Identification of Emotion and Age from Audio Signal

Aparna Pradhan
Registration No. - 137327 of 2016-2017

Examination Roll No. - MCA196017

Supervisor: **Prof. Sanjoy Kumar Saha**

Department of Computer Science & Engineering
Jadavpur University
Kolkata - 700032

This project report is submitted for the partial fulfillment of the degree of
Master of Computer Application

May 2019

Declaration

I hereby declare that the contents of this dissertation are original and has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation has been composed solely by myself and it has not been submitted. It contains nothing which is the outcome of work done in collaboration with others, except as specified in the references and Acknowledgements.

Aparna Pradhan
Registration No. - 137327 of 2016-2017
Examination Roll No. - MCA196017

Jadavpur University

Department of Computer Science

Jadavpur, Kolkata – 700032

CERTIFICATE

This is certify that the project entitled **"Identification of Emotion and Age from Audio Signal",** submitted by **Aparna Pradhan** is a record of bona-fide work carried out by her, in the partial fulfillment of the requirement for the award of Degree of **Master of Computer Application** of the Department of Computer Science and Engineering, Jadavpur University.This work is done during the academic year 2018-2019, under my guidance.

_____

(**Prof. Sanjoy Kumar Saha**)

Project Supervisor
Computer Science and Engineering
Jadavpur University, Kolkata - 700032

Countersigned:

_____

(Dr. Mahantapas Kundu)

Head of Department
Computer Science and Engineering
Jadavpur University, Kolkata - 700032

_____

(Prof. Chirnajib Bhattacharjee)

Dean
Faculty Of Engineering and Technlogy
Jadavpur University, Kolkata - 700032

Jadavpur University

Department of Computer Science

Jadavpur, Kolkata – 700032

CERTIFICATE OF APPROVAL

The foregoing project report entitled "**Identification of Emotion and Age from Audio Signal**" is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood by this approval the undersigned do not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn therein but approve the report only for the purpose for which it has been submitted.

_____ _____

Internal Examiner **External Examiner**

Acknowledgements

It is a matter of pleasure for me to be assigned with this project work. I have put my knowledge and effort in the best possible manner.

First and foremost, I want to express my sincere thanks and gratitude to Prof. Sanjoy Kumar Saha for his persistent interest, construction criticism and encouragement throughout the project. And I greatfully acknowledge my deepest gratitude to Rajib Sarkar for his guidance and input which made this project successful. Finally, I want to thank my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you

Abstract

Automatic identification of emotion and age from speech is a prominent research area in recent past. In this project, we have investigated the acoustic properties of speech related to basic human emotions such as happiness, sadness, anger, fear, neutral etc. as well as how voice property changes over the ages of a person delivers the speech. It helps to find out the emotion quotient of the speaker. Automatic identification of emotion and age group has both social and commercial impact like human computer interaction, commercial product promotions, tele-marketing, analysing emotional state or mental health of a person to prevent suicidal tendency or other social violence. In this work speech signal is classified based on fundamental aspects like emotion and age. Low level features are extracted as descriptors. SVM with SMO is used for classification purpose. The classification results for individual as well as the combination of the aspects are satisfactory.

Table of contents

Chapter 1

Introduction

Emotion is the representation of the mental state of the person through their thought, feeling, mood and gesture. Emotion is very complex in nature. This state of feelings influences the behavioural change. Emotion is correlated with mood, temperament, behavioural pattern and personality. The basic emotions are happiness, sadness, anger, neutral and fear. Emotion plays a vital role in understanding one's thinking and behaviour. The three components of emotions are that how we experience the emotions, how our bodies reacts to the emotions and how we response to the emotion. Emotion helps to understand each other, taking wise decision, surviving and avoiding danger and responding to the situation.

Emotions of a person can be analysed or understood by their speaking voice or speech. It mainly focuses on the tone and pitch of the voice and other non-verbal elements used for communication such as gestures used while talking or the distance maintained by the communicators while talking. When a person is happy or angry there is a longer utterance duration while speaking, shorter inter-word silence, with higher pitch and energy level with very high ranges. This shows that both emotions shares the same acoustic properties. Similarly, in the case of boredom and sadness has less power(low RMS energy), low pitch, narrow pitch range and speaking rate helps in determining these emotions.

