A Survey on Gender and Emotion Recognition Using Voice

Poonam Rani[1], Mr.Bhupender Yadav[2]

[1]Gurgaon institute of Technology and Management, MDU University, Rohtak Haryana India
[2]Head of Department (CSE), Gurgaon institute of Technology and Management, MDU University, Rohtak Haryana India

Abstract: **The voice of a person plays an important role in analyzing people. From voice of the person we not only recognize the gender of the person but also detect the emotion of the person. Gender recognition and emotion detection both has its importance in forensics, games, in security purposes and of course in our day to day life. This paper proposes a system that allows recognizing a person's emotional state starting from audio signal registrations. Identifying the gender and emotion of a speaker from speech has a variety of applications ranging from speech analytics to personalizing human-machine interactions. While gender identification in previous works has explored the use of the statistical properties of the speaker's pitch features, in this project, we explore the impact of using acoustic features on identifying gender. In addition to gender we will also predict the emotion of speaker using the same acoustic values. We present a novel approach that models acoustic properties in the interest of identifying the speaker's gender and emotion with as little speech as possible. In this project we will investigate two datasets containing voice samples of over 3000 people for gender and over 1000 voice samples for emotions**.

Keywords: **Gender and Emotion Recognition, Voice**

I. INTRODUCTION

Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future peoples life. A peculiar and very important developing area concerns the remote monitoring of elderly or ill people. Indeed, due to the increasing aged population, HCII systems able to help live independently are regarded as useful tools. Despite the significant advances aimed at supporting elderly citizens, many issues have to be addressed in order to help aged ill people to live independently. In this context recognizing people emotional state and giving a suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot research area in both industry and academic field.

There is much research in this area and there have been some successful products [1].Usually, emotion recognition systems are based on facial or voice features. This paper proposes a solution, designed to be employed in a Smart Environment, able to capture the emotional state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smartphones. Main problems to be faced concern: the concept of emotion, which is not precisely defined for the context of this paper; the lack of a widely accepted taxonomy of emotions and emotional states; the strong emotion manifestation dependency of the speaker. Emotion recognition is an extremely difficult task. This paper presents the implementation of a voice-based emotion detection system suitable to be used over smartphone platforms and able to recognize six emotions (anger, boredom, disgust, fear, happiness, sadness) and the neutral state, as widely used for emotion recognition. Particular attention is also reserved to the evaluation of the system capability to recognize a single emotion versus all the others. For these purposes, a deep analysis of the literature is provided and state-of-the-art approaches and emotion related features are evaluated. In more detail, to capture emotion information, 182 different features related to speech signals' prosody and spectrum shape is used; the classification task is performed by adopting the Support Vector Machine (SVM) approach.

The survey of this study elaborates analysis of different techniques used for detection of gender and emotion of human. Main contribution to this paper concerns different possible solutions for gender and emotion detection. This survey paper is composed of i) Feature extraction ii) Gender detection iii) Emotion detection. It briefly explains various techniques available for gender and emotion detection through speech signal such as MFCC, autocorrelation method and different classification methods. It also includes figure analysis and summarizing sections.

II. EXISTING SYSTEMS (Problems)

People can identify gender and emotions of other people easily just by listening to their voice but training a computer program to this is a difficult task. Building a computer program to identify gender and emotion can be used in various technologies for making great user experiences. Voice recognition can be used in artificial intelligent systems. In general identification of a speaker gender is important for increasingly natural and personalised dialogue systems.

III. PROPOSED SYSTEM

The proposed system provides the facility to determine a person's gender and emotion from his/ her speech. The system is provided with 3000+ voice samples for gender recognition and 800+ voice samples for emotion recognition. The system extracts the features from the voice samples and stores them in 'CSV' files. These files are used to build models for prediction of gender and emotion. The gender and emotion of new voice sample received is predicted using these models.

