International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018 www.csjournals.com

Gender and Emotion Recognition Using Voice

[1]Poonam Rani, [2]Ms.Geeta
[1]Student, [2]Assistant professor
Gurgaon institute of Technology and Management, MDU University Rohtak Haryana India
surliyapoonam@gmail.com, geetgitm@gmail.com

Abstract: 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 paper, 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. Finally, we present various models for gender and emotion detection using the programming language R.

Keywords: Human-computer intelligent interaction, gender recognition, emotion recognition, acoustic properties, support vector machine.

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 people 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]. Determining a person's gender as male or female, based upon a sample of their voice seems to initially be an easy task. Often, the human ear can easily detect the difference between a male or female voice within the first few spoken words. However, designing a computer program to do this turns out to be a bit trickier. This paper describes the design of a computer program to model acoustic analysis of voices and speech for determining gender and emotion. The model is constructed using more than 3,000 recorded samples of male and female voices, speech, and utterances plus over 1000 recorded samples for different emotions. The samples are processed using acoustic analysis and then applied to an artificial intelligence/machine learning algorithm to learn gender-specific traits. The resulting program achieves 89% accuracy on the test set.

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 personalized dialogue systems.

RELATED WORK

Voice

Voice (or vocalisation) is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. Voice is not always produced as speech, however. Your voice is as unique as your fingerprint. It helps define your personality, mood, and health.

R Programming Language [3]

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS

R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

Shiny

Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

CSV (Comma Separated Value)

In computing, comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

TuneR

Analyze music and speech, extract features like MFCCs, handle wave files and their representation in various ways, read mp3, read midi, perform steps of a transcription, ... Also contains functions ported from the 'rastamat' 'Matlab' package.

Seewave

Functions for analyzing, manipulating, displaying, editing and synthesizing time waves (particularly sound). This package processes time analysis (oscillograms and envelopes), spectral content, resonance quality factor, entropy, cross correlation and autocorrelation, zerocrossing, dominant frequency, analytic signal, frequency coherence, 2D and 3D spectrograms and many other analyses.

R Random Forest Algorithm[6]In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model.

Random Forest pseu docode:

1. Randomly select **"k"** features from total **"m"** features.
1. Where **k << m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

The beginning of random forest algorithm starts with randomly selecting **"k"** features out of total **"m"** features. In this, you can observe that we are randomly taking features and observations.

In the next stage, we are using the randomly selected **"k"** features to find the root node by using the best split approach. The next stage, we will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node.

Finally, we repeat 1 to 4 stages to create **"n"** randomly created trees. This randomly created trees forms the **random forest.** The R package "random Forest" is used to create random forests.

CART Model

When utilizing an algorithm such as logistic regression, it can be difficult to determine which exact properties indicate a target gender of male or female. We could guess that it likely one of the statistically significant features, but ultimately this decision breakdown is masked within the model. To gain an understanding of a trained model, we can apply a classification and regression tree model (CART) to our dataset to determine how these properties might correspond to a gender classification of male or female.

Building a CART Model of Voice Acoustics

When utilizing an algorithm such as logistic regression, it can be difficult to determine which exact properties indicate a target gender of male or female. We could guess that it likely one of the statistically significant features, but ultimately this decision breakdown is masked within the model. To gain an understanding of a trained model, we can apply a classification and regression tree model (CART) to our dataset to determine how these properties might correspond to a gender classification of male or female.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018 www.csjournals.com