In theoretical as well as in practical, researchers defines emotion in one or more dimensions. Most of the models incorporates valence and arousal dimen-

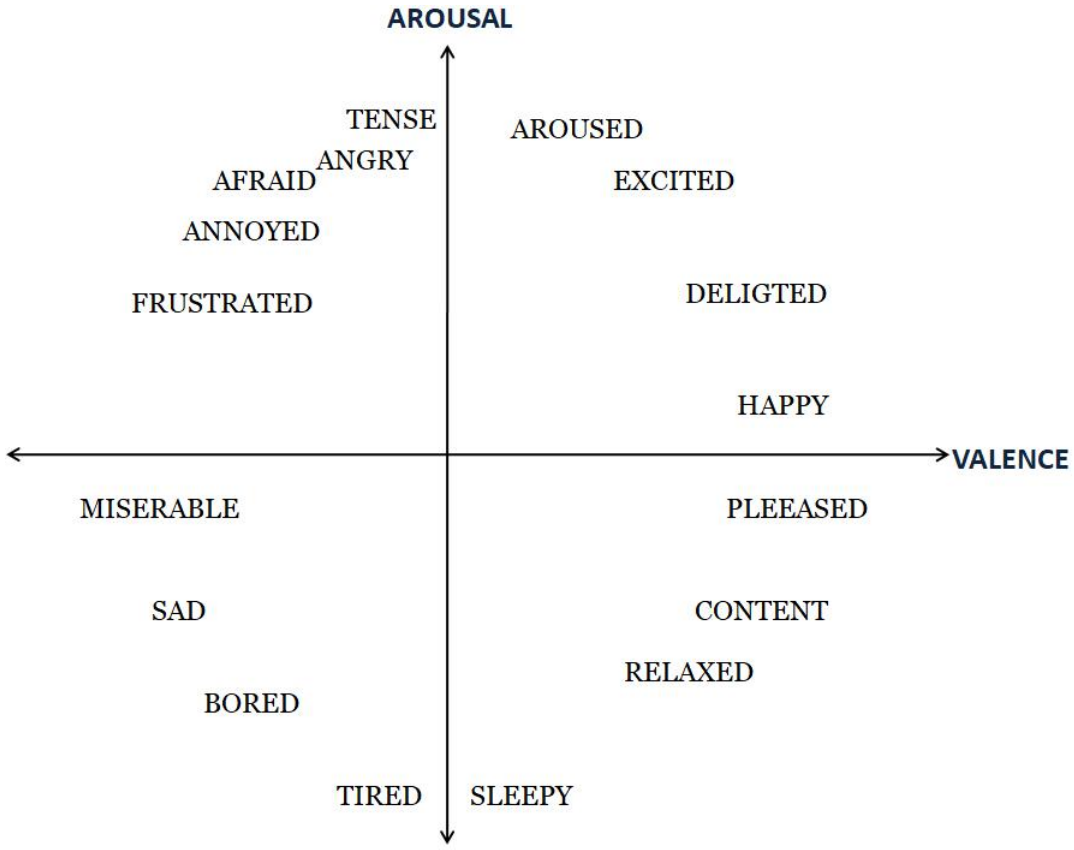

Fig. 1.1 Two Dimensional Emotion Plane: Valence vs. Arousal

sions. These dimensional model helps to find inter-connection and common properties between different emotional states. One of the model is circumplex model of emotion which was developed by James Russell [1] and shown in Figure 1.1. This model defines an emotional space along with unpleasant-pleasant axis (valence axis) and deactivation-activation axis (arousal axis). Russell's model is referred to further researches on different fields such as psychology and computer graphics[2]. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states[3]. Researchers have adopted Russell's model to use regression based model, where valence and arousal are regressed and plotted Arousal-Valence plane to estimate the perceived emotion. Beside Russell's model, researchers have experimented with single emotion tagging method. In this method, each

speech signal is tagged with a emotion category, and a feature based classification method is used to classify into that emotion category.

Speech Emotion Recognition system has derived the new aspect regarding the emotion inherent in the speech which helps in determining the state of mind of the speaker. It has the wide range of applications in the field of Human Computer Interaction(HCI), determining neuro-cognitive disorder etc.