The main contributions of this paper concern: i) a system able to recognize people emotions composed of two subsystems, Gender Recognition (GR) and Emotion Recognition (ER); ii) a gender recognition algorithm, based on pitch extraction, and aimed at providing a priori information about the gender of the speaker; iii) a SVM-based emotion classifier, which employs the gender information as input. Reduced feature sets, obtained by feature selection, performed through Principal Component Analysis (PCA), have been investigated and applied. In order to train and test the mentioned SVM based emotion classifier, a widely used emotional database (called Berlin Emotional Speech Database, BESD) has been employed. Experimental results show that the proposed system is able to recognize the emotional state of a speaker with an accuracy level often higher than the evaluated methods taken from the literature, without applying any pre-processing on the analysed speech signals. The obtained results show also that adopting a feature selection algorithm assures good recognition rate levels also when a consistent reduction of the used features is applied. This allows a strong limitation of the number of operations required to identify the emotional content of a particular audio signal. These peculiarities make the proposed solution suitable to operate on mobile platforms such as smartphones and tablets, in which the availability of computational resources and the energy consumption constitute issues of primary relevance. The obtained results also show a strong dependency of the overall system reliability on the database adopted for training and testing phases: the use of a simulated database (i.e., a collection of emotion vocal expressions played by actors) allows obtaining a higher level of correctly identified emotions. In addition, the performed tests show that the SVM based emotion classifier can be reliably used in applications where the identification of a single emotion (or emotion category) versus all the other possible.

Existing systems for feature extraction:-

Paper titled "Feature Extraction from Speech Data for Emotion Recognition" authored by S. Demircan and H. Kahramanlı published in 2014 describes which features are best suited for gender and emotion recognition. It implements extraction of Mel Frequency Cepstral Coefficients (MFCC) from the signals and classification with kNN algorithm to recognize emotion. Statistics of Mel-Frequency Cepstral Coefficients which are computed over three phoneme type classes of interest are stressed vowels, unstressed vowels and consonants in the speech. The paper clearly indicates that indeed both the richer set of spectral features and the differentiation between phoneme type classes are helpful for the task. The method uses short time log frequency power coefficients (LFPC) to represent a speech signals and a discrete hidden Markov model (HMM) as the classifier. Performance of the LFPC feature parameters is compared with performance of linear prediction Cepstral coefficients (LPCC) and melfrequency Cepstral coefficients (MFCC) feature used in speech recognition systems. This paper suggests that LFPC is a better choice as feature parameters for gender and emotion classification.

Paper titled " Improved MFCC-Based Feature for Robust Speaker Identification" ,authored by WU Zunjing , CAO Zhigang, published in 2005 describes, the Mel-frequency cepstral coefficient (MFCC) is the most widely used feature in speaker recognition. It presents a method which combines robust representations in the feature space sample and speech enhancement technique in signal space where the important aim for feature space is to extract the acoustic features of the input speech. The paper also shows that standard MFCC feature analysis is very successful except when the noise is present. The log function in MFCC is sensitive to noise, so it replaced the logarithmic transformation in the MFCC analysis by a combined function to improve the noise.

1) GENDER RECOGNITION FEATURES

Together with the Mel Frequency Cepstral Coefficients (MFCC) [7] pitch is the most frequently used feature since it is a physiologically distinctive trait of a speaker's gender. Other employed features are formant frequencies and bandwidths, open quotient and source spectral tilt correlates energy between adjacent formants, fractal dimension and fractal dimension complexity jitter and shimmer (pitch and amplitude micro-variations, respectively) harmonics-to-noise-ratio, distance between signal spectrum and formants.

The paper titled "Gender classification by pitch analysis" ,authored by BhagyaLaxmi Jena & Beda Prakash Panigrahi studies about developing a gender classifier using speech signal. It mainly concentrates on pitch analysis of speech signal for gender classification. It consist analysis of pitch values of male and female voice samples. It implements pitch determination through autocorrelation method. The paper states that there is sufficient difference between the average pitch value of male and female voice sample value obtained by auto-correlation method. The table1 shows the experimental average pitch values of male and female voice samples. It uses difference in pitch value to develop gender classifier by setting threshold pitch value for male and female voice samples.

TABLE 1: Pitch comparison between male and female voice Sample

Female Voice	Average pitch(in Hz)	Male Voice	Average pitch(in Hz)
Female 1	236.89	Male 1	154.68
Female 2	273.7	Male 2	184.22
Female 3	310.01	Male 3	192.92
Female 4	258.98	Male 4	190.71
Female 5	178.40	Male 5	137.35

.

