanced driver assistance system

H. D. Vankayalapati[1], K. R. Anne[2] and K. Kyamakya[1]

[1]University of Klagenfurt, Austria.
[2]VR Siddhartha Engineering college, India

**Summary.** Poor attention of drivers towards driving can cause accidents that can harm the driver or surrounding people. The poor attention is not only caused by the drowsiness of the driver but also due to the various emotions/moods (for example sad, angry, joy, pleasure, despair and irritation) of the driver. The emotions are generally measured by analyzing either head movement patterns or eyelid movements or face expressions or all the lasts together. Concerning emotion recognition visual sensing of face expressions is helpful but generally not always sufficient. Therefore, one needs additional information that can be collected in a non-intrusive manner in order to increase the robustness of the emotion measurement in the frame of a non-intrusive monitoring policy. We find acoustic information to be appropriate, provided the driver generates some vocal signals by speaking, shouting, crying, etc. In this paper, we propose a decision level fusion technique, to fuse the combination of visual sensing of face expressions and pattern recognition from driver's voice. The result of the proposed approach significantly increase the performance of the automatic driver emotion recognition system.

## 1.1 Introduction

Driving is one of the most dangerous tasks in our everyday lives. In 2001, according to the Australian government Road Traffic Report, 20% of all accident or major crashes are due to the driver behavior. In trucking industry, 57% of the truck accidents are also due to the driver fatigue [6]. This shows that the driving scenario needs supervision. Here manual supervision is impractical or impossible, and drivers must monitor themselves to ensure that they do not fall asleep and inattentive. The supervision is more important in case of commercial drivers, who drive large vehicles for long periods of time. The drivers may be working at night. Recent research shows that six out of ten crashes are due to the late reaction (fraction of second) of the driver. Therefore, to improve road safety, we need to control, record and monitor the driver status and behavior related parameters. So the emotion recognition is a growing field in developing friendly human-computer interaction systems. Thus, the necessity of the driver monitoring systems is increasing day by day.

Driver monitoring plays a major role in order to assess, control and predict the driver behavior. The research concerning driver monitoring systems was started

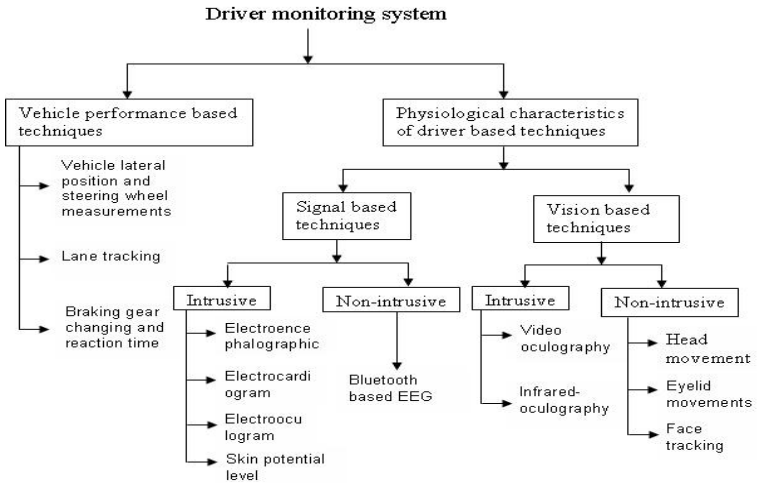


**Fig. 1.1.** General classification of driver monitoring systems

nearly from the 1980's. The driver monitoring systems can be mainly classified as shown in Fig. 1.1. In the first stages of this research, researchers developed driver monitoring systems based on inferring both driver behavior and state from the observed/measured vehicle performance. However, these indirect approaches heavily depend upon vehicle and road conditions (e.g. quality of lane markings, alternate lane markings during road repairs) as well as on environmental conditions (e.g. shadow, rain and night vision) [12]. These drawbacks have drawn the researcher's interest to directly monitoring the driver behavior. Thus, a second class of approaches does directly measure driver physiological characteristics but in an intrusive way by involving measurement systems such as the Electroencephalogram (EEG) which monitors brain activities, the Electrocardiogram (ECG) which measures heart rate variation, the Electrooculogram (EOG) which monitors eye movement, the skin potential level measurement techniques, etc [4]. These methods of the second class of approaches do need the driver's cooperation as the electrodes are attached directly to the driver's body. Due to an expected very limited user acceptance of these intrusive methods in normal vehicles, they are more realistic for a daily use rather only in health care or similar special vehicles. A further problem is that the intrusive apparatus involved in these methods may itself contribute to the driver's distraction and fatigue.