Speaker Age Estimation can be determined using speech signals. Voice aging is caused by normal anatomical and physiological changes associated with this phase of life [4]. In particular, physiological changes occur in the larynx, the vocal tract, and the respiratory system. From a perceptual point of view, the aged voice has been associated with increased hoarseness and breathiness, vocal fatigue, instability, and crackling [5]. Identification of speaker age is mostly used in commercial and technologies during product promotions and investigation of crimes. It is comparatively difficult for the machines to identify from the audio signal as it depends upon the speaker's weight, language, gender, etc.

Automatic identification of emotion and age group has both social and commercial impact like HCI(Human Computer Interaction), commercial product promotions, tele-marketing etc. Analysing emotional state of different age group subjects can help us to find mental health of a person. Identification and proper monitoring of a person with depression emotional state for a long period of time can help us to prevent suicidal tendency or other social violence.

Chapter 2

Past work

Recently acoustic investigation of emotions expressed in speech has attracted increasing attention partly due to the potential value of emotion recognition for spoken dialogue management[6][7][8]. For instance, displeasure or anger due to frequent system errors in understanding user's requests could be dealt with smoothly by transferring the user to a human operator before premature man machine dialogue disruption. However, in order to reach such a level of performance we need to identify a reliable acoustic feature set that is largely immune to inter- and intra-speaker variability in emotion expression[9]. A prerequisite for this is to accumulate knowledge on how acoustic parameters of speech are modulated when emotion changes from normal to a certain emotional state. Such knowledge is also valuable for emotional speech synthesis through speech modification[10]. The speech database introduced and analyzed in this study has been designed and is currently being expanded with such purposes in mind. Some preliminary results of the acoustic analysis of the emotional speech database is presented[9].

In general, emotion has been described in a two dimensional space where arousal (activation or energy) and valence (pleasure) represent each dimension[9, 1]. Commonly analyzed acoustic parameters for such a description of emotion in speech have been pitch, duration at phoneme or syllable level, inter-word silence duration and voiced/unvoiced duration ratio in utterance level, energy related to the waveform envelop, the first three formant frequencies and spectral

moment or balance. These are parameters related to speech prosody, vowel articulation and spectral energy distribution. Detailed reviews can be found in [9].

Specifically, previous studies have shown that anger and happiness/joy are generally characterized by high mean pitch, wider pitch range, high speech rate, increases in high frequency energy, and usually increases in rate of articulation [9]. Sadness is characterized by decrease in mean pitch, slightly narrow pitch range, and slower speaking rate[11]. Kienast et al.[12] analyzed spectral and segmental changes due to emotion in speech. Their study on segmental reduction and vowel formants showed that anger has the highest accuracy of articulation compared to other emotions that they analyzed. They also analyzed the spectral balance of fricative sounds. Their analysis revealed that two different groups can be observed, one containing fear, anger and happiness (increased spectral balance compare to neutral), and the other containing boredom and sadness (decreased spectral balance compare to neutral). Yang and Lugger [13] calculated various spectral gradients as voice quality (VQ) features. Further they have extracted pitch interval features and an auto-correlation of pitch features. Kotti et al. [14] have investigated wide range of acoustic features along with video processing features. They have extracted a total of 2,327 features for speech emotion recognition.

Many researchers have proposed important speech features which contain emotion information, such as the patterns inherent in a speech signal provide the perception of emotion [15]. Energy or the power of a music clip is frequently used [16–20] as it has very correlation with arousal [21]. A audio clip with fast tempo is often correlated with positive valence and slow tempo is correlated with negative valence [21]. Hence, use of tempo is also very common [22, 20, 23, 24]. Timbral features, captured in different forms are also utilized by the researchers. Such features include mel-frequency cepstral coefficients (MFCC) [25, 26], daubechies wavelets coefficient histogram (DWCH) [22, 27, 24]. Zero crossing rate (ZCR) and pitch [16] are also useful. Variants of spectral features [16, 28]

like spectral rolloff, spectral flux as well as tonality [27], fundamental frequency (f_0) [16] are also considered in various works.

