Vocal Gender Classification

Using Piecewise Gaussian Modeling

Instructor : Juan Pablo Bello

Github: https://github.com/jason2003120/2016_ACA_Final

Cheng Hsun Lee (chl468@nyu.edu)

Bryan Cheng (bc2197@nyu.edu)

***Abstract*—This project used PGM (Piecewise Gaussian Modeling) to extract the vocal feature from the music, and classified the vocal's gender. Decision Tree and Multi Layer Perceptron were used to classify the feature with WEKA machine learning tool box. The pure vocal sound were used first to develop the classifier. Vocal isolation from the music will be evaluate for the future work.**

Keywords—Piecewise Gaussian Modeling; Vocal Isolation; Decision Tree; Multi Layer Perceptron; WEKA; MedleyDB

I. Introduction

Record the gender of singer is an important part for music annotation. However, most of voice gender identification research only devote in speech sound, rarely in vocal sound. Therefore, we try to find a appropriate feature extraction method that can classify the gender for the singers.

The goal of this project is to design a system that can recognize the gender of the vocal in music. In order to analyze the vocal features from the music, the instrumental source should be removed, since it will interfere the results. However, to completely remove instrument and from a mixed track is considered incredibly difficult. Many research have been done for isolating the vocal and the instrument, but it is still not clear enough for gender classification. Therefore, pure vocal tracks have been used as our dataset to develop our classifier first. Vocal gender classification with instrumental source will only be discussed with simple tests.

The audio data samples used in this project were provided by MedleyDB, which have roughly 29 male vocal and 37 female vocals in the dataset. Vocal sounds were edited into a four continuous audio data tracks that generate 1200 seconds of training data and 600 seconds of testing data for each gender.

To extract the features, we used Piecewise Gaussian Modeling (PGM), which was mentioned in Voice-Base Gender Identification in Multimedia Application by Hadi Harb and Liming Chen. There are three main reasons we choose this method to develop our feature:

First, We want the feature to be linguistic independent, that is, we don't need the information which relate to words or phonemes. Some research use HMM-based (Hidden Markov Models) gender classifier which are not effective since they are phoneme-related. Our system should be able to recognize the gender not only for songs in English, but also other in language. Hence we use PGM that analyze the frequency distribution in certain time.

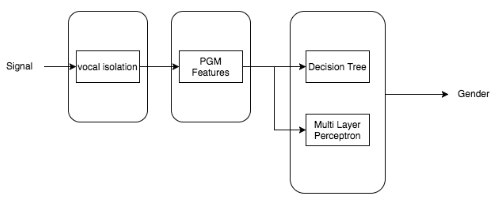
Second, we are not using pitch feature which is used in some speech recognition, because the pitch range of singing is larger then speaking. In addition, there will be more pitch overlapped region which will confused our classifier and lower its performance.

Last but not least, relating to human perception manner, it's hard to classify a short term audio segment signal. Instead, we could easier recognize the audio feature in long term time frame. In PGM, we've modeled several short term features into a long term window.

Finally, to classify the gender, we used both Muti Layer Perceptron and Decision Tree classifier to analyze the performance. Further data analysis and discussion and be mentioned in this report.

II. Method

A. System Overview



In our design, the input signal will have preprocessing before extraction the feature, and then sending the feature vector into the classifier, and then identify the gender finally.

B. *Feature Analysis with PGM (Piecewise Gaussian Modeling)*

The PGM (Piecewise Gaussian Modeling) is a method that use a long term structure to represent the features, called Integration Time Windows (ITW). In every ITW, it have a mean vector and a variance vector which are modeled from a set of short term spectral vectors.

To be more specifically, first, we compute Mel power spectrum by multiply power spectrum and mel filter bank, it called MFSC (Mel Frequency Spectral Coefficient), it's similar to MFCC but without DCT part. Where t is the time index for MFSC.

$$\overrightarrow{X_t}, t = 1, 2, 3, ...N * T$$

T refers to the number of spectral vector contained in a ITW. For example, in our design, we set MFSC parameter: window size = 512, hop_size = 256, ITW = 1 sec, fs = 44100 Hz. The sample rate of MFSC is fs / hop size = 172.26 Hz. Hence, we will get 172 of MFSC feature vectors in a second, so T = 172. N refers to the number of ITW windows. If we have 1000 sec samples with 1 sec ITW, the N should be 1000.

The PGM consist of modeling a set of T consecutive MFSC vectors by one Gaussian model. That is, N*T of MFSC will be modeled by N Gaussians:

$$\left\{\overrightarrow{X_1}, \overrightarrow{X_2}, ..., \overrightarrow{X_{NT}}\right\} \rightarrow \{M_1(\overrightarrow{u_1}, \overrightarrow{\sigma_1}), M_2(\overrightarrow{u_2}, \overrightarrow{\sigma_2}), ...M_N(\overrightarrow{u_N}, \overrightarrow{\sigma_N})\}$$

The the mean and the variance vectors will be computed by:

$$\overrightarrow{u_i} = 1/T \sum_{t=(i-1)N+1}^{iN} \overrightarrow{X_t}$$

$$\overrightarrow{\sigma_i} = 1/T \sum_{t=(i-1)N+1}^{iN} (\overrightarrow{X_t} - \overrightarrow{u_i}) \cdot (\overrightarrow{X_t} - \overrightarrow{u_i})^T$$

When calculate the covariance matrix, we only take diagonal index of the covariance matrix to generate the variance vector.

Finally, The PGM features are the concatenation of the mean and variance normalized by their respective maximum and minimum by

$$y_{norm} = (y - min(y))/(max(y - min(y))$$

Therefore, the PGM feature dimension will be (2 * Number of MFSC features) * N. It capture the dynamics of the speech and the distribution of the energy in each frequency channel contained in a time sliding long term window.

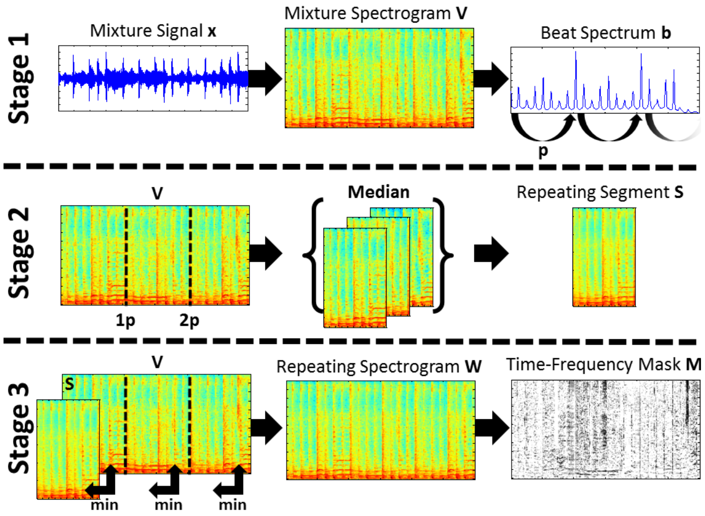
C. *Vocal Isolation*

Vocal isolation has always been an important but difficult issue. Three approaches were discussed in this paper as following.

The first solution uses phase to isolate vocal. It is assumed that under most circumstances vocals are in the middle and instruments are spread around them. Ideally, in a stereo audio signal, vocals which are in the middle will have the same amplitude in both left and right channels. By taking advantage of destructive interference, we inverted one track and subtracted it with the other one. This will eliminate the vocals in the signal. And with the instrumental track, the vocals can be extracted from the original signal.

However, in practical cases, vocals are not perfectly in the middle, and numerous other instruments such as bass are also panned in the middle. This will all defect the performance. Also, another drawback is that this method has to turn a stereo signal into a mono signal.

Another solution is center channel extractor. It does phase invert only to a specific range of frequencies. This will reduce the instruments even more, and keep more phase. Nevertheless, it will still damage the quality of the signal unavoidably.

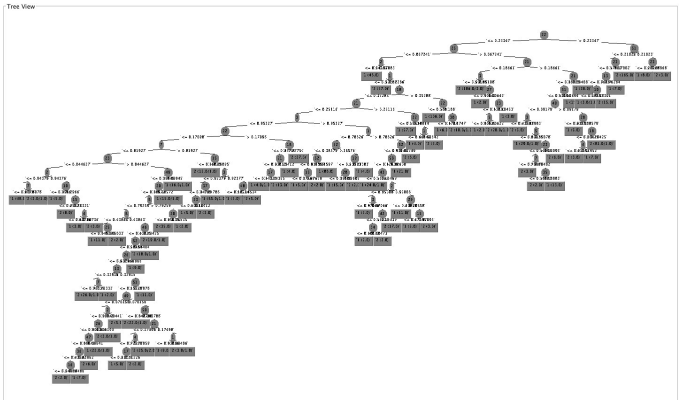


Due to the restrictions, Repeating Pattern Extraction Technique(REPET) was introduced to solve this problem. Repetition plays an important role in generating and perceiving structure. In this algorithm, we assume that all the instruments are in the repeating background and can be detected and removed. First, calculates the best spectrum and estimate a repeating period. Then separate the mixture spectrogram and calculate the repeating segment. Finally, calculate the repeating spectrogram and derive the time-frequency mask M. Furthermore, there is an extension of REPET. REPET-SIM extends REPET and can handle non-periodically repeating structures. It uses a similarity matrix to identify the repeating elements.

D. *Classifier*

Decision Tree: Decision trees are well known data mining methods. While each tree nodes is a question of a value of a parameter. And the paths represent different set of parameters. We used J48 in WEKA which implemented C4.5 algorithm introduced my Ross Quilan. The splitting rule used by C4.5 is Entropy and Gain Ratio.

Multi Layer Perceptron: Multi layer perception is an algorithm that imitates humans nervous system. Since it is more complex, the main disadvantage is that the training time could be relatively long. We use MLP because PGM features are not extremely rely on the correlation of the input features as the statistical classifiers do.



III. Discussion

A. *Using the dataset from MedleyDB*

In this experiment, 29 male and 37 female of pure vocal source are included with total 60 minutes long. Every second in the audio data are treated as a single instance, and both male and female have the same amount of data. We separate data into two parts, training and testing, 20 minutes and 10 minutes respectively. Therefore, we have 1200 instances as training and 600 instances for both gender. For convenience, we edited all source into four tracks, train and test for both gender. Most of the blank without vocal sound had been remove to improve the accuracy.

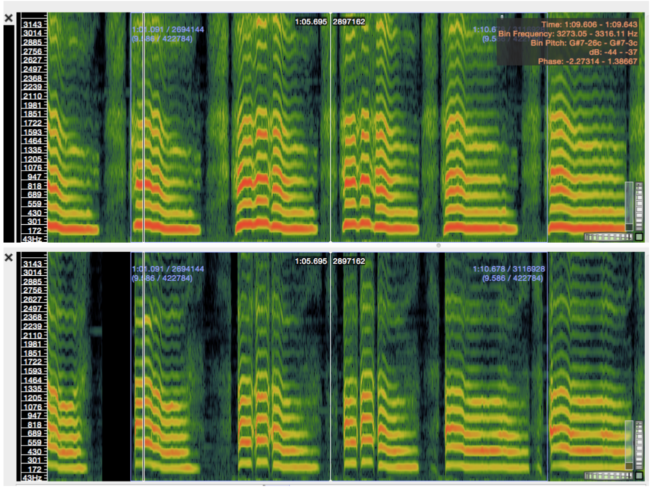
B. *Using the Weka GUI:*

Weka is an extremely convenient tool developed by Machine Learning Group at the University of Waikato. It is a collection of machine learning algorithms for data mining tasks. With a user-friendly GUI, datasets can be loaded and classified easily.

Weka has numerous features. In this experiment, we used mainly the preprocess training without filter, and classify for output result. Both classifier J48 and MultilayerPerceptron are used to classified the data, where J48 is under the root of trees and MultilayerPerceptron is under the root of functions. Weka can also shows information such as decision tree in graphic as following.

C. *Characteristics of Speech and Vocal:*

Although the sound speaking and singing are all generated by human, there are still few differences. Singing is more dynamic comparing to speaking. The frequency range of singing is larger and change more rapidly. Also, the magnitude of singing varied more than speaking. Timbre changes as well. Therefore, there is still a huge gap between analyzing speech and vocal.



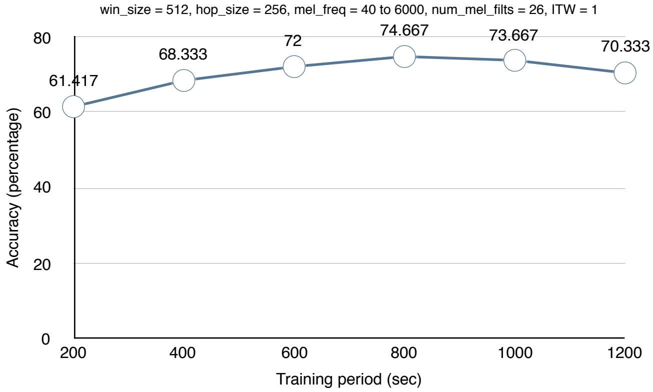
D. *Male Vocal vs Female Vocal*

It's the spectrogram with male and female vocal that sing in the same note (top is female vocal, bottom is male vocal). We can see that female's harmonic in higher frequency area are more intensive than male's. In human perception view, large weight in higher harmonic frequency could be sound more sharp, vice versa, in lower harmonic frequency, the sound will be more rough.

IV. THE RESULT

A. *Decision Tree*

Different duration of training data are set up to compare the accuracy.

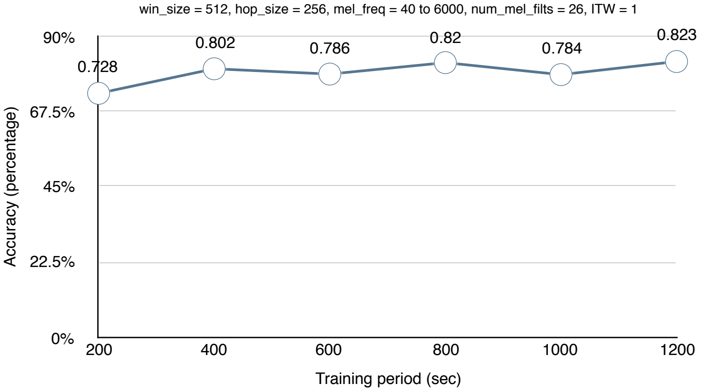


From the plot above, the result reach to maximum when we train with 800 seconds from our training data set with decision tree classifier.

```
=== Summary ===

Correctly Classified Instances        896          74.6667 %
Incorrectly Classified Instances      304          25.3333 %
Kappa statistic                        0.4933
Mean absolute error                    0.2598
Root mean squared error                0.4928
Relative absolute error               51.9564 %
Root relative squared error           98.5513 %
Total Number of Instances           1200

=== Detailed Accuracy By Class ===

        TP Rate  FP Rate  Precision  Recall  F-Measure   ROC  Class
         0.697    0.203    0.774     0.697    0.733     0.746   1
         0.797    0.303    0.724     0.797    0.759     0.746   2
Avg.     0.747    0.253    0.749     0.747    0.746     0.746

=== Confusion Matrix ===

  a   b   <-- classified as
 418 182  |    a = 1
 122 478  |    b = 2
```

Above is the summary that generated by Weka. Class 1 is for male, class 2 is for female.

B. *Multi Layer Perceptron*



Different duration of training data are set up to compare the accuracy.

From the plot above, the result reaches the maximum when we train with 1200 seconds from our training data set with multi layer perceptron.

```
=== Summary ===

Correctly Classified Instances        988          82.3333 %
Incorrectly Classified Instances      212          17.6667 %
Kappa statistic                        0.6467
Mean absolute error                    0.1809
Root mean squared error                0.3921
Relative absolute error               36.1873 %
Root relative squared error           78.4138 %
Total Number of Instances           1200

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
           0.743    0.097    0.885     0.743    0.808     0.906     1
           0.903    0.257    0.779     0.903    0.836     0.906     2
Avg.       0.823    0.177    0.832     0.823    0.822     0.906

=== Confusion Matrix ===

  a   b   <-- classified as
 446 154 |   a = 1
  58 542 |   b = 2
```

In our experiment, MLP preformed better accuracy with the same data set and parameters.

C. *Vocal Instance Accuracy*

Male

male vocal	sec	#instance	err	accuracy (%)
1	0-82	1-82	-21	74.3
2	82-154	82-154	-13	82.1
3	154-219	154-219	-21	68.1
4	220-246	220-246	-6	77.8
5	247-280	247-280	-12	64.7
6	280-345	280-345	-34	46.8
7	346-442	346-442	-12	87.6
8	443-491	443-491	-26	45.8
9	491-599	491-599	-37	65.4

Female

female vocal	time(sec)	#instance	err	accuracy (%)
1	0-24	1-24	-2	91.6
2	25-59	25-59	-12	62.7
3	60-83	60-83	-1	95.8
4	84-126	84-126	-11	74.41
5	127-158	127-158	-4	87.5
6	159-206	159-206	0	100
7	207-263	207-263	-7	87.7
8	264-301	264-301	-13	65.7
9	302-329	302-329	-2	96.4
10	330-366	330-366	-11	70.2
11	367-438	367-438	-16	77.7
12	439-600	439-600	-42	74.0

We've separated the result of male and female, and print out the predictions for the test set. In our testing data, we have 12 female vocal and 9 male vocal. We found there are two main reasons that led to error prediction. First, there were some errors happened at the time when the male vocal sing in a high pitch with light strength which sounded vigorous and sometimes sounded more like a female singer. For human perception, sometimes it's also hard to tell the difference by gender when the male vocal sings like female or the female vocal sings like male. In our data set, there are some male vocal singing like female that led to the wrong prediction.

The other reason is that due to the breathing sound between the singing interval, the breathing sound is hard to tell the gender that effect to the result for both male and female identification. Moreover, breathing is very important and happened frequently when we sing a song, so it will lose some accuracy due to this inevitable factor.

Above of all, if remove these two kind of miss detecting situation, our performance is quite good with this feature extraction method.

IV. CONCLUSION AND FUTURE WORK

The result of the accuracy is about 74.6% (79.6% for female and 69.6% for male) with decision tree classifier and 82% (86.8% for female and 77.1% for male). We've found that most of the paper that classify the speech gender can reach over 90% accuracy. Therefore, our result tells us that to classify the vocal sound is more difficult than speech sound, maybe we need more efficient method to obtain the feature. However, we have discovered the factors that decrease the accuracy. Hence, to increase the accuracy in the future is to find the solution to remove those factor.

First, we may need to add more training data, it may help to the classifier predict more clearly when we face to high pitch male sound or low pitch female sound. In addition, to reduce the impact of singing interval, we may do some preprocessing such as setting up a volume threshold to eliminate breathing sound before we analyze the signal.

For instrument removal part. We are looking forward in the future to find a better solution to isolate the vocal and integrate it to our system. We hope that our system can be tested with realtime conditions, and can be fed with any kind of sources and still maintain high performance.

In conclusion, this project implements the gender classifier, which can classify music from the gender of the vocals automatically. We use speech recognition method to develop our vocal recognition result, which we think is reasonable. Although speaking and singing have some different phonate way, there are more pitch overlap space between male and female vocal, the result can still be acceptable, after all.

IV. REFERENCE

[1] H. Harb, L. Chen, J. Auloge, Speech/