As reported in papers such as [3], [5], audio-based Gender Recognition (GR) has many applications. For example: gender-dependent model selection for the improvement of automatic speech recognition and speaker identification, content-based multimedia indexing systems, interactive voice response systems, voice synthesis and smart human-computer interaction. In this paper, the recognition of the gender is used as input for the emotion recognition block. As shown in the numerical result section, this pre-filtering operation improves the accuracy of the emotion recognition process. Different kinds of classifiers are used to identify the speaker gender starting from features: e.g., Continuous Density Hidden Markov Models [7], Gaussian Mixture Model (GMM) [4], Neural Networks [1], Support Vector Machines [2]. The percentages of correct recognition of the speaker gender are reported in Table 2 for most classifiers referenced above.

TABLE 2. Classification accuracy (percentage) obtained by the evaluated GR methods.

Refrence	Accuracy
[10]	100
[33]	100
[14]	98.35
[12]	95
[32]	91.7
[13]	90.5
[34]	90

The paper titled, "Gender classification using pitch and formants" is presents information about gender detection using speech signals. It describes methods which are based on pitch, formants and combination of both. It briefly explains Pitch Detection Algorithms based on the autocorrelation function, the average magnitude difference function, cepstral analysis and formants extraction. The paper concludes that autocorrelation method shows better results for pitch estimation as compared to other techniques. On the basis of above survey of gender survey of gender detection techniques, it can be concluded that pitch calculation is very important factor in finding gender through speech signal. For more accuracy we can combine it with formant, MFCC.

2) EMOTION RECOGNITION FEATURES

This feature will take input a voice sample from user and analyse it to predict the emotion of the user. This feature will train from a dataset of over 1000 voice samples which is made by extracting acoustic properties of sample through seewave package in R and storing in a CSV file. Following are the various steps of implementation of this functionality.

Recording Voice Samples-

In this phase, over 1000 voice samples are collected from users either by recording them or by downloading them from internet and storing them separately for emotions -

1 Neutral 2 engry 3 Sad 4 fear.

Extracting Acoustic Properties such as duration: length of signal meanfreq: mean frequency (in kHz) sd: standard deviation of frequency median: median frequency (in kHz) Q25: first quantile (in kHz). Coherently with the wide literature in the field, in this paper a set of 182 features has been analysed for each the recorded speech signal, including: Mean, variance, median, minimum, maximum and range of the amplitude of the speech Mean, variance, minimum, maximum and range of the formants; Energy of the Bark sub-bands Mean, variance, minimum, maximum and range of the Mel-Frequency Cepstrum Coefficients Spectrum shape features Centre of Gravity, Standard Deviation, Skewness and Kurtosis; Mean and standard deviation of the glottal pulse period, jitter local absolute, relative average perturbation, difference of difference period and (–ve) point period perturbation quotient.

Pitch detection:

Pitch is one of the essential components of emotion recognition from audio signal. It defines rate of vibration of speaker's vocal cord. Although different sub features like fundamental frequency, pitch, harmony etc. are used. In this work the features selected are: cepstral fundamental frequency, harmony and pitch contour.

Formant frequency:

The formant feature specifies phonetic content of speech signals. As we know, that Hindi is more phonetic compared to English. So considering this point formant frequency is selected.

Feature combination:

The features obtained from the audio signal are of different sizes. In order to make them uniform the feature vector is resized. The duration and ZCR is resized into one pixel, whereas energy, fundamental frequency, pitch, and formant frequency are resized to 20 pixels. In this way total feature length of 102 pixels is generated for each audio signal. These features are further classified by using statistical (KNN) and Neural Network based classifiers.

Paper titled "Recognizing emotion in speech" authored by Frank Dellaert, Thomas Polzin and Alex Waibel describes statistical pattern recognition techniques to classify speech signal according to their emotional content. They have conducted a small and informal experiment in order to assess how well a human does in classifying emotions. In this paper for classification they have used only pitch information extracted from speech. The paper compares classification methods such as Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR) and K-nearest neighbors

(KNN) for emotion recognition. It studies feature selection methods like population hillclimbing and forward selection which gives better result with above mentioned classification methods. It interprets the features selected by these methods as the likely correlates of emotion.