For age detection, researchers have investigated the behaviour of anatomical and physiological changes associated human voice. They have used various acoustic features to model that changes. Li et al. [29] uesd MFCC, prosodic and voice quality information, pitch, energy, harmonic structure of spectrogram etc. features. Dobry et al. [30] used MFCC features. They have experimented with principal component analysis (PCA) and weighted-pairwise principal components analysis (WPPCA) based on the nuisance attribute projection (NAP) technique. In their experiment, Dupuis et al. [31] used fundamental frequency (f_0) to differentiate age groups.

In classifier based approach audio clips are first represented by a set of features. Thereafter, feature vector is fed as input to the classifier for emotion or age group recognition. Commonly used classifiers include support vector machine (SVM) [17, 26], artificial neural network (ANN), radial basis function ANN (RBF-ANN) [32], Gaussian mixture model (GMM) [19], random forest [16] etc. Researchers have experimented with different parameter and kernel setups for the classifiers. In some cases, principal component analysis (PCA) [32] and linear discriminant analysis (LDA) have been used for reduction of feature dimension.

It is observed that variety of features and classifiers/regression models have been considered by the researchers. But success of all such systems are quite limited. Hence, emotion and age based categorization still remains an active area of research.

Chapter 3

Methodology

The proposed methodology focused on identification of emotion expressed through speech signal as well as the age group of the speakers. First of all, speech signal is classified using emotion categories and age groups independently. Then, emotion and age aspects are combined together to see the performance of the proposed system. Suitable features has been extracted from the speech excerpts based on the aspects. For classification, Support Vector Machines (SVM) with Sequential Minimal Optimization (SMO) is used. In this work we have experimented with two speech databases that is EmoDB(German Speech Emotion Database) of different age group speakers and TESS(Toronto Emotional Speech Set). TESS contains speech signals of different emotional states of two different age. These are detailed in the following sections.

3.1 Feature Extraction

3.1.1 Emotion based features

Emotion is a mental state associated with the brain. Emotions can be described as a good or bad experience that is associated with certain pattern of physiological activity. The way spoken accents are patterned along with the frequency pattern produced by vocal cord, through time leads listeners to anticipate the emotional essence spread over the speech signal. In this work, feature set is designed considering the relation between emotional response and speech

structure. Spectrogram of a speech signal represents the frequency domain behaviour of the speech signal over the time. We have investigated various pattern and texture based features from spectrogram as emotion is concealed in speech signal. The following MFCC and spectral features are extracted from spectrogram and used for emotion classification.

MFCC based feature

The Mel Frequency Cepstral Coefficients (MFCCs) is considered as listener end feature as it takes the functionality of cochlea in human auditory system into consideration. The Mel scale is related to perceived frequency of a pure tone to its actual measured frequency. Human ear can detect small changes at low frequencies very efficiently. But cannot detect small changes at high frequency. Human cochlea vibrates at different locations depending on the frequencies of the audio signal that ear receives. Accordingly different nerves of the brain are fired to provide the perception of the frequency. In audio signal processing, the frequency perception technique of human ear is performed by Mel filterbank. The shape of the filterbanks is triangular. The initial filters are very narrow as the human ear can sense the small differences. Higher the frequencies, corresponding Mel filters get wider, to become less concerned about small variations. In short, MFCC is a compact description of the shape of the spectral envelope of an audio signal from perceptual perspective. The steps for computing MFCC are elaborated in [33]. First of all the signal is divided into frames. Corresponding to each frame, log of amplitude spectrum is computed. The spectrum is then transformed into Mel scale. Mel frequency $m(f)$ corresponding to the signal frequency f is computed as

$$m(f) = 1125 * log_e(1 + \frac{f}{700})$$

It may be noted that there is a nonlinear relationship between the actual frequency scale and the Mel scale to incorporate the perception model. Finally, discrete cosine transform (DCT) is applied on the Mel spectrum to obtain the

co-efficients. In our work, frame size is taken as 256. First thirteen co-efficients of all the frames are considered. The frame level co-efficients may be concatenated to represent the signal characteristics in detail. But the dimension becomes prohibitive. We have considered mean and standard deviation of the 13 co-efficients over the frames.

Spectral Flux (SF):

Spectral flux indicates the amount of changes or variations reflected in spectral shape. For n-th frame, the spectral flux is computed as:

$$SF(n) = \frac{\sqrt{\sum_{i=0}^{K-1}(|\mathscr{S}(i,n)| - |\mathscr{S}(i,n-1)|)^2}}{K}$$

It captures changes in the power of spectral components over the successive frames.