And more recently, a significant research has been focusing on developing non-intrusive techniques. These Non-intrusive approaches generally involve machine vision as an alternative to a direct measurement of physiological characteristics and they do not need any cooperation from the driver; they monitor the driver behavior and status directly through visual sensors [16]. Video sensors are placed on the dash board to measure, for example, eyelid movements (open/close interval of eyelid), head movements, mouth movements (yawning) and face expression.

The first investigations to emotion recognition from speech were conducted around the mid of the 1980s using statistical properties of certain acoustic features [19]. Later, the evolution of computer architectures introduced the recognition of

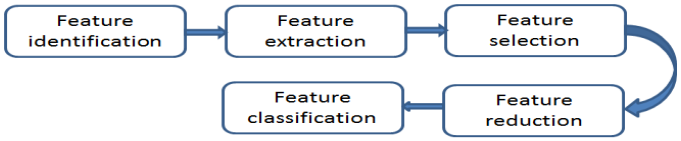


**Fig. 1.2.** The overall architecture of the emotion recognition system

more complicated emotions from the speech. Certain features in the voice of a person can be used to infer the emotional state of the particular speaker. The real-time extracting the voice characteristics conveys emotion and attitude in a systematic manner and it is different from male and female [3]. The research towards detecting human emotions is increasingly attracting the attention of the research community. Nowadays, the research is focused on finding powerful combinations of classifiers that increase the classification efficiency in real-life speech emotion recognition applications. Some of these techniques are used to recognize the frustration of a user and change their response automatically.

By using these multidimensional features, we recognize the emotions such as drowsiness (sleepy), fatigue (lack of energy) and emotions/stress (for example sad, angry, joy, pleasure, despair and irritation). In this work, we recognize emotions based on the visual and acoustic features of the driver. Here we calculate the visual emotion and acoustic emotion separately and fuse them by using linear distance measure.

## 1.2 Feature based emotion recognition

Automatic emotion recognition plays a major role in human computer interaction and speech processing. Facial expressions and speech characteristics of the driver form as crucial information in assessing the emotion of the driver. The overall approach of emotion recognition is illustrated in Fig. 1.2. As shown in Fig. 1.2, identifying the important features which can improve the performance of recognition systems is a key issue. Generally in case of visual features, features are classified as local features and global features. local features means eyes, nose, mouth etc and global features means transformation coefficients of global image decomposition. after identifying the features, appropriate feature extraction and feature selection is essential for achieving good performance in emotion recognition. After feature extraction, high dimensional feature vector is obtained from the visual and acoustic information. So we have to reduce the dimensionality of the feature vector by using dimensional reduction technique like PCA and LDA. By using these low dimensional feature vector, we classify the emotion from the visual and acoustic features separately. By combining these results at decision level, we estimate the emotion. The overall approach of emotion recognition algorithm is shown in Fig. 1.2.

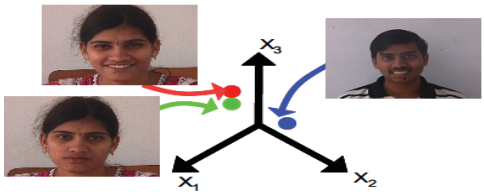


**Fig. 1.3.** Image representation in the high dimensional space

## 1.3 Feature extraction

### 1.3.1 Visual Feature extraction

In this work, we mainly concentrate on the global features of the driver's face. In general, all emotion recognition algorithms use any one or the combinations of the local and global features namely shape, texture, color, or intensity to represent the facial image structure. It has been seen from previous works that the appearance based representations that uses the intensity or pixel values produces the better result compared with other techniques [8]. In these techniques, driver face images are stored as two dimensional intensity matrix. The vector space contains different face images and each point in the vector space represents an image as shown in Fig. 1.3. Almost all appearance based techniques use statistical properties like mean and covariance to analyze the image.