Paper titled, "Emotion and Gender Recognition of Speech Signals Using SVM", presents implementation of emotion detection through speech signal system. It describes two support vector machines (SVM'S) which are distinctly used for the male and female speaker's emotion recognition such as anger, fear, happiness, sad, neutral. It highlight that the knowledge of the speaker's gender allows a performance increase. The paper focuses on the evaluation of the system capability to recognize the single emotion with and without Gender recognition (GR). It also shows that the features selection technique gives a satisfying recognition rate and also allows reduction in employed features.

From above survey, it is clear that for emotion detection there are various classification technologies available like Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR) and K-nearest neighbors (KNN), Support Vector Machine (SVM). But SVM is more efficient among all as it has high accuracy. As per the literature survey, it was found that females usually have shorter and thinner vocal cords than males. And the fundamental frequency (F0) of female voices is typically higher than fundamental frequency (F0) of male voices. This makes fundamental frequency (F0) a prospective choice for gender recognition. And hence we choose the fundamental frequency as the feature to extract. And for emotion detection pitch, energy and speaking rate are the main features to extract from the audio signal. As by getting the pitch, speaking rate and energy of the voice of the person one can detect the emotion of the person easily For example, speech produced in a state of fear, anger or joy becomes faster, louder, precisely enunciated with a higher and wider pitch range.

IV. CONCLUSION

The proposed system, able to recognize the emotional state of a person starting from audio signals registrations, is composed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former has been implemented by a Pitch Frequency Estimation method, the latter by two Support Vector Machine (SVM) classifiers (fed by properly selected audio features), which exploit the GR subsystem output.

The performance analysis shows the accuracy obtained with the adopted emotion recognition system in terms of recognition rate and the percentage of correctly recognized emotional contents. The experimental results highlight that the Gender Recognition (GR) subsystem allows increasing the overall emotion recognition accuracy from 77.4% to 81.5% due to the a priori knowledge of the speaker gender.

V. FUTURE WORK

This work can be expended further in future by increasing the dataset as we have taken the voice samples of 800 people, one can take the more number of voice samples. Further samples can be taken in the vacuum, as the noise affects the accuracy of the gender recognition and emotion detection, it may give the better results. Moreover different feature set can be included as we have taken pitch, energy and speaking rate for emotion detection and fundamental frequency (Fo) for gender recognition, in future more features can be added like MFCC coefficients, LPC coefficients etc. This work can be further expended by using the other emotions like anger, boredom etc as we have worked on only three emotions, happy, normal and sad.

REFRENCES

1. Mehmet Cenk Sezgin, Bilge Gunsel* and Gunes Karabulut Kurt, "Perceptual audio features for emotion detection", EURASIP Journal on Audio, Speech, and Music Processing, 2012.
2. carlos busso,sungbok lee shrikanth narayanan,"analysis of emotionally sailent aspects of fundamental frequency for emotion detection", IEEE transansactions on audio, speech, and language processiong, vol. 17, no.4, may 2009.
3. David Philippou-H¨ubner, Bogdan Vlasenko, Ronald B¨ock, Andreas Wendemuth," The Performance of the Speaking Rate Parameter in Emotion Recognition from Speech", IEEE International Conference on Multimedia and Expo Workshops, pp. 296-301, 2012.
4. WU Zunjing , CAO Zhigang, "Improved MFCC-Based Feature for Robust Speaker Identification", 2005
5. BhagyaLaxmi Jena & Beda Prakash Panigrahi, "Gender classification by pitch analysis", 2012
6. Pawan Kumar, Nitika Jakhanwal, Anirban Bhowmick, and Mahesh Chandra, "Gender classification using pitch and formants", 2011
7. Aastha Joshi, Rajneet Kaur, "Study of speech emotion recognition method", 2013 6. Frank Dellaert, Thomas Polzin and Alex Waibel, "Recognizing emotion in speech
8. Y.-L. Shue and M. Iseli, ''The role of voice source measures on auto- matic gender classification,'' in Proc. IEEE ICASSP, Mar./Apr. 2008, pp. 4493–4496.