Voice Gender Identification System

Leanne Feng
University of Victoria
leanne_feng@yahoo.com

Xinyue Liu
University of Victoria
lichtbalabala@outlook.com

Zhenyu Zhang
University of Victoria
kurohasky0925@gmail.com

Abstract

This project is aimed at detecting whether an audio sample contains a male or female voice. To analysis and identify the gender of a person in a recording, we plan to use collections of voice and speeches as training dataset. This report is a brief explanation of this project including the purpose and ideal result, the progress of system design and compile and a discussion on future works.

1. Introduction

Voice gender detection, being one of the most frequently studied topics in MIR, is interesting for programmers to investigate and helpful in security activities and in rescue operations. Detecting a person's voice gender seems to be a easy task for human. However, it is necessary to teach computer to predict voice gender. As human vocal range is around 80 - 1400HZ,[7][8] we can extract human voice from noisy background voice and identify voice gender, which can be used for criminal investigation.

In this project, we proposed a computer program for automatic gender detection in both monophonic and polyphonic recordings. This database consists of over 3000 audio pieces, collected from male and female recordings. Analyzed and generated a model based on a set of audio data, the classifier could apply the model on the input audio file which need to be detected. Given enough training data and sufficient time to train, models with high accuracy should be able to classify the gender of a voice in a recording with a quite high accuracy. [1]

2. Dataset

The dataset is formed by features of voice audio samples from each gender. We found this dataset from Kaggle datasets, which consists one target variable and 20 independent variables [19]:

1 target variable:

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)

meanfun: mean fundamental frequency measured across acoustic signal

minfun: minimum fundamental frequency measured across acoustic signal

maxfun: maximum fundamental frequency measured across acoustic signal

meandom: mean 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

modindx: modulation index

Although the vocal range turns out to be different between different language

speakers[14][15], we will use English language be a major language this time. [3] There will be some pre-progress on the dataset such as filter the statistics by deleting rows with empty feature/ null column[16].

3. Method

After some pre-progress on dataset, we will try to generalizing and analysing those variables and compared with gender (which the voice gender of samples is known). We tend to use some math function in kiwi which is a java-based data analysing software and a python library called scikit-learn which contains many classifier models can be used to predict.

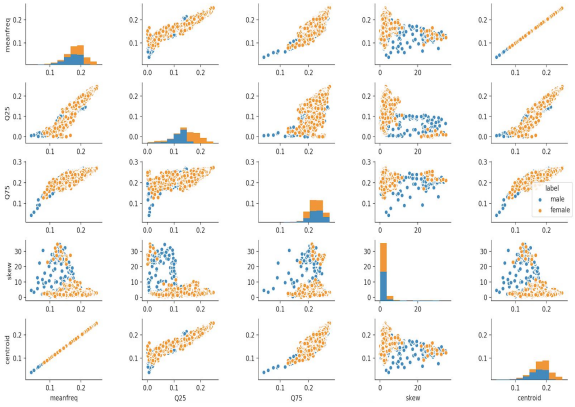
3.1 Pre-processes

3.1.1 Dataset Filtering

Although we deleted the null rows in dataset, 20 features are still too much for a simple voice identification system. The next step is filtering the features in dataset and exclude some features with less relationship.[16] Those features can be figured out from the plot of feature-gender. (Figure. 1)

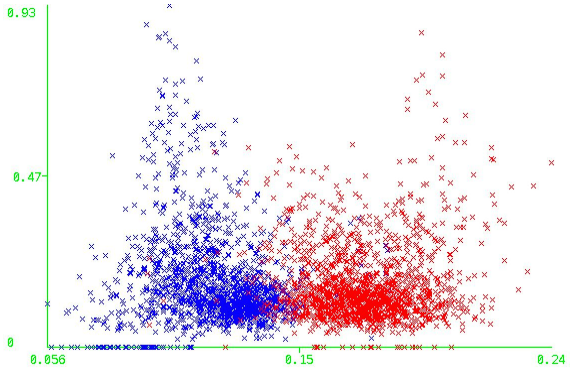
```
ds = pd.read_csv('/Users/llicht/Documents/csc475/project/voice.csv')
ds.head()


g = seaborn.pairplot(ds[['meanfreq', 'Q25', 'Q75', 'skew',
                         'centroid', 'label']],hue='label', size=2)
plt.show()
```