### 1.3.2 Acoustic Feature extraction

Humans recognize emotions by observing what we say and how we say it. Here "how" is even more important than the "what". Many features are present in acoustic information. The important acoustic features for emotion recognition are pitch, zero crossing rate, short time energy, Mel Frequency Cepstral Coefficients (MFCCs) etc.

The architecture of the emotion recognition system based on acoustic features is shown in Fig. 1.4. The architecture depicts the process of transforming given input speech signals to driver emotions.

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

The main goal of pre-emphasis is to boost the amount of energy in the higher frequencies with respect to lower frequencies. Mainly boosting is used to get more information from the higher frequencies available to the acoustic model and to improve the recognition performance [1].This pre-emphasis is done by using a first-order high-pass filter.

**Frame Blocking:** When we analyze audio signals or speech, most of the audio signals are more or less stable within a short period of time. When we do frame blocking, there may be some overlaps between neighboring frames to capture subtle change in the audio signals [5]. For frame blocking, windowing operation is used. In

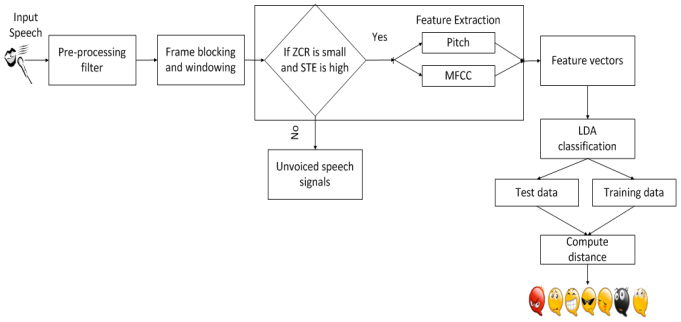


**Fig. 1.4.** The overall architecture of the acoustic emotion recognition system

the window operation, the large input data is divided into small data sets and stored in sequence of frames. While dividing, some of the input data may be discontinuous. In order to keep the continuity of the first and the last points in the frame(to reduce the spectral leakage in the input data) hamming window method is used.

**Feature Extraction:** Features are extracted from the real time data by performing time and frequency domains algorithms. These algorithms extract temporal features, spectral features. These features are extracted based on the amplitude and spectrum analyzer of the audio data. After windowing, we perform the feature extraction methods for estimating the acoustic features that are mostly used in emotion detection.

**Zero-crossing rate:** Zero-crossing rate is measure of number of times the amplitude of the speech signals passes through a value of zero in a given time interval/frame. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced [19].

**Short Time Energy:** The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations.

A reasonable generalization is that if the Short time energy is high, the speech signal is voiced, while if the Short time energy is low, the speech signal is unvoiced. Based on zero crossing rate and short time energy, voiced sounds are identified. We can extract the following features from the identified voice speech signal.

**Pitch:** Pitch is the fundamental frequency of audio signals, which is equal to the reciprocal of the fundamental period [18]. This is mainly explained in terms of highness or lowness of a sound. Pitch in reality can be defined as the rate at which peaks in the autocorrelation function occur. Autocorrelation function is used to estimate pitch, directly from the waveform. xcorr function is used to estimate the statistical cross-correlation sequence of random process. We can estimate the fundamental frequency by using autocorrelation function, peaks at delay intervals corresponding to the normal pitch range in speech,

**Mel frequency cepstral coefficient (MFCC):** MFCC are the most widely used spectral representation of speech. MFCC is based on human hearing percep-

