VOICE FORENSICS REPORT

An Intelligent Tool for Criminal Investigations

OBJECTIVE

To build a voice forensics system that would identify bodily features such as height, weight, age, sex, region of origin and various other demographic information about a miscreant from the voice evidence collected. The end objective is to build an extensive, if not comprehensive, one-of-a-kind voiceprint database to enable authorities to track criminals.

ABSTRACT

Security has become a great concern for the citizens of our nation. With incidents such as bomb attacks, ransom calls, and threat calls to life and property occurring more frequently, it is important to develop a mechanism to help curb them. It is vital for the government to devise a mechanism to deal with threats and ransom calls in an effective and promising way. Voice Forensics has potential to help the law enforcement agencies by providing valuable information such as height, weight, age, sex of the suspect from the voice evidence available. In the current scenario of crime investigation in India, we are technologically ill-equipped to investigate cases that have only audio as their evidence. Our project tries to solve this problem. Through this project we wish to explore and improvise the area of Voice Forensics. The ultimate aim of the project is to equip law enforcement agencies with the tools to process voice samples and provide physical and demographic information about the miscreant that could be used as an important evidence for investigation purposes. We propose to build a unique one-of-a-kind voice print database for further research and analysis.

INTRODUCTION

During the process of criminal investigations, it is imperative to extract as much information as possible from the available evidences. Currently, the National Crime Records Bureau cites two methods of Criminal identification, one using fingerprints, and the other is a portrait building system. Fingerprint matching could provide accurate information about the criminal, but in cases where evidence is not available, or if the person is not recorded in the database, we will not be able to make any predictions. In case of fingerprints, it is impossible to approximate predictions about the person, if he/she is not recorded in the database. Presently, there are 11 divisions under the CBI for forensics and crime investigations in India. Surprisingly, voice forensics is not one of them yet. With the technology we are developing, it would be possible for the CBI to investigate cases with the evidences obtained from voice and speech also. With this tool, voice could be used as a reliable evidence in a court proceeding as per Section 65B of the Indian Evidence Act, 1972.

Exploration of voice as a possible evidence is quite recent, and there are some advanced voice identification software's being developed, such as VoiceGrid. While, voice based technologies such as Siri and Cortana are used as personal digital assistants in mobile phones, VoiceGrid is a database intensive tool that has been adopted by various state police organizations in the USA and Russia for identification of miscreants based on the voice sample captured. These systems rely largely on an existing database to make exact or close-to matches. However, in cases when the exact voice samples cannot be matched, or is unavailable in the database, it is very useful to extract physical and geographical information of the miscreant from the voice sample available. Hence, there is scope to develop much smarter and efficient systems for the purpose of voice forensic study. In addition to this problem, there are no publicly available benchmarks to test an attribute identification method. This is mainly due to the difficulty in procuring a large dataset for the models to work on and the absence of a framework for testing. Moreover, there exists no framework that does the work of:

1. Collecting a large quantity of audio data from the citizens of our nation
2. Storing, Analyzing, Validating the audio samples collected and managing it securely.
3. Perform formal research on the collected voice samples. There is no framework that allows for testing different models that predict physical attribute of a person from their voice.
4. Provides aids to the work of researchers across the country to use this nationwide audio database for other interesting applications. (Anonymity of persons will be maintained for security purposes).

Within this project, we propose to build a framework that would solve these problems. We aim to build the necessary technology for voice forensics and investigation. The long term aim of this project is to equip law enforcement agencies with the required tools to perform voice forensics and provide necessary evidence for enforcement of law and order. With the system we build, the officials should be able to estimate with good accuracy, the physical and geographical features of the suspect.

DATA COLLECTION

Voice samples were collected from 40 students who participated in the IPTSE CMU-NITK Winter School 2015. The age group of the participants was in the range of 19-22 years. The height, weight, age and sex of the students were recorded. Each student was asked to speak a set of 25 phonetically rich sentences randomly selected from the large TIMIT database. Thus, there were 25 recordings per person, making a total of 1000 recordings. The samples were recorded using an external microphone on Audacity, in a relatively quiet room. We ensured that the recordings were lossless. All other necessary conditions like distance between the speaker and the microphone were taken care of while recording the voice samples.

METHODOLOGY

The Framework we are developing consists of machine learning tools, classification and regression algorithms that extract and analyze features of the voice and learn the correlations of the physical features and voice of the speaker. The framework depicts the pipeline of computations and analysis. The pipeline mainly consists of the following:

1. Feature Extraction
2. Normalization of data
3. Clustering (Bag of Words Model)
4. Machine Learning Algorithms
   - Classifier Models
   - Regression Models

The pipeline followed is depicted in the picture demonstrated below (see Fig 1.1). The following sections will explain the above sections in detail.