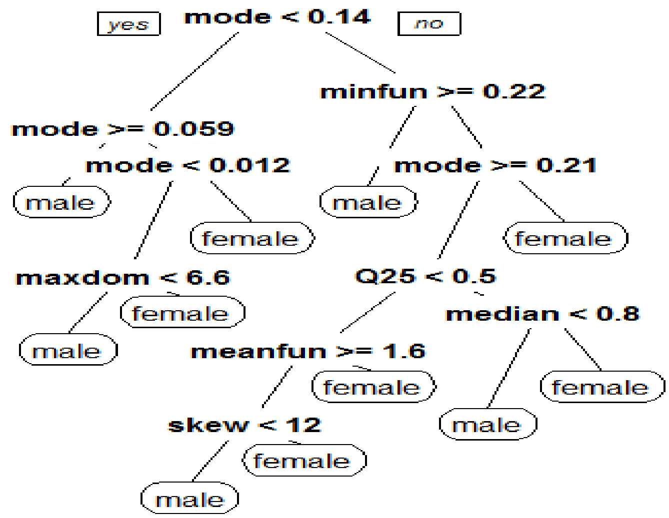


Fig:1 Classifications and Regression Decision Tree (CART) Model

Classification Decision Tree Algorithm

In a classification problem, we have a training sample of n observations on a class variable Y that takes values 1, 2, ... , k, and p predictor variables, X1,..., Xp. Our goal is to find a model for predicting the values of Y from new X values In theory, the solution is simply a partition of the X space into k disjoint sets, A1, A2,..., Ak, such that the predicted value of Y is j if X belongs to Aj , for j = 1, 2,..., k.X takes ordered values, the set S is an interval of the form $(-\infty, c]$. Otherwise, S is a subset of the values taken by X. The process is applied recursively on the data in each child node. Splitting stops if the relative decrease in impurity is below a prespecified threshold. Algorithm 1 gives the pseudo code for the basic steps.

Algorithm 1 Pseudo code for tree construction

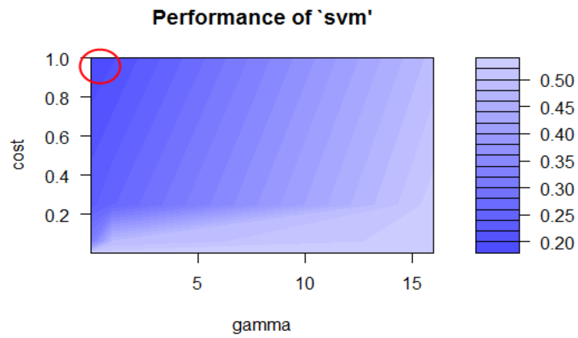
1. Start at the root node.
2. For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split $\{X* \in S*\}$ that gives the minimum overall X and S.
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

SVM Model

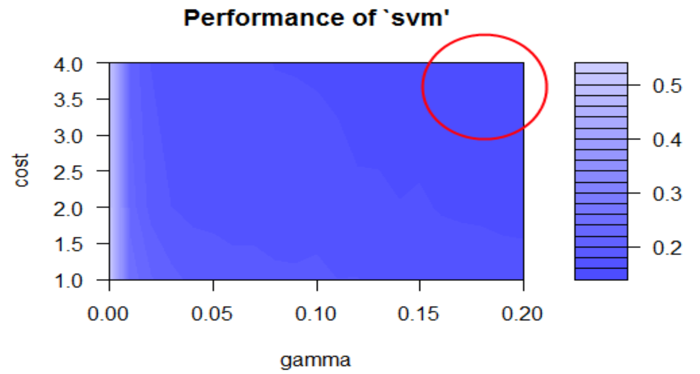
Our next model is a support vector machine, tuned with the best values for cost and gamma. To determine the best fit for an SVM model, the model was initially run with default parameters. A plot of the SVM error rate is then printed, with the darkest shades of blue indicating the best (ie. lowest) error rates. This is the best place to choose a cost and gamma value. You can fine-tune the SVM by narrowing in on the darkest blue range and performing further tuning. This essentially focuses in on the section, yielding a finer value for cost and gamma, and thus, a lower error rate and higher accuracy. The following performance images show how this progresses.