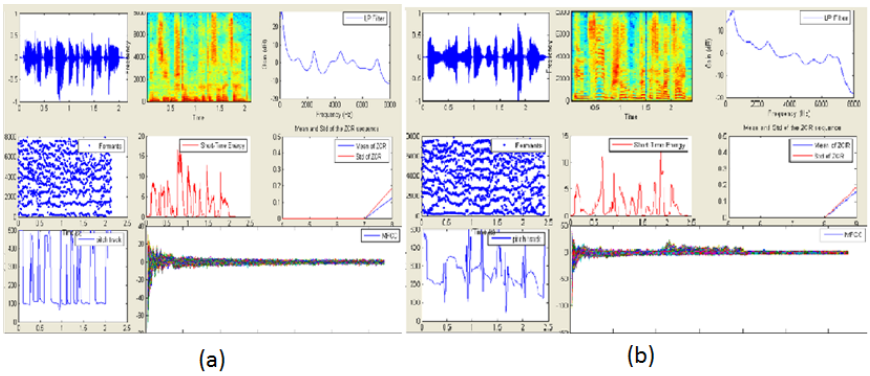


**Fig. 1.5.** Illustration of audio feature extraction (a) Extracted features from the sad emotional audio file (b) Extracted features from the happy emotional audio file

tions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [10]. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. It turns out that humans perceive sound in a highly nonlinear way. Basic parameters like pitch and loudness highly depend on the frequency, adding weight to components at lower frequencies. MFCC consists of several computational steps. MFCC is the most widely used spectral representation of speech. MFCC parameters are calculated by taking the absolute value of the FFT, warping it to a Mel frequency scale, taking the DCT of the log-Mel spectrum and returning the first 13 (12 cepstral features+energy) coefficients [9]. The variation in the different acoustic features for different emotions are shown in Fig. 1.5.

## 1.4 Feature reduction

The performance of emotion recognition heavily depends upon the quality and size of the extracted feature set from visual and acoustic information of the driver. The appearance based linear subspace techniques use the statistical properties like the mean and variance of the image/audio [15]. The dimensionality of the feature set is reduced by using these statistical techniques.

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the appearance based linear subspace techniques. Among these techniques, LDA gives high recognition rate and speed when compared with PCA [8]. So LDA is used for dimensionality reduction from both visual and acoustic features.

The emotion detection needs a database with a significant number of variables. This means a high dimensionality database is required [8]. This high dimensionality database contains more similar features. In such situations, we need to reduce the dimensionality by only selecting the non-correlated features (information loss is very

less) from the database. Linear Discriminant Analysis (LDA) is one of the important and popular dimensionality reduction technique [15].

**Linear discriminant analysis (LDA):** The main objective of LDA is minimizing the within class variance and maximizing the between class variance in the given data set [15, 11]. In other words it groups the same class wave files and separates the different class wave files. A class means the collection of data belonging to same object or person. LDA finds the optimal transformation matrix as to preserve most of the information that can be used to discriminate between the different classes. The LDA helps for better understanding of feature data [15].

In order to find the best match, we make use of the distance measure classifier. The training set feature vector with least distance gives the best match emotion with the test sample. The Euclidean distance is commonly used linear distance measure classifier in many applications. This distance gives the shortest distance between the two sample files or vectors [17]. But this is sensitive to both adding and multiplying the vector with some factor or value. So in this section we used a special nonlinear metric which is able to compute the distance between different sized matrices having a single common dimension, like the visual/acoustic matrices representing our sample feature vectors. It derives from the Hausdorff metric for sets [13, 14].

**Hausdorff distance:** The Hausdorff distance (HD) is a non-linear operator, which measures the mismatch of the two sets. The Hausdorff distance measures the extent to which each point of a 'model' set lies near some point of an 'sample' set and vice versa. Unlike most vector comparison methods, the Hausdorff distance is not based on finding corresponding mode and speech points. Thus, it is more tolerant of perturbations in the location of points because it measures proximity rather than exact superposition [2].However, the Hausdorff distance is extremely sensitive to outliers.