Spectral Rolloff (SR):

It is defined as the q^{th} percentile of the power spectral distribution [34]. SR is identified as the frequency bin for which the overall power spectrum of $\mathscr{S}(i,n)$ covers q percent of the total power spectrum. In our case q is taken as 85.

Spectral Centroid (SC):

The Spectral Centroid of an audio signal represents the center of gravity of the spectral power. SC is a commonly accepted as a measure for brightness of the music signal. It is the ratio of the frequency weighted magnitude spectrum with unweighted magnitude spectrum.

$$SC(n) = \frac{\sum_{i=0}^{K-1} K \times |\mathscr{S}(i,n)|^2}{\sum_{i=0}^{K-1} |\mathscr{S}(i,n)|^2}$$

Spectral Spread (SSP):

Spectral spread also known as instantaneous bandwidth. It measures the centralism of the spectral power about the spectral centroid (SC). It is calculated as

$$SSP(n) = \sqrt{\frac{\sum_{i=0}^{K-1}(i - SC(n))^2 \times |\mathscr{S}(i,n)|^2}{\sum_{i=0}^{K-1}|\mathscr{S}(i,n)|^2}}$$

Spectral Slope (SSL):

It is the measurement of slope of a spectral shape. SSL is measured by taking linear approximation of magnitude spectrum. It is calculated as

$$SSL(n) = \frac{\sum_{i=0}^{K-1}(i - \mu_i)(|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})}{\sum_{i=0}^{K-1}(i - \mu_i)^2}$$

where, $\mu_{\mathscr{S}}$ is the overall mean of spectral magnitude of the spectrogram and μ_i is the spectral component.

Once the frame level spectral features are computed, those are summarized to obtain the clip level descriptors. For each feature, its mean and standard deviation over the frames are considered.

Spectral Flatness Measure (SFM):

It is the proportion of geometric mean and arithmetic mean of a magnitude spectrum [35, 36], as shown below,

$$SFM(n) = \frac{K \times \sqrt[K]{\prod_{i=0}^{K-1}\mathscr{S}(i,n)}}{\sum_{i=0}^{K-1}\mathscr{S}(i,n)}$$

For uniform (flat) distribution of power spectral component it provides higher value.

Spectral Crest Factor (SCF):

It is the measurement of the quality of a acoustic signal [36]. It is computed as the proportion of highest of the power spectrum with total power spectrum.

$$SCF(n) = \frac{\max\limits_{0 \leq i \leq K-1} |\mathscr{S}(i,n)|}{\sum_{i=0}^{K-1} |\mathscr{S}(i,n)|}$$

Spectral Kurtosis (SK):

The spectral kurtosis summarizes the existence of series of momentary variation in frequency and their locations in a spectrogram. The spectral kurtosis is the normalized fourth-order moment of the spectrogram. SK indicates how Gaussian the magnitude spectrum distribution looks like. It is calculated as

$$SK(n) = \frac{\sum_{i=0}^{K-1} (|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})^4}{K \times \sigma_{\mathscr{S}}^4}$$

where, $\mu_{\mathscr{S}}$ is the mean of spectral magnitude and $\sigma_{\mathscr{S}}$ is the standard deviation of the spectrogram.

Spectral Skewness:

It is the ratio of third central moment of the spectral components and the cube of its standard deviation. It is calculated as

$$SSK(n) = \frac{\sum_{i=0}^{K-1} (|\mathscr{S}(i,n)| - \mu_{\mathscr{S}})^3}{K \times \sigma_{\mathscr{S}}^3}$$

Here also, mean and standard deviation of individual frame level features are considered at the clip level.

3.1.2 Age based features

The relationship between vocal characteristics and perceived age of the speaker is of interest in various contexts, as is the possibility to affect age perception

through vocal manipulation. The human voice changes from childhood and throughout an individual's lifespan because of biochemical and physiological changes affecting the speech mechanism. Regularities in this variation allow listeners to make fairly accurate assessments of the speaker's age from his or her voice and may also be used by speakers to give the impression of being younger or older than s/he actually is. In this work, we have tried to capture the age based properties from the bark scale bands of a spectrogram. The following Bark Scale based feature is discussed in following section.

Bark Scale based features

Bark scale is a psychoacoustical scale which is defined as the measurement of loudness where each width is of one bark. There are in total 24 critical bands of hearing which ranges from 1 to 24 Barks. The Bark band edges and band centres are [0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500] Hertz. These center-frequencies and band-widths are to be interpreted as samplings of a continuous variation in the frequency response of the ear to a sinusoidal or narrow-band noise process [37].

We have divided the spectrogram considering 19 bark scale bands. For each band, frame wise log magnitude spectrum is captured. Finally, for all 19 bark scale bands, mean and standard deviation over the frames are calculated and used as feature vector.

3.2 Classifier

3.2.1 Support Vector Machines (SVM):

We have used Support Vector Machines for classification. SVMs [38] are widely used in classification, regression or novelty detection problem. SVMs are large-margin classifiers. Given training data containing f features with only two possible output labels, SVM finds a hyperplane in f-dimensional hyperplane

that maximizes margin between two classes. This margin is calculated as the difference between the decision border line and the nearest input vectors. This binary classification strategy can be extended to solve multi-class classification problem with one against one approach. Sequential Minimal Optimization (SMO) [39] is used to train the SVM. The SVM training algorithm requires to solve a very large quadratic programming (QP) optimization problem. SMO divides this QP problem into a series of small sub-problems, which are then solved analytically. In this way SMO avoids the expensive QP optimization during the training of SVM.

We have used the implementation of the classifiers from the WEKA [40] framework.

Chapter 4

Results

4.1 Dataset

In order to carry out the experiment, we have used two benchmark datasets *EmoDB* and *TESS*. The dataset details are discussed below.

4.1.1 EmoDB(German Speech Emotion Database)

EmoDB (German Speech Emotion Database) [41] database of emotional speech contains 535 audio clips of different emotional utterance spoken by actors in a happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version it contains the utterance from 10 different actors of various age group (5 males and 5 females). It contains high quality audio clips with minimizing background noise. To ensure its correctness perception test are also done.

4.1.2 TESS(Toronto Emotional Speech Set)

TESS (Toronto Emotional Speech Set) [42] contains a set of 200 target words where each sentences starts with *"Say the word* ———— *"* by two professional actresses (aged 26 and 64 years) and the audios consists of seven different emotional categories (anger, disgust, fear, happiness, sadness and neutral). There are in total 1400 audio files of each old and young database with above mentioned seven different emotions. These all audio clips are in English

language. Audiometric test has been conducted for measuring the intensity, tone of the sound and other balance issues.

Both *EmoDB* and *TESS* dataset consists of speeches of speakers of different ages. We have combined both of the dataset to prepare our *Age Dataset.* The speech excerpts are categorized into four age groups and is shown in Table 4.1.

Table 4.1 Table for showing *Age Groups* on Age Dataset.

SL No.	Age Range (in Years)	Termed as
1	21 to 25	21-25
2	26 to 30	26-30
3	31 to 35	31-35
4	61 to 65	61-65

4.2 Result and Discussion

The above experiment is done using SVM classifier with SMO. All the results reported are based on five fold cross validation testing i.e. 60% data used for training, 20% data used for validation and 20% data used for testing. The results are represented below using a confusion matrix, which gives a visualization of the performance of our proposed system. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class.

4.2.1 Age group classification

For age group classification, out of 2933 excerpts, 2681 excerpts were correctly identified. The classification accuracy is 91.4081%. The confusion matrix for Age Group based classification is shown in Table 4.2.

Table 4.2 Confusion matrix for Age based classification.

Age group	21-25	26-30	31-35	61-65
21-25	0.80	0.03	0.17	0
26-30	0	0.92	0.04	0.04
31-35	0	0.10	0.90	0
61-65	0	0.01	0	0.99

4.2.2 Emotion classification

EmoDB(German Speech Emotion Database)

For this database, 454 excerpts out of 383 were correctly identified, which results an accuracy of 84.36%. The result is as shown in Table 4.3.

Table 4.3 Confusion matrix for classification of Emotion on EmoDB dataset.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	0.96	0	0.04	0	0	0
Disgust	0	0.81	0.15	0	0	0.04
Fear	0.14	0	0.82	0	0	0.04
Happiness	0.17	0.03	0	0.73	0.07	0
Neutral	0.01	0.05	0.10	0.08	0.76	0
Sadness	0	0	0	0	0.10	0.90

TESS (Toronto Emotional Speech Set)

A. TESS for older subjects:

For this database, 1200 excerpts out of 1150 were correctly identified, which results an accuracy of 95.8333%. The result is as shown in Table 4.4.

Table 4.4 Confusion matrix for classification of Emotion for the database TESS for old person.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	0.94	0	0.06	0	0	0
Disgust	0	0.95	0	0.03	0	0.02
Fear	0.03	0.01	0.96	0	0	0
Happiness	0	0.03	0	0.97	0	0
Neutral	0	0	0	0	0.98	0.02
Sadness	0	0.01	0	0	0.04	0.90

B. TESS for younger subjects:
For this database, 1200 excerpts out of 1188 were correctly identified, which results an accuracy of 99.00%. The result is as shown in Table 4.5.

Table 4.5 Confusion matrix for classification of Emotion for the database TESS for young person.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	0.98	0	0	0.02	0	0
Disgust	0	1.00	0	0	0	0
Fear	0.01	0	0.98	0.01	0	0
Happiness	0	0	0.01	0.99	0	0
Neutral	0	0.01	0	0	0.99	0
Sadness	0	0	0	0	0	1.00

4.2.3 Classification of Emotion and Age together

In order to find out speech excerpts which correctly identified with both emotion and age, we have build two models. First model identifies age groups and second model identifies emotions. Only correctly identified age group excerpts are tested with second model for emotion identification. The first model for age group classification, out of 2933 excerpts, 2681 excerpts were correctly identified. The second model for emotion classification, correctly identified 2371 excerpts out of 2681 excerpts. Thus total 2371 excerpts out of 2933 are correctly identified as both age and emotion category, resulting an accuracy of 80.84%.

4.2.4 Performance Comparison

We have compared our method to other reported results on both datasets. A comparison of performance on EmoDB dataset is summarised in Table 4.6. It is clear that performance of our method is better compared to other established methods. It is also noted that Kotti et al.[14] have used feature set of dimension 2317, which is more than four times of total number of speech excerpts present

in the dsataset. In our method, feature dimension for emotion identification is 44.

Table 4.6 Comparison with other standard methods on EmoDB dataset.

Dataset	Method	Feature Set	Classifier	Accuracy (in %)
EmoDB	Ours	Spectral features and MFCC	SVM	84.36
	Semwal et al.[26]	ZCR, MFCC Spectral, Chroma harmonic features	SVM	80.00
	Yang and Lugger [13]	spectral gradients, pitch features	GMM	73.62
	Kotti et al.[14]	a total of 2317 dimensional acoustic and video features	SVM	83.30

We did not find any reported result on automatic classification of emotion on TESS dataset, best of our knowledge. But, Dupuis and Pichora-Fuller [31] have experimented on how efficiently human being can identify basic emotions. They have experimented on TESS dataset. We have compared our automatic emotion identification method to Dupuis and Pichora-Fuller's [31] human (or manual) emotion identification system and is shown in Table 4.7.

Table 4.7 Comparison with other standard methods on TESS dataset.

Dataset	Method	Target Age Group	Accuracy (in %)
TESS	Ours	Younger	99.00
		Older	95.83
	Dupuis and Pichora-Fuller [31]	Younger	82.10
		Older	65.80

Chapter 5

Conclusion

Despite the fuzzy nature of emotion boundaries and age group, classification can be performed automatically with results significantly better than chance, and performance comparable to human classification. Using the proposed features set 91.41% classification accuracy achieved based on age groups. For emotion classification, we have achieved an accuracy of 84.36% German data (EmoDB). For English data (TESS), we have achieved an accuracy of 95.83% and 99% on older and younger subjects respectively. Automatic identification of emotion and age group has both social and commercial impact like promotion of commercial products, HCI, monitoring mental health of a person to perpetuate social health. In this work speech signal is classified based on two fundamental aspects emotion and age. Spectral features are extracted as descriptors. The classification results and performance comparison with other existing systems validates proposed method.