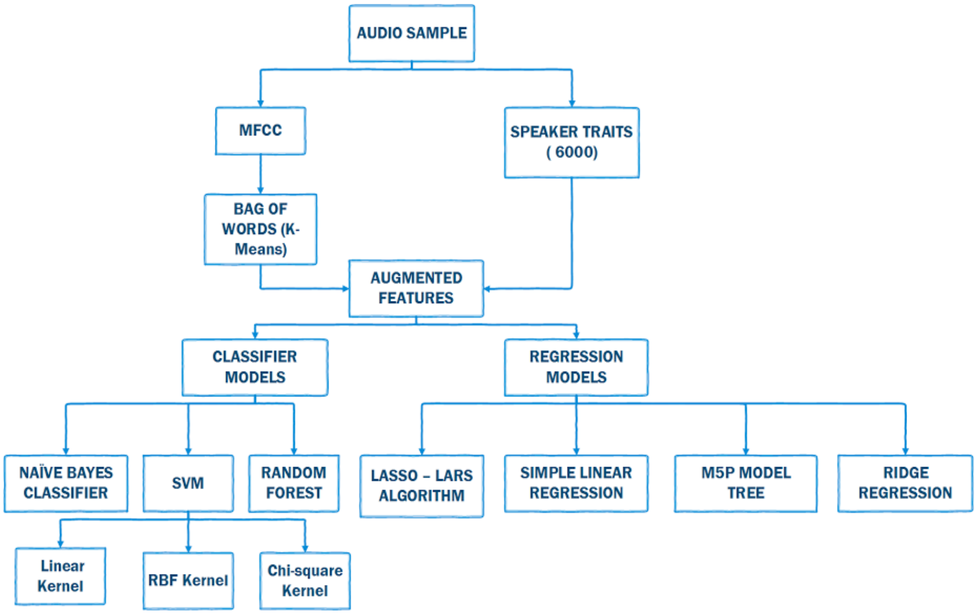


Fig 1.1. The Pipeline of algorithms and analysis performed.

Please Note: This framework will be deployed on a public platform and will be available for download and use for research purposes. (We will maintain anonymity of the voice samples for security reasons).

FEATURE EXTRACTION

To extract features from the voice, we used an open source tool called OpenSMILE. This enables us to extract large audio feature spaces in real time. The speech related features were Signal energy, Loudness, MFCC, PLP-CC, Pitch, Voice quality (Jitter, Shimmer), Formants, LPC, Line Spectral Pairs (LSP) and Spectral Shape descriptors. For our system, we have mainly used Mel frequency Cepstral Coefficients (MFCC) and IS12 Speaker traits which include around 6000 features for each audio sample. The 39 MFCC coefficients extracted were used directly in Weka Tool, without implementing the bag-of-words model. As expected, this resulted in lower accuracies.

NORMALIZATION OF DATA

The data was normalized across each coefficient before applying k-means. MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values to lessen the influence of noise.

M = Sample mean of old values across a coefficient
V = Variance of old values across a coefficient
O = Old value of a coefficient in a coefficient vector
N = New value of a coefficient in a coefficient vector

N = (O-M)/V

A 200 bag model was used, and the histogram generated for each sample, was considered to be the contribution for that sample. The data was normalized across the 200 bins in the histogram in the following way. For a histogram,

New value in bin = Old value/sum of old values across all the bins in histogram

Data was also normalized in each bin of the histograms to make the mean 0 and the variance 1.For a particular bin in a histogram,

M = Sample mean of old values in bin across all histograms
V = Variance of old values in bin across all the histograms
O = Old value of bin in a histogram
N = New value of bin in a histogram

Then we have,

N = (O-M)/V

The bag of words model was implemented in C++, using octave for normalizations. Further calculations were carried out on the bag of words only and not the original data.

CLUSTERING & BAG OF WORDS

The extracted features are clustered using a vector quantization method called K-means clustering. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

39 MFCC coefficients are extracted per frame, and the frame interval is 25ms and the audio is sampled every 10ms. Since the length of each recorded sample was different, we obtained varying number of feature vectors for each sample. We could average the coefficients over all the frames of a sample to obtain a single feature vector, but we would lose out on a lot of information in the process. Thus, we resorted to a bag of words model for MFCC coefficients. The bag of words model is implemented over the clustered features and each audio clip is represented in the form a histogram. The bins of the histogram consist of our new feature space and the histograms themselves become our new datasets.

The 6000 features are obtained over the entire length of the sample, by averaging some features out. Therefore we obtain only one feature vector for each recording, but we might lose information about the distribution of these features, since they are averaged out Using K-means and Bag of Words model on MFCC coefficients, we can get a single feature vector for one file. This can then be augmented with the 6000 features and the combined set of features can be used for training any machine learning model.

MACHINE LEARNING ALGORITHMS

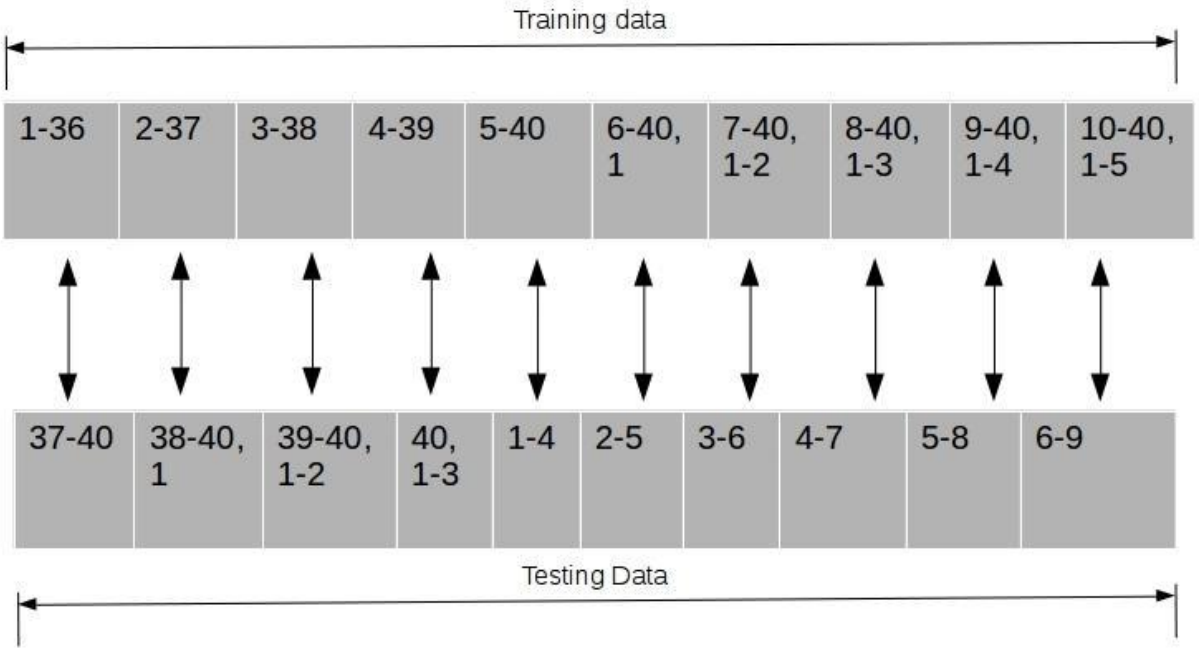
Our next step was to test various relevant machine learning algorithms and find that one algorithm or a combination of such algorithms that works the best on our system. We set up a platform consisting of all these algorithms and devised an iterative pipeline for the data flow. The algorithms were grouped based on their outcome: Classification and Regression. We fed the histograms that were generated from our bag of words model as the input to these algorithms. The various algorithm used are listed below along with their results.

TRAINING AND TESTING DATA

MODIFIED CROSS VALIDATION

Weka does stratified cross validation by default. However, there is a problem with this for our data set. Given a certain training set, we want the algorithm to test for efficiency on a “new” test set. In cross validation, there could be a possibility of a certain speaker’s voice sample occurring both in the test and training set. For ex: out of the 25 samples of any one subject, 12 could be in training and 13 in the test set. It would be very easy for the model to get a high accuracy on such test samples. To avoid this, the data was split manually into 10 folds and cross validated. To accomplish this, we developed a Java API for Weka. We incorporated the Java API in our framework to automate the whole process of training and k-fold cross validation on our dataset.

The diagram is an example of 10 fold splitting (k=10) of data set into training and testing pairs, for modified cross validation. The numbers in the table indicate subject numbers (1-37 indicates subject 1 to subject 37).



Training data	1-36	2-37	3-38	4-39	5-40	6-40, 1	7-40, 1-2	8-40, 1-3	9-40, 1-4	10-40, 1-5
Testing Data	37-40	38-40, 1	39-40, 1-2	40, 1-3	1-4	2-5	3-6	4-7	5-8	6-9

CLASSIFICATION MODELS

The classification algorithms used were Support Vector Machines, Naive Bayes Classification and Random Forest.

Random Forests

The Random Forest classifier provided by Weka was used for this. A forest of 500 trees and modified cross validation was used for testing the performance of the model.

SVM with chi-square kernel

We used libsvm toolkit to implement the Support Vector Machine. The C-SVM classification or classification type 1 with a pre-computed chi-square kernel was selected for our purpose. This was implemented for the bag of words model only and not for 6000 features, since chi-square kernel is generally used only with histograms. The chi-square kernel was used to transform data from the input (independent) to the feature space. The larger the C, the more the error is penalized. Thus, C was chosen with care to avoid over fitting.

Following are the results for different values of C's for classification based on gender:

C	AUC	Accuracy	Recall	Precision
10 ^ -5	0.418667	68.00%	16.00%	26.67%
10 ^ -4	0.416	66.00%	20.00%	26.32%
10 ^ -3	0.408	65.00%	20.00%	25.00%
10 ^ -2	0.408	65.00%	20.00%	25.00%
10 ^ -1	0.425067	65.00%	20.00%	25.00%
10 ^ 0	0.425067	65.00%	20.00%	25.00%
10 ^ 1	0.425067	65.00%	20.00%	25.00%
10 ^ 2	0.425067	65.00%	20.00%	25.00%
10 ^ 3	0.425067	65.00%	20.00%	25.00%
10 ^ 4	0.425067	65.00%	20.00%	25.00%
10 ^ 5	0.425067	65.00%	20.00%	25.00%

On doing the same thing for height based classification, we got the following table:

C	AUC	Accuracy	Recall	Precision
10 ^ -5	0.3904	72.00%	0.00%	0.00%
10 ^ -4	0.414933	52.00%	12.00%	10.34%
10 ^ -3	0.4144	53.00%	12.00%	10.71%
10 ^ -2	0.4176	52.00%	12.00%	10.34%
10 ^ -1	0.4176	52.00%	12.00%	10.34%
10 ^ 0	0.4176	52.00%	12.00%	10.34%
10 ^ 1	0.4176	52.00%	12.00%	10.34%
10 ^ 2	0.4176	52.00%	12.00%	10.34%
10 ^ 3	0.4177	52.00%	12.00%	10.34%

10 ^ 4	0.4177	52.00%	12.00%	10.34%
10 ^ 5	0.4177	52.00%	12.00%	10.34%

Note: C is the penalty parameter and AUC is the Area under the curve in ROC.

Naive Bayes Classification

We used the classifier provided by Weka for this. The assumption is that the continuous values associated with both classes are distributed according to a Gaussian distribution.

REGRESSION MODELS

We used regression to predict the height from the voice sample. We used the following regression model:

Simple Linear Regression

We performed Simple linear regression (in Weka) on the data set for height estimation. Coefficient estimates for the models described in Linear Regression rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the matrix $(XTX)^{-1}$ becomes close to singular. As a result, the least-squares estimate $\beta = (XTX)-1XTy$ becomes highly sensitive to random errors in the observed response y, producing a large variance. Hence we had to try out a different algorithm.

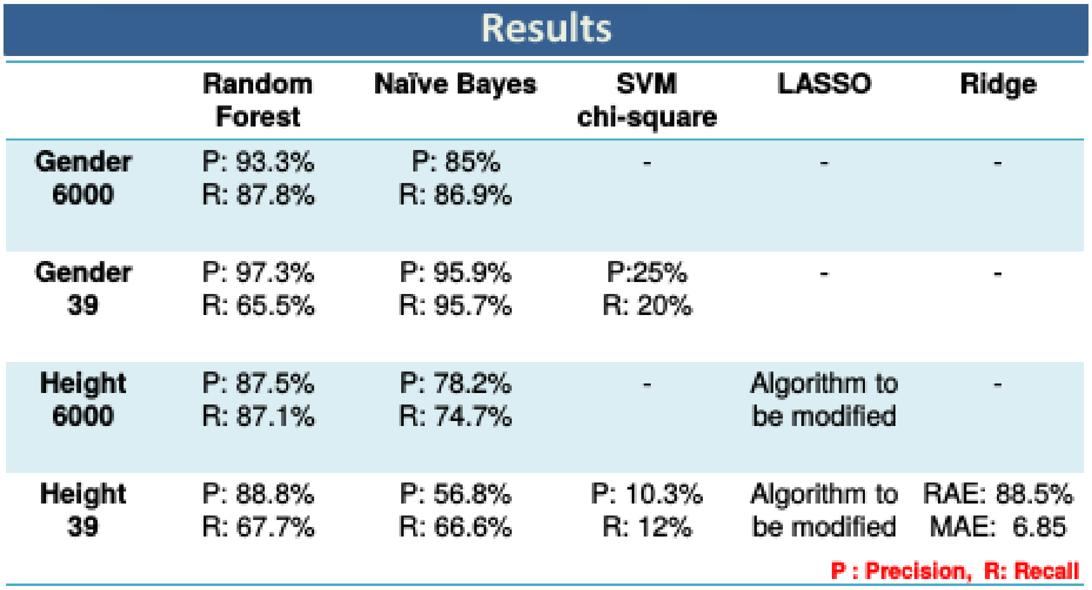
Ridge Regression

A code was written in octave to implement ridge regression. The results obtained using Simple linear regression were sufficiently accurate. Hence, we combined it with a form of regularization function in order to pull out the important features that correlate the most with height. The problem of the matrix becoming singular is also resolved by using ridge regression.

RESULTS

The initial results that we obtained was itself a proof of concept for what we were trying to build. Given that the data set we used to test our system was meagre and biased (male-female ratio was 3:1), we were still able to generate results with good accuracy.

We could predict the gender of an unknown person’s voice with an accuracy of 95.2% and predict his/her height with an error of 6.5cm. With more data, and fine-tuning, our system could become reliable enough to finally reach our desired goals.



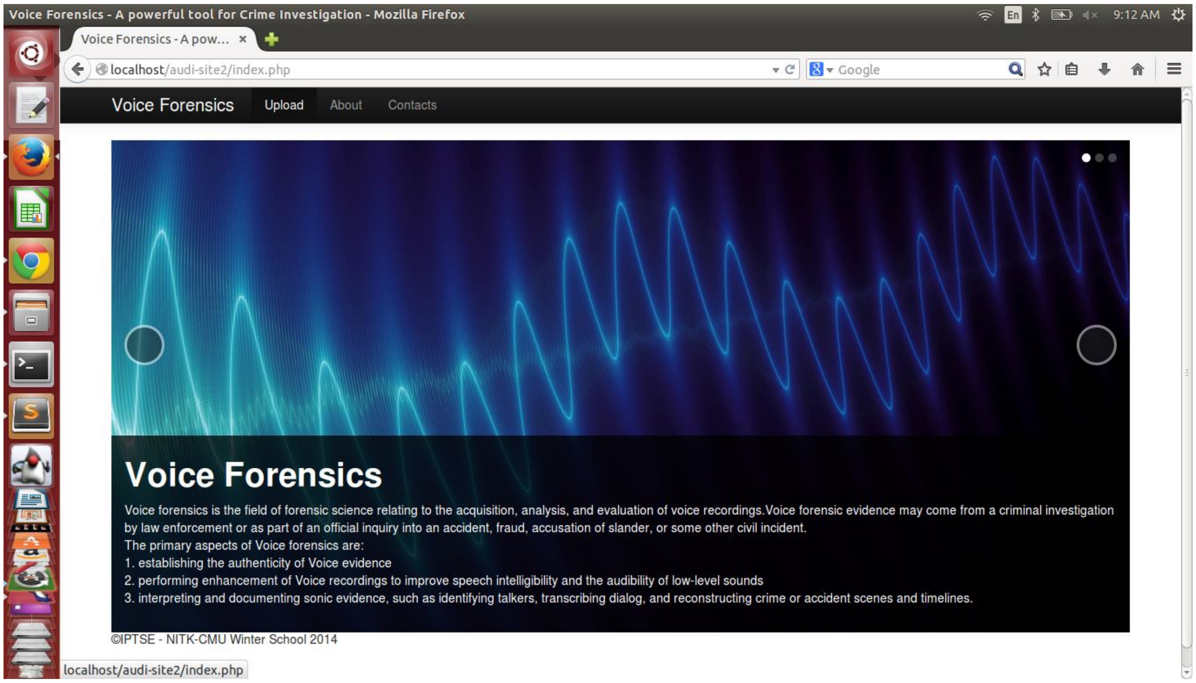
Results

	Random Forest	Naïve Bayes	SVM chi-square	LASSO	Ridge
Gender 6000	P: 93.3% R: 87.8%	P: 85% R: 86.9%	-	-	-
Gender 39	P: 97.3% R: 65.5%	P: 95.9% R: 95.7%	P:25% R: 20%	-	-
Height 6000	P: 87.5% R: 87.1%	P: 78.2% R: 74.7%	-	Algorithm to be modified	-
Height 39	P: 88.8% R: 67.7%	P: 56.8% R: 66.6%	P: 10.3% R: 12%	Algorithm to be modified	RAE: 88.5% MAE: 6.85

P : Precision, R: Recall

WEBSITE

To make our tool publicly usable, we have developed a website. The website allows a user to upload a voice sample (only .wav files are accepted as of now) and outputs the physical characteristics of the owner of that voice in the sample. To predict the physical features, the voice sample inputed is run on the already trained model. In future, we intend to make a provision for users to contribute training data as well. To ensure authenticity and security, only validated users shall be allowed to upload their voice samples and their physical characteristics. After inspection of the samples collected from the website for genuineness, it will be used for training of our models.

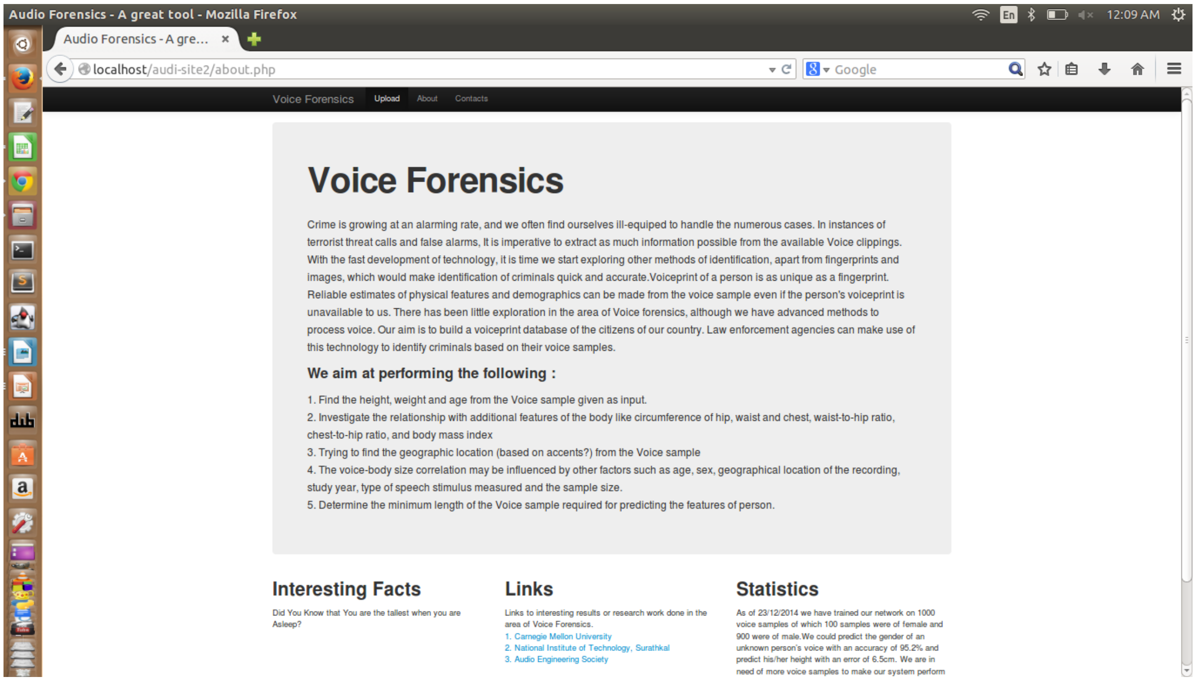

Voice Forensics - A powerful tool for Crime Investigation - Mozilla Firefox
En
9:12 AM
Voice Forensics - A pow... ×
localhost/audi-site2/index.php
Google
Voice Forensics
Upload
About
Contacts
localhost/audi-site2/index.php

Voice Forensics

Voice forensics is the field of forensic science relating to the acquisition, analysis, and evaluation of voice recordings.Voice forensic evidence may come from a criminal investigation by law enforcement or as part of an official inquiry into an accident, fraud, accusation of slander, or some other civil incident.
The primary aspects of Voice forensics are:
1. establishing the authenticity of Voice evidence
2. performing enhancement of Voice recordings to improve speech intelligibility and the audibility of low-level sounds
3. interpreting and documenting sonic evidence, such as identifying talkers, transcribing dialog, and reconstructing crime or accident scenes and timelines.

©IPTSE - NITK-CMU Winter School 2014

Fig 2.1. The Landing page of the website.


Audio Forensics - A great tool - Mozilla Firefox
En
12:09 AM
Audio Forensics - A gre... ×
localhost/audi-site2/about.php
Google
Voice Forensics
Upload
About
Contacts

Voice Forensics

Crime is growing at an alarming rate, and we often find ourselves ill-equiped to handle the numerous cases. In instances of terrorist threat calls and false alarms, It is imperative to extract as much information possible from the available Voice clippings. With the fast development of technology, it is time we start exploring other methods of identification, apart from fingerprints and images, which would make identification of criminals quick and accurate.Voiceprint of a person is as unique as a fingerprint. Reliable estimates of physical features and demographics can be made from the voice sample even if the person's voiceprint is unavailable to us. There has been little exploration in the area of Voice forensics, although we have advanced methods to process voice. Our aim is to build a voiceprint database of the citizens of our country. Law enforcement agencies can make use of this technology to identify criminals based on their voice samples.

We aim at performing the following :

1. Find the height, weight and age from the Voice sample given as input.
2. Investigate the relationship with additional features of the body like circumference of hip, waist and chest, waist-to-hip ratio, chest-to-hip ratio, and body mass index
3. Trying to find the geographic location (based on accents?) from the Voice sample
4. The voice-body size correlation may be influenced by other factors such as age, sex, geographical location of the recording, study year, type of speech stimulus measured and the sample size.
5. Determine the minimum length of the Voice sample required for predicting the features of person.

Interesting Facts

Did You Know that You are the tallest when you are Asleep?

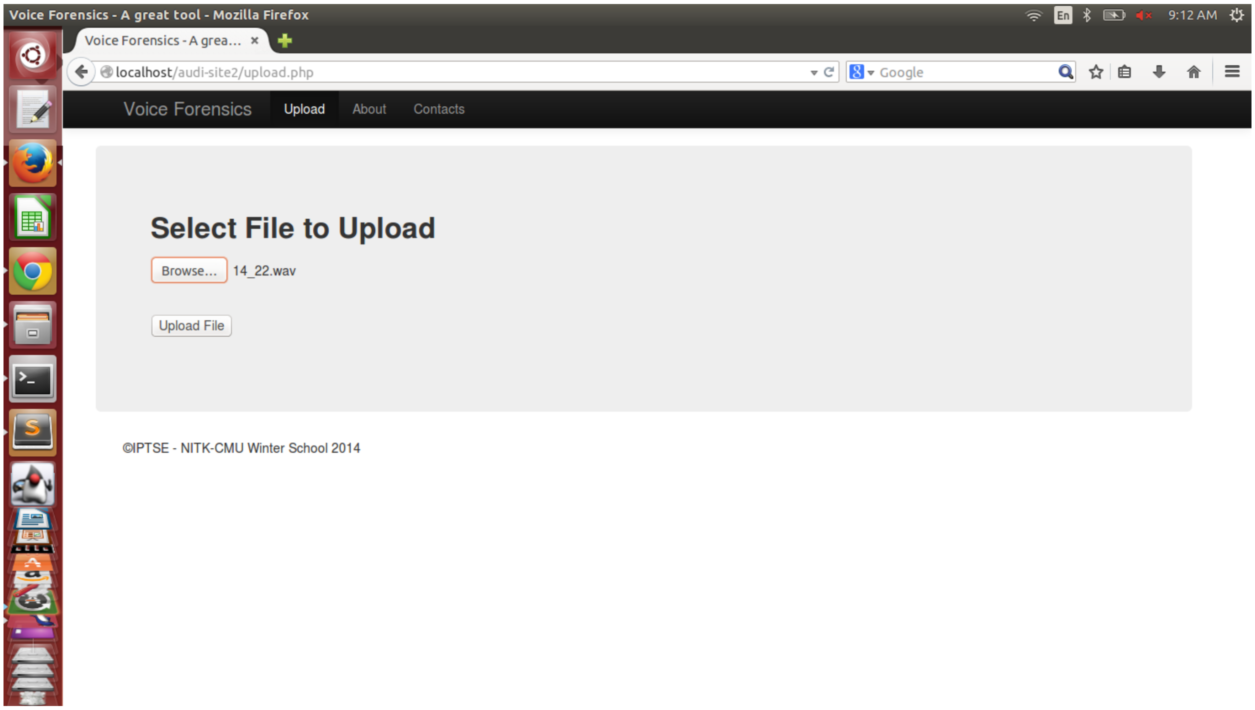
Links

Links to interesting results or research work done in the area of Voice Forensics.
1. Carnegie Mellon University
2. National Institute of Technology, Surathkal
3. Audio Engineering Society

Statistics

As of 23/12/2014 we have trained our network on 1000 voice samples of which 100 samples were of female and 900 were of male.We could predict the gender of an unknown person's voice with an accuracy of 95.2% and predict his/her height with an error of 6.5cm. We are in need of more voice samples to make our system perform

Fig 2.2. The about page

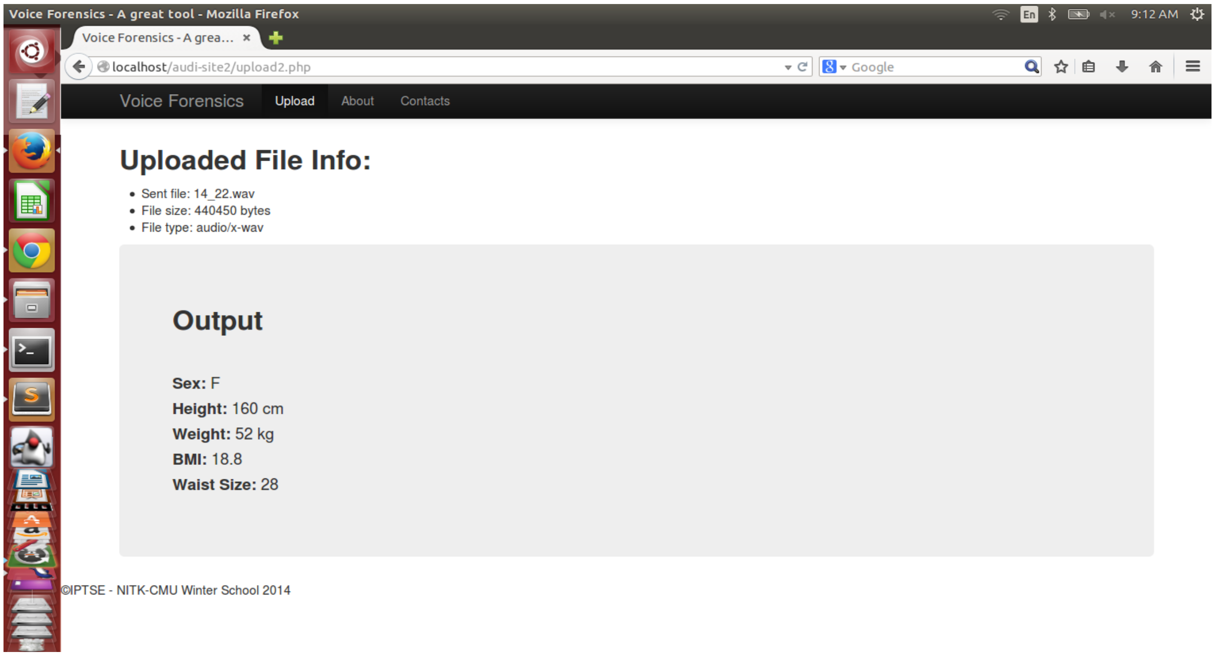

Voice Forensics - A great tool - Mozilla Firefox
9:12 AM
Voice Forensics - A grea...
localhost/audi-site2/upload.php
Google
Voice Forensics
Upload
About
Contacts
Browse...
Upload File

Select File to Upload

14_22.wav

©IPTSE - NITK-CMU Winter School 2014

Fig 2.3. The upload page for voice samples


Voice Forensics - A great tool - Mozilla Firefox
9:12 AM
Voice Forensics - A grea...
localhost/audi-site2/upload2.php
Google
Voice Forensics
Upload
About
Contacts

Uploaded File Info:

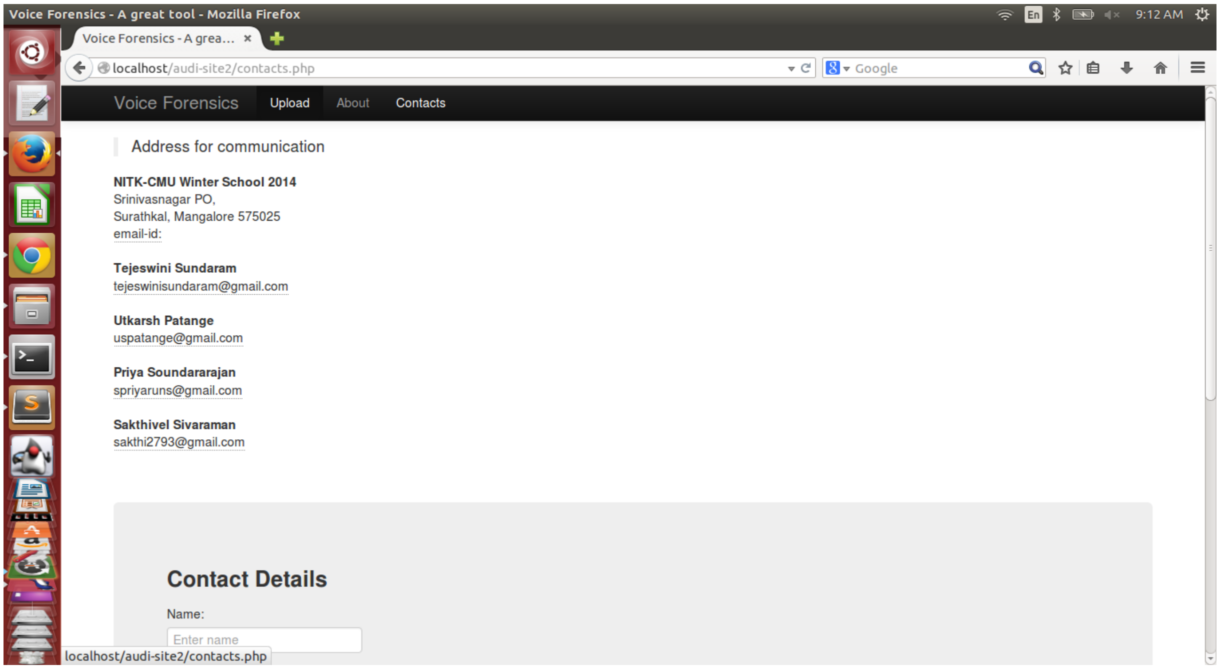
- Sent file: 14_22.wav
- File size: 440450 bytes
- File type: audio/x-wav

Output

Sex: F
Height: 160 cm
Weight: 52 kg
BMI: 18.8
Waist Size: 28

©IPTSE - NITK-CMU Winter School 2014

Fig 2.4. The results of the voice sample uploaded processed by our voice forensics tool.


Voice Forensics - A great tool - Mozilla Firefox
9:12 AM
Voice Forensics - A grea...
localhost/audi-site2/contacts.php
Google
Voice Forensics
Upload
About
Contacts
localhost/audi-site2/contacts.php

Address for communication

NITK-CMU Winter School 2014
Srinivasnagar PO,
Surathkal, Mangalore 575025
email-id:

Tejeswini Sundaram
tejeswinisundaram@gmail.com

Utkarsh Patange
uspatange@gmail.com

Priya Soundararajan
spriyaruns@gmail.com

Sakthivel Sivaraman
sakthi2793@gmail.com

Contact Details

Name:
Enter name

Fig 2.5. The contact page.

FURTHER WORK

- LASSO regression for height estimation,
- Augmenting 6000 features(speaker traits) with bags of words features,
- Since we have got high accuracy for gender classification, we would now hope to see better results by using the predicted gender itself as a feature for height prediction.
- The data collected was biased, we had a girls to boys ratio of 1:3. We need to test our models on a larger dataset with unbiased inputs and check for the performance.

ACKNOWLEDGMENT

We would like to extend our gratitude to our guides Prof Bhiksha Raj, Prof Rita Singh from CMU and Mr. Pulkit Agrawal, PhD student from University of California, Berkeley. A special thanks to the entire IPTSE Winter School Team for providing us the opportunity and resources to work on this project.

Team Voice Forensics:

1. Tejeswini Sundaram, BTech Computer Science, MIT Manipal
2. Priya Soundararajan, Int. M.Sc. Applied Mathematics, IIT Roorkee
3. Utkarsh Patange, BTech Computer Science, IIT Kanpur
4. Sakthivel Sivaraman, BTech Mechanical Engineering, NITK Surathkal