The distance between two points a and b is defined as $d(a,b) = ||a-b||$. Here, we not only compute the distance between the point $a$ in the finite point set A and the same value $b$ in the finite point set $B = b_1, ..., b_{N_b}$, but also compute the distances between the $a_t$ and its two neighbor values $b_{t-1}$ and $b_{t+1}$ in the finite point set B, respectively, and then minimize these three distances as shown in in Equation (1.1) [14].

$$d(a,B) = \min_{b \in B} d(a,b) = \min_{b \in B} ||a-b|| \tag{1.1}$$

The directed Hausdorff metric h(A,B) between the two finite point set $A = a_1, ...., a_{N_a}$ and $B = b_1, ...., b_{N_b}$is defined in Equation (1.2,1.3) :

$$h(A,B) = \max_{a \in A} d(a,B) = \max_{a \in A} \min_{b \in B} d(a,b) \tag{1.2}$$

$$h(A,B) = \left\{ \max_{a \in A} \left\{ \min_{b \in B} ||a-b|| \right\} \right\} \tag{1.3}$$

## 1.5 Feature set classification

### 1.5.1 Classification based on visual features

The recognition performance has been systematically evaluated by using different sizes of the database with differen appearance based techniques like PCA and LDA.

The results of the evaluation have shown that the recognition rate of LDA is considerably increased when compared to PCA. The performance of these feature extraction approaches are systematically evaluated in our previous work over FERET database for face recognition application [8].

### 1.5.2 Classification based on acoustic features

From the literature emotion recognition based on acoustic information has been implemented on a variety of classifiers including maximum likelihood classifier (MLC), neural network (NN), k-nearest neighbor (k-NN), Bayes classifier, support vector classifier, artificial neural network (ANN) classifier and Gaussian mixture model (GMM) etc [19].

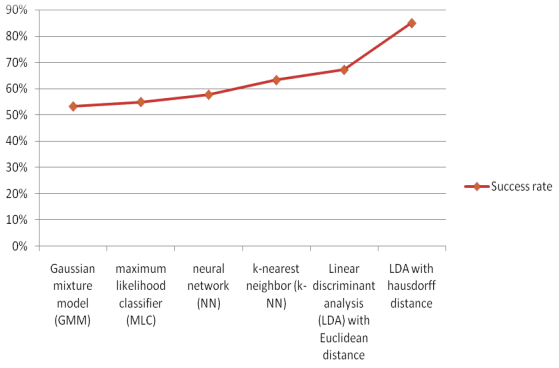


**Fig. 1.6.** Graphical representation of success rate of different classifiers

The LDA performs considerably better when compared to above classifiers for Berlin emotional database (EMO-DB database) [7]. But the performance of LDA with Euclidean distance is also not sufficient for real world applications. In order to improve the performance (success rate and process speed), we propose the nonlinear Hausdorff metric based LDA. By considering the Hausdorff distance measure instead of the linear Euclidean distance measure, the success rate of the LDA algorithm is increased by around 20% as shown in Fig. 1.6.

## 1.6 Multi-dimensional feature fusion

Emotions can be classified into discreet classes (like anger, happiness, disgust or sadness). Neutral emotion means no expression/emotion is present. In this work, we classify different expression based on neutral as shown in Fig. 1.7.

Features can be fused at different levels i.e., after feature selection, after feature reduction, and at decision level. In this work, the major focus is at identifying the emotion of the driver in a real world scenario. In real world scenario, acoustic information is present in bursts based on the mood of the driver, where as visual

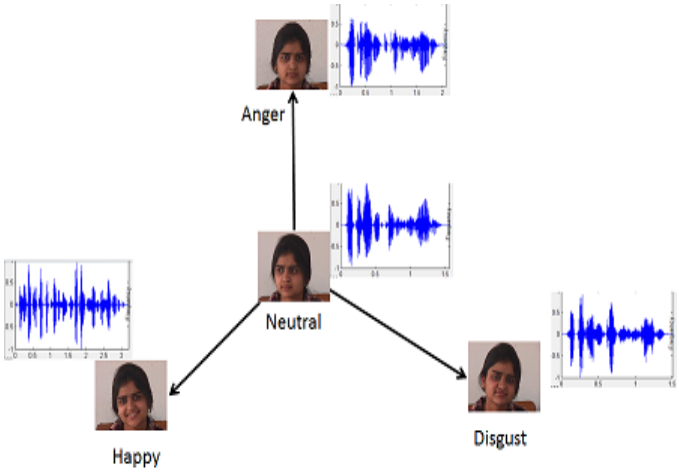


**Fig. 1.7.** Illustration of fusion based emotion classification

information is present throughout. By considering this aspect, we proposed the feature fusion at the decision level. To validate the performance, the probability for each of the emotions was calculated for the audio and visual features separately and were multiplied to get the final result and projected emotional vector space as shown in Fig. 1.7. In order to evaluate the recognition algorithm with fused features, we

| Data set | Success rate with LDA |
|---|---|
| only acoustic | 87% |
| only visual | 79% |
| 40 acoustic + 40 visual | 96% |
| 20 acoustic + 40 visual | 92% |

**Table 1.1.** Performance evaluation of LDA with different data sets

have used the Berlin emotional database for acoustic information and Indian face database for visual emotional information. The performance evaluation of the LDA over different data sets is shown in Table 1.1.

## References

1. Tobias Andersson. Audio classification and content description. Master's thesis, Lulea University of Technology, Multimedia Technology, Ericsson Research, Corporate unit, Lulea, Sweden, March 2004.
2. T. Barbu. Discrete speech recognition using a hausdorff-based metric. In *Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks, ICETE 2004*, volume 3, pages 363–368, Setubal, Portugal, Aug 2004.

3. R. Van Bezooijen. The characteristics and recognizability of vocal expression of emotions. Foris, Drodrecht, The Netherlands, 1984.
4. A. Broggi. Vision-based driving assistance. *IEEE Intelligent Transport Systems*, 13:22– 23, 1998.
5. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32 –80, jan 2001.
6. Regulation (EC) No 561-2006 EU. of the european parliment and of the council of 15th march 2006 on the hramoniyation of certain social legislation relating to road transport and amending council regulations (eec). *Offical journal of the european union*, 102, 2006.
7. M. Rolfes W. Sendlmeier F. Burkhardt, A. Paeschke and B. Weiss. A database of german emotional speech. In *Interspeech*, pages 1517–1520, 2005.
8. H.D.Vankayalapati. Nonlinear feature extraction approaches for scalable face recognition applications. In *ISAST Transactions on Computers and Intelligent Systems*, volume 2, 2010.
9. K.Kyamakya H.D.Vankayalapati, K. Anne. Extraction of visual and acoustic features of the driver for monitoring driver ergonomics applied to extended driver assistance systems. volume 81, pages 83–94. Springer Berlin / Heidelberg, 2010.
10. J. Arnott I. Murray. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.
11. M. Nikravesh I.M. Guyon, S.R. Gunn and L. Zadeh. *Feature Extraction, Foundations and Applications*. Springer, 2006.
12. Chen L. Fletcher, Apostoloff and Zelinsky. Computer vision for vehicle monitoring and control. pages 67–72, Sydney, 2001.
13. A. K. Jain M. P. Dubuisson. Pattern recognition - conference a: Computer vision image processing., proceedings of the 12th iapr international conference on. volume 1, pages 566 –568, 1994.
14. Klaus J. Kirchberg Oliver Jesorsky and Robert W. Frischholz. Robust face detection using the hausdorff distance. *Third International Conference on Audio- and Video-based Biometric Person Authentication*, page 9095.
15. J. P.Hespanha P. N.Belhumeur and D. J.Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.
16. Lan P. Qiang Ji, Zhiwei Zhu. Real time non-intrusive monitoring and prediction of driver fatigue. *Vehicular Technology, IEEE Transactions*, 53:1052 – 1068, 2004.
17. V. Perlibakas. Distance measures for pca-based face recognition. *Pattern Recogn. Lett.*, 25(6):711–724, 2004.
18. Johannes Wagner Thurid Vogt, Elisabeth Andr. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. pages 75–91, 2008.
19. Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.