References

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

[2] Junghyun Ahn, Stephane Gobron, Quentin Silvestre, and Daniel Thalmann. Asymmetrical facial expressions based on an advanced interpretation of two-dimensional russell's emotional model. *Proceedings of ENGAGE*, 2010.

[3] Peter Warr. How to think about and measure psychological well-being. In *Research methods in occupational health psychology*, pages 100–114. Routledge, 2012.

[4] Leonardo A Forero Mendoza, Edson Cataldo, Marley MBR Vellasco, Marco A Silva, and Jose A Apolinario Jr. Classification of vocal aging using parameters extracted from the glottal signal. *Journal of Voice*, 28(5):532–537, 2014.

[5] Naomi D Gregory, Swapna Chandran, Deborah Lurie, and Robert T Sataloff. Voice disorders in the elderly. *Journal of Voice*, 26(2):254–258, 2012.

[6] Chul Min Lee, Shrikanth S Narayanan, et al. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

[7] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[8] Diane Litman and Kate Forbes. Recognizing emotions from student speech in tutoring dialogues. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 25–30. IEEE, 2003.

[9] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*, 2004.

[10] Janet Elizabeth Cahn. *Generating expression in synthesized speech*. PhD thesis, Massachusetts Institute of Technology, 1989.

[11] Iain R Murray and John L Arnott. 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.

[12] Miriam Kienast and Walter F Sendlmeier. Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[13] B. Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90(5):1415 – 1423, 2010. Special Section on Statistical Signal & Array Processing.

[14] Margarita Kotti and Fabio Paternò. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International Journal of Speech Technology*, 15(2):131–150, Jun 2012.

[15] Carol L Krumhansl. Music: A link between cognition and emotion. *Current directions in psychological science*, 11(2):45–50, 2002.

[16] Fan Zhang, Hongying Meng, and Maozhen Li. Emotion extraction and recognition from music. In *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1728–1733, 2016.

[17] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.

[18] Ali Hassan, Robert Damper, and Mahesan Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.

[19] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2006.

[20] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *IEEE Trans. Audio, Speech & Language Processing*, 19(6):1802–1812, 2011.

[21] Alf Gabrielsson and Erik Lindström. *The influence of musical structure on emotional expression.* Oxford University Press, 2001.

[22] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Paiva. Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset. In *Proceedings of the International Workshop on Music and Machine Learning*, 2016.

[23] Byeong Jun Han, Seungmin Rho, Roger B Dannenberg, and Eenjun Hwang. SMERS: Music Emotion Recognition Using Support Vector Regression. In *Proceedings of the International Society for Music Information Retrieval*, pages 651–656, 2009.

[24] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. Music emotion classification: A regression approach. In *Proceedings of the International Conference on Multimedia and Expo*, pages 208–211, 2007.

[25] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools and Applications*, 76(5):6785–6799, 2017.

[26] N. Semwal, A. Kumar, and S. Narayanan. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–6, Feb 2017.

[27] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457, 2008.

[28] Qi Lu, Xiaoou Chen, Deshun Yang, and Jun Wang. Boosting For Multi-Modal Music Emotion. In *Proceedings of the International Society for Music Information and Retrieval Conference*, page 105, 2010.

[29] Ming Li, Kyu J. Han, and Shrikanth Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151 – 167, 2013. Special issue on Paralinguistics in Naturalistic Speech and Language.

[30] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1975–1985, Sep. 2011.

[31] Kate Dupuis and M Kathleen Pichora-Fuller. Aging affects identification of vocal emotions in semantically neutral sentences. *Journal of Speech, Language, and Hearing Research*, 58(3):1061–1076, 2015.

[32] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014.

[33] Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR*, 2000.

[34] Tong Zhang and C-C Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on speech and audio processing*, 9(4):441–457, 2001.

[35] A Gray and J Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3):207–217, 1974.

[36] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 1st edition, 2012.

[37] Julius O Smith and Jonathan S Abel. Bark and erb bilinear transforms. *IEEE Transactions on speech and Audio Processing*, 7(6):697–708, 1999.

[38] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[39] John Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical report, apr 1998.

[40] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, nov 2009.

[41] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[42] Kate Dupuis and M Kathleen Pichora-Fuller. *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department, 2010.