Fig:2 Performance of SVM

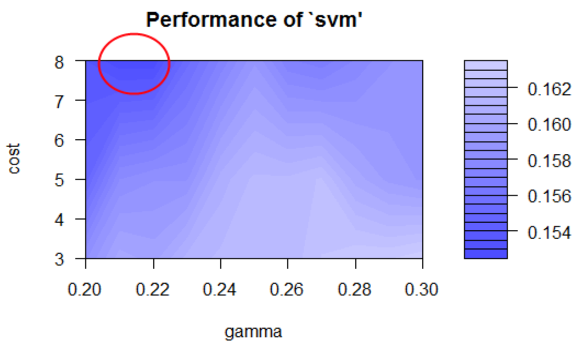
International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018 www.csjournals.com



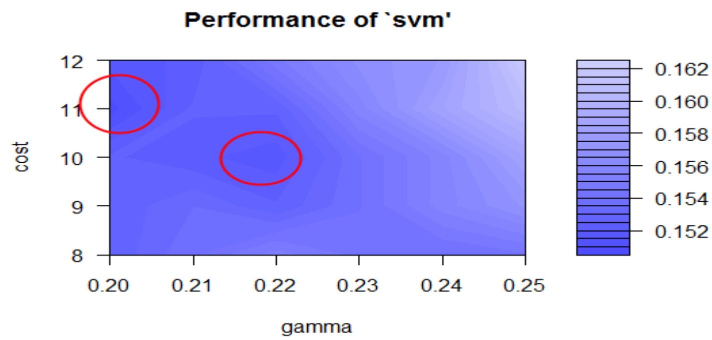
First pass of tuning the SVM.
values are Our best values are around cost 1 and gamma 0.2



Further fine-tuning and our best around cost 4 and gamma 0.2



Zooming in further, our best values are around cost 8 and 0.21 gamma



One more final pass, our best values are around cost 10 and 0.22 gamma

Gender and Emotion Recognition Diagram

All these recognizers are trained in the same manner as the basic emotion recognition system, only the input data or class definition (emotions vs. gender) changes. For the training of the gender-specific emotion systems, only those utterances of the training set that were classified to the respective gender by the gender detection system were used. In the following, the combined gender and emotion detection system will be compared to an emotion recognition system without gender information and to one with information about the correct gender.

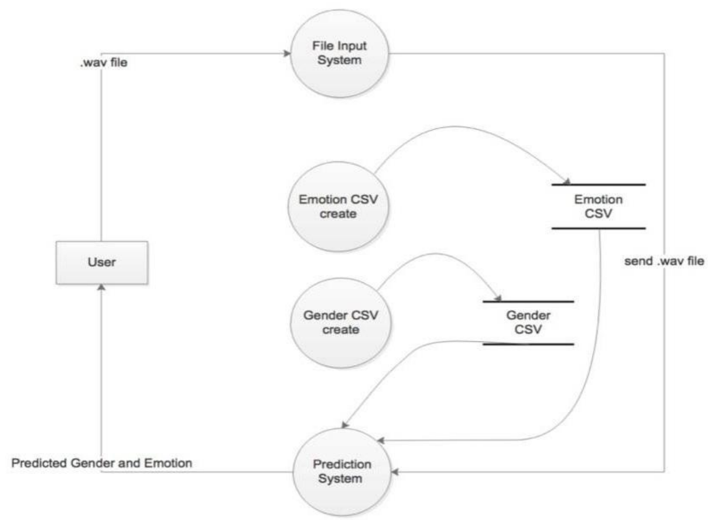


Fig:3 Gender and Emotion Prediction System using in data flow diagram

A Voice emotion recognition system consists of three principal parts, as shown in figure 3: signal processing, feature calculation and classification. Signal processing involves digitalization and potentially acoustic preprocessing like filtering, as well as segmenting the input signal into meaningful units. Feature calculation is concerned with identifying relevant features of the acoustic signal with respect to emotions. Classification, lastly, maps feature vectors onto emotion classes through learning by examples. Voice is converted into .wav file. Voice should be small in size

This paper involves following two functionalities

1. Gender Recognition
2. Emotion Recognition

1. GENDER RECOGNITION

This feature will take input a voice sample from user and analyze it to predict the gender of the user. This feature will train from a dataset of over 3000 voice samples which is made by

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018 www.csjournals.com

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.

1.1 Recording Voice Samples

In this phase, over 3000 voice samples are collected from users either by recording them or by downloading them from internet and storing them separately for male and female voices in folders- Male and Female.

1.2 Extracting Acoustic Properties

In this phase various acoustic properties are extracted from all the voice samples in one go and storing the values in a CSV file.

Acoustic properties measured-

- Duration: length of signal
- Meanfreq: mean frequency (in kHz)
- SD: standard deviation of frequency
- Median: median frequency (in kHz)
- Centroid: frequency centroid (see specprop)
- Peakf: peak frequency (frequency with highest energy)
- Meanfun: average of fundamental frequency measured across acoustic signal
- Minfun: minimum fundamental frequency measured across acoustic signal
- Maxfun: maximum fundamental frequency measured across acoustic signal
- Meandom: average of dominant frequency measured across acoustic signal
- Mindom: minimum of dominant frequency measured across acoustic signal
- Maxdom: maximum of dominant frequency measured across acoustic signal
- Dfrange: range of dominant frequency measured across acoustic signal
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (see note in specprop description)
-

1.4 Creating CSV files

meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.0597809849598081	0.0642412677031359	0.032026913372582	0.0150714886459209	0.0901934398654331	0.0751219512195122	12.86346	274.4029	0.893369416700807	0.491917766397811	0	0.0597809849598081	0.084279106440321	0.0157016683022571	0.275862068965517	0.007812	0.007812	0.007812	0	0	male
0.066008740387572	0.0673100287952527	0.040228734810579	0.0194138670478914	0.0926661901358113	0.0732523230879199	22.42328	634.6138	0.892193242265734	0.513723842537073	0	0.066008740387572	0.107936553670454	0.0158259149357072	0.25	0.00901442307692308	0.007812	0.054687	0.046875	0.0526315789473684	male
0.0773155026958227	0.0838294209445061	0.0367184586699814	0.00870105655686762	0.131908017402113	0.123206960845246	30.75715	1,024.927	0.846389091878782	0.478904979116727	0	0.0773155026958227	0.0987062615673936	0.0156555772994129	0.271186440677966	0.00799005681818182	0.007812	0.015625	0.007812	0.0465116279069767	male
0.151228091724635	0.0721105872627985	0.158011187072716	0.0965817277812306	0.207955251709136	0.111373523927906	1.232831	4.177296	0.963322461535984	0.727231798861951	0.0838781852082039	0.151228091724635	0.0889648485504597	0.017797552836485	0.25	0.201497395833333	0.007812	0.5625	0.554687	0.247119078104994	male
0.135120387296677	0.0791461004935869	0.124656228727025	0.0787202178352621	0.206044928522805	0.127324710687543	1.101173	4.333713	0.971955076212905	0.783568057553871	0.104261	0.135120387296677	0.106397844620363	0.0169312169312169	0.266666666666667	0.712812	0.007812	5.484375	5.476562	0.208273894436519	male
0.132786407306188	0.0795568659729794	0.119089848308051	0.0679579929988331	0.209591598599767	0.141633605600933	1.932562	8.308895	0.963181	0.738307	0.112555425904317	0.132786407306188	0.110131920122721	0.0171122994652406	0.253968253968254	0.298221982758621	0.007812	2.726562	2.71875	0.125159642401022	male

Figure 4 Gender CSV.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018 www.csjournals.com

1.3 Training with Models

In this phase we train the program on this data set using various models and predict gender on a test set. Various models used are—

- Random Forest
- CART Model
- SVM Model

1.5 Shiny APP

In this phase we create web application using Shiny to take input voice sample from user and showing the predicted value of gender using the above models

2. EMOTION RECOGNITION

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.

2.1 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.

2.2 Extracting Acoustic Properties

In this phase various acoustic properties are extracted
from all the voice samples in one go and storing the values in a CSV file.
Acoustic Properties Measured—
•IQR: interquantile range (in kHz)
•skew: skewness (see note in specprop description)
•kurt: kurtosis (see note in specprop description)
•sp.ent: spectral entropy
•sfm: spectral flatness
•mode: mode frequency
•centroid: frequency centroid (see specprop)
•peakf: peak frequency (frequency with highest energy)
•meanfun: average of fundamental frequency measured across acoustic signal
•minfun: minimum fundamental frequency measured across acoustic signal

meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.187383870066221	0.0469853445774683	0.199933110367893	0.193377926421405	0.205083612040134	0.0117056856187291	5.743882	43.30470	0.776211	0.272873522821225	0.199933110367893	0.187383870066221	0.188044346373421	0.0362225519287834	0.265369565217391	1.468654	0	10.37118	10.37118	0.128113026819923	neutral
0.174112385700404	0.0552044792425174	0.190329	0.185494505494505	0.198241758241758	0.0127472527472528	5.743627	43.27521	0.786805505242834	0.304910581770547	0.189890	0.174112385700404	0.175665340434886	0.0247857868020305	0.256989	1.946315	0.190734	8.773781	8.583046	0.0722800925925926	neutral
0.192817169909523	0.0470657844827469	0.2044	0.199266666666667	0.209066666666667	0.0097999999999998	4.896726	28.83642	0.751075544565975	0.248368159314965	0.2058	0.192817169909523	0.182474086649909	0.0345807365439093	0.230320754716981	1.974895	0.035762	9.036041	9.000278	0.0993377483443709	neutral
0.172378824415843	0.0524192429465304	0.186501766784452	0.182049469964664	0.191448763250883	0.0093992932862190 6	6.218212	49.00626	0.772328871141711	0.299858	0.185512367491166	0.172378824415843	0.157103995890483	0.0260833333333333	0.246606060606061	2.041065	0.178813	9.107566	8.928752	0.128170894526035	neutral
0.179164510452416	0.0540929617993851	0.192820512820513	0.187692307692308	0.201538461538462	0.0138461538461538	5.137214	36.33489	0.780520216220941	0.251559858641525	0.191282051282051	0.179164510452416	0.177343949331316	0.025431	0.241722772277228	1.423769	0	9.131408	9.131408	0.135176833610254	neutral
0.170448750021394	0.0509054710366123	0.184044943820225	0.178801498127341	0.187715355805243	0.0089138576779026 3	7.952093	79.84042	0.743390150407703	0.279726381018614	0.183520599250936	0.170448750021394	0.164328305653116	0.0261672025723473	0.230320754716981	1.885669	0.178813	9.238696	9.059882	0.128258145363409	neutral

FIG5. Emotion CSV

2.3 Creating CSV File

All the acoustic properties extracted for all voice samples are stored in a CSV file.

2.4 Training with Models

In this phase we train the program on this data set using various models and predict gender on a test set.

Various models used are—

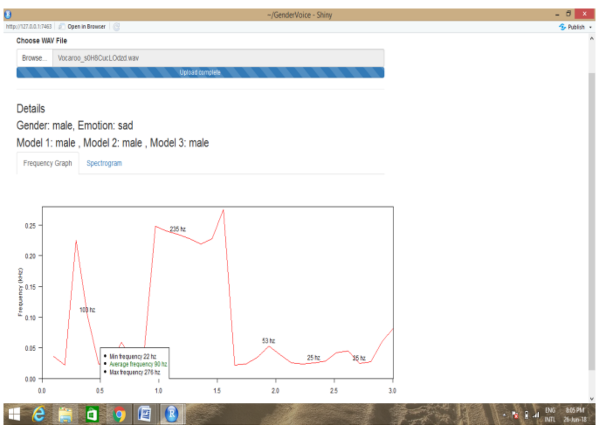
- Random Forest
- CART Model
- SVM Model

2.5 Shiny APP

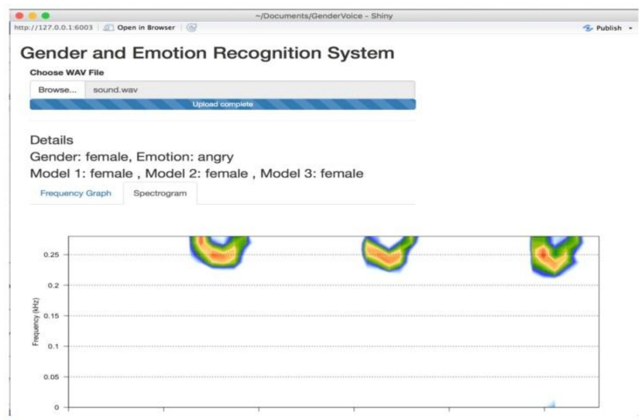
In this phase we create web application using Shiny to take input voice sample from user and showing the predicted value of gender using the above models

EXPERIMENTAL RESULTS

Shiny App Output 1



Shiny App Output 2



The results show that with the employment of a features selection algorithm, a satisfying recognition rate level can still be obtained also reducing the employed features and, as a consequence, the number of operations required to identify the emotional contents. This makes feasible future development of the proposed solution over mobile devices. The obtained results underline that our system can be reliably used to identify a single emotion, or emotion category, versus all the other possible ones.

CONCLUSION

The proposed system, able to recognize the emotional state of a person starting from audio signals registrations, is com- posed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former has been implemented by a Acoustic properties 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 system provides facility to determine a person's gender and emotion from their voice sample provided in '.wav' format. The system predicts the gender and emotion accurately for most cases. The system is provided with 3000+ voice samples divided as male, female for gender recognition model building. It has an accuracy of 97%. Over 1000 voice samples are provided for emotion recognition model building. It predicts between 4 emotions which are neutral, angry, sad, fear. This makes feasible future development of the proposed solution over mobile devices. The obtained results underline that our system can be reliably used to identify a single emotion, or emotion category, versus all the other possible ones.

REFERENCES

[1] A. Vinciarelli, M.Pantic, and H.Bourlard. (2009, Nov.).''Social signal processing: Survey of an emerging domain,'' Image Vision Comput. [Online]. 27(12), pp. 1743–1759.Available: http://dx.doi.org/10.1016/j.imavis.2008.11.007

[2]Z.Zeng, M. Pantic, G. Roisman, and T. Huang, ''A survey of affect recognition methods: Audio,visual, and spontaneous expressions,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[3] A. Gluhak, M. Presser, L. Zhu, S. Esfandiyari, and S.Kupschick, ''Towards mood based mobile services and applications,'' in Proc. 2nd Eur. Conf. Smart Sens. Context, 2007, pp. 159–174.

[4]K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C.Longworth, and A. Aucinas, ''Emotionsense: A mobile phones based adaptive platform for experimental social psychology research,'' in Proc. UbiComp, 2010, pp. 281–290.

[5]M. El Ayadi, M. S. Kamel, and F. Karray, ''Survey on speech emo- tion recognition: Features, classification schemes, and databases,'' Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.

[6]R. Fagundes, A. Martins, F. Comparsi de Castro, and M. Felippetto de Castro, ''Automatic gender identification by speech signal using eigenfil- tering based on Hebbian learning,'' in Proc. 7th Brazilian SBRN, 2002, pp. 212–216.

[7]Y.-M.Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, ''Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech,'' in Proc. Int. Conf. Mach. Learn. Cybern. 2006, pp. 3376–3379.

[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.

[9]F. Yingle, Y. Li, and T. Qinye, ''Speaker gender identification based on combining linear and nonlinear features,'' in Proc. 7th WCICA 2008, pp. 6745–6749.

[10]H. Ting, Y. Yingchun, and W. Zhaohui, ''Combining MFCC and pitch to enhance the performance of the gender recognition,'' in Proc. 8th Int. Conf. Signal Process., vol. 1. 2006.