(Figure. 1)



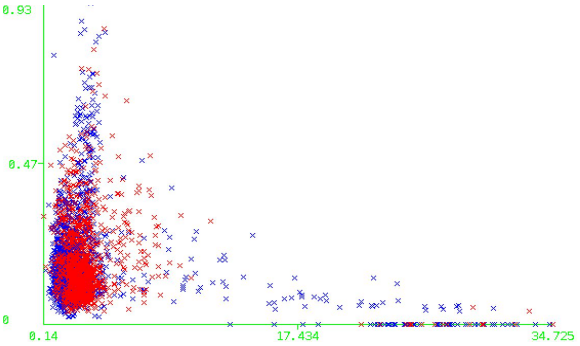
(Figure. 2)

As the graph shown above (Figure. 2), the relationship between two features with gender is not clearly distinguished. So we change another way to make plot and observe. We import the dataset into weka and use visualizing function to see the feature-feature plot with gender in different color. The plot are shown below:



(Figure. 3)

Variable “meanfun” considers to be good to use since it clearly shows the difference between male and female. (Figure. 3)



(Figure. 4)

Above is a variable that can barely observe the difference. In the dataset, the variables “skew”, “smf”, “mode”, and “meandom” with such overlapping feature will not consider to be used in the code.

3.1.2 Input Processing

The main part of input processing is standardize vocal features. Standardizing

vocal features eliminate the dimension relation between features, so we can compare the data. In our work, directly using tuneR, seewave and fftw library from R approaches the purpose of standardizing the data. [9] [10] A function specprop() to return a list of statistical properties of a frequency spectrum (Figure. 5), then fund() is applied to track the fundamental frequency of "meanfun", "minfun" and "maxfun". (Figure. 6) To obtain features' data of "meandom", "mindom" and "maxdom", by using, dfreq(), returning the dominant frequency of a time wave. (Figure. 7)

```
tuneWave <- readWave(file.path(getwd(), wave),
                     from = start, to = end, units = "seconds")
waveSpec <- spec(tuneWave, f = tuneWave@samp.rate, plot = F)
analysis <- specprop(waveSpec, f = tuneWave@samp.rate,
                     flim = c(0, humanFrequency / 1000), plot = F)
```

(Figure. 5)

```
fundamental <- fund(tuneWave, f = tuneWave@samp.rate, ovlp = 50,
                    threshold = 5, wl = 2048, ylim = c(0, humanFrequency / 1000),
                    fmax = humanFrequency, plot = F)
```

(Figure. 6)

```
dom <- dfreq(tuneWave, f = tuneWave@samp.rate, wl = 2048,
                  ylim = c(0, humanFrequency / 1000),
                 ovlp = 0, threshold = 5, bandpass = b * 1000,
                           fftw = T, plot = F)[, 2]
```

(Figure 7)

With the standardized features, we can then applied our input data with Random Forest model to be predicted.

3.2 Modeling and Predicting

Data set first should be separated with training data and test data for evaluating trained model. Several classifiers such as naive bayes Bernoulli, random forest, CART and Linear Discriminant Analysis, will be used in the aim of producing the best accuracy.

3.2.1 Classifiers

The models below are all applied on the dataset and get stats of accuracy. Compared these accuracy with each other, random forest with the highest accuracy is chosen to predict gender.

Classifier name	accuracy
rpart	0.9589905
knn	0.6899054
glm	0.9744479
C5.0	0.9791798
ctree	0.9678233
ida	0.9694006
Svm linear	0.9763407
Random forest	0.9823344

The query below (Figure. 8) creates a random forest model relying on the input file "voice.csv". After generating this model, R can apply the model to classify if the testing audio is from male or female.

```
model.forest <- randomForest(label ~ ., data = file)
```

(Figure. 8)

This "model.forest" is based on the features of audio in dataset. It can classify the testing audio in random forest format by its features using predict() function (Figure. 9) in random forest library.

```
prediction <- predict(model.forest, analyzedVoice)
```

(Figure. 9)

If the random forest classify the testing audio into male label, it means the system predict the speech is from male, and vice versa.

3.2.2 Discussion on Predictions

Following is a table on 20 samples tested in the system. The stats indicates that the accuracy of chosen classifier random forest is really high where in 20 samples there just 1 sample not match the reality.

The reason that the predicted result not match the reality might be the quality of audio. Audio with noise, background sound or multi-person voices might have a inaccurate prediction.

No.	predict		No.	predict	
1	female	yes	2	female	yes
3	female	yes	4	female	yes
5	female	yes	6	female	yes
7	female	yes	8	female	no
9	female	yes	10	female	yes
11	male	yes	12	male	yes
13	male	yes	14	male	yes
15	male	yes	16	male	yes
17	male	yes	18	male	yes
19	male	yes	20	male	yes

4. Interface

We made an interface prototype.Users can operate this system on the Website: https://lichtbalabala.wixsite.com/voicegender. On the index page(figure), users can press “find” button to choose a .wav file form their local documents and then press “->” for prediction.

(Figure. 10)

After calculating, an icon that represents the gender will pop-up on the result page(figure. 10). Users can click on the “thumbs up” or “thumbs down” button to agree or disagree the result. Doubleclick can cancel the operation. The feedback will be collected for calculating the accuracy of this system. Beside, there is a home button for returning to index page. (Figure. 11)

(figure. 11)

On each page there are facebook button and twitter button that allow users to share this website.

5. Timeline

Data	Workload
Oct. 9 - Oct.16	Project proposal/ideas
Oct.16-Oct.26	Consider users’ requirements
Oct.25-Nov.2	Search dataset and start on algorithm code.

Nov.2-Nov.9	Analyze the vocal features in dataset
Nov.9-Nov.16	Finish coding part Input voice recording and debug.
Nov.16-Nov.23	Design interface based on users' requirements
Nov.24-Nov.31	Combine system with interface

5.1 Work have done

- Read some research paper about the topic and found speech audio features and proper algorithms for predicting voice gender
- Explored voice datasets and familiarized it for future use
- Researched online for more information about the project:
    1. "Evaluation of gender classification model" in python notebook[16]
    2. "Voice gender recognition with many different models"[17]
    3. "Compared among classifier models on their accuracy"[18]
    4. "Gender-voice-recognition" in R[19]
- Made a prototype for this system: https://lichtbalabala.wixsite.com/voicegender
- Calculated and compared the accuracy of each classifier in R
- Choose a classifier that have highest accuracy
- Used the database "voice.csv" to generate a random forest model for classifying voice gender
- Debugging the code.
- Discussed work deviation and future work

6. Discussion of Work

6.1 Future goal

To approach a better functionality of our system, following aspects can be made in future work:

6.1.1 Interface

The prototype should be turned into a real frontend website, which can be combined with voice gender prediction system. Besides, the interface should be improved to offer a better user experience. A built-in music player for playing uploaded audio file is needed. Thus users can determine whether this is the correct audio file that they want to predict.

6.1.2 Method

The two methods: XGBoost, and Neural Network, are worth to consider in predicting voice gender. XGBoost, short for eXtreme Gradient Boosting package. The advantages of XGBoost is, it has several features to help with viewing how the learning progress internally and might maintain a quite high accuracy. By researching, Neural Network seems to maintain the highest, however, a worse time cost. Keras and tensorflow can be applied in neural network.[1][21]

6.1.3 Target

Songs, long speeches and multi-person audio are the next target of the system. For songs, rhythm and different instruments obstruct the prediction of gender of singer. The system might add some option in audio preprocessing on removing accompaniment.

6.2 Work Deviation

Workload turns out to be different from the original planned timeline due to the reason that the project workload is much heavier than what we think. Tons of time have spend on researching and deciding a best dataset and model to use. The development of interface and combination of website and system have not been finished due to the limit of time.

6.3 Work Distribution

Group Member	Workload
Leanne	Idea/ Proposal Intro/Dataset/Method/Timeline/Discussion of work/Conclusion
Zhenyu	Idea/Abstract/ Intro/Dataset/Method/Timeline/Discussion of work/Reference
Xinyue	Idea/Proposal/Dataset/Timeline/Interface/Discussion of work/Final edition

7. Conclusion

This paper has presented an approach to predict speeches that contain whether a male or female voice using Random Forest method in R. Eight methods have been investigated in the context of voice gender recognition, and their accuracy is evaluated based on the prediction of each model using the predict function. The result showed that Random Forest classifier outstands with an accuracy of 98%. By filtered the useful vocal features and then applying default R functions, we preprocessed the input audio data and predict the result using Random Forest model. The model can correctly predict the gender approximately 95% of the time on average.

We identify possible directions for future work as: First, combining our voice detection system with website as well as improve the interface looking. Second, trying more method to reach a higher accuracy. Third, detecting on more types of audio files such as songs and long speeches. Filtering noises to improve the quality of audios will also be considered in future work. To fully implements these expectations, many challenges and questions still need to be answered. With current advances in feature design and feature learning, however, we expect significant progress to be made in the near future.

References

[1] G. Tzanetakis, "Music Information Retrieval", ConneX, 2017. [Online]. Available: https://connex.csc.uvic.ca/access/content/group/1532dc0b-f665-443b-bfa8-995ce6fded43/mirBook_Jan12_2017.pdf. [Accessed: 12- Oct- 2017].

[2] M. Alhussein, Z. Ali, M. Imran and W. Abdul, "Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System", Hindawi, 2016. [Online]. Available: https://www.hindawi.com/journals/misy/2016/7805217/.

[3] M. Kumari and I. Ali, "An efficient algorithm for Gender Detection using voice samples", Ieeexplore.ieee.org, 2015. [Online]. Available: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7437912.

[4] H. Sheikh, "WHO IS SPEAKING? MALE OR FEMALE", https://studentnet.cs.manchester.ac.uk, 2013. [Online]. Available: https://studentnet.cs.manchester.ac.uk/resources/library/thesis_abstracts/MSc13/FullText/Sheikh-HassamUllah-fulltext.pdf.

[5] M. bot1, "Voice Gender Detection using GMMs : A Python Primer", Machine Learning in Action, 2017. [Online]. Available: https://appliedmachinelearning.wordpress.com/2017/06/14/voice-gender-detection-using-gmms-a-python-primer/.

[6] K. Becker, "Identifying the Gender of a Voice using Machine Learning", Primary Objects, 2017. [Online]. Available: http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/.

[7] "Human voice", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Human_voice.

[8] "Scientific pitch notation", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Scientific_pitch_notation.

[9] Lay, C. and James, N. (2017). Cite a Website - Cite This For Me. [online] Projapps.com. Available at: http://www.projapps.com/CS5240.doc.

[10] Bakshi, T. (2017). Gender detection using MATLAB. [online] Slideshare.net. Available at: https://www.slideshare.net/ronwinstanmay/gender-detection-using-matlab.

[11] Rakesh, K., Dutta, S. and Shama, K. (2017). GENDER RECOGNITION USING SPEECH PROCESSING TECHNIQUES IN LABVIEW. [online] citeseerx. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.9166&rep=rep1&type=pdf.

[12]Manual.audacityteam.org. (2017). Tutorial - Vocal Removal and Isolation - Audacity Development Manual. [online] Available at: http://manual.audacityteam.org/man/tutorial_vocal_removal_and_isolation.html.

[13]Wiki.audacityteam.org. (2017). Noise Reduction - Audacity Wiki. [online] Available at: http://wiki.audacityteam.org/wiki/Noise_Reduction.

[14] BEZOOIJEN, R. (2017). "Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women", [online]: http://journals.sagepub.com/doi/pdf/10.1177/002383099503800303.

[15] Liu, H. (2017). "Effect of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations", [online]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037774/.

[16]Sasaki, K. (2017). Evaluation of gender classification model | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/lewuathe/evaluation-of-gender-classification-model.

[17]Garziano, G. (2017). Voice gender recognition with caret package | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/giorgiogarziano/voice-gender-recognition-with-caret-package.

[18]Mesh, U. (2017). Simple Benchmarking of Classifiers - Accuracy | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/umeshnarayanappa/simple-benchmarking-of-classifiers-accuracy.

[19]https://github.com/primaryobjects/voice-gender/blob/master/sound.R

[20]Chen, J. (2017). Predict gender with voice and speech data – James Chen – Medium. [online] Medium. Available at: https://medium.com/@jameschen_78678/predict-gender-with-voice-and-speech-data-347f437fc4da.

[21] Palet, M. (2015). Voice gender identification using deep neural networks running on FPGA. https://upcommons.upc.edu/bitstream/handle/2117/86673/113166.pdf?sequence=1&isAllowed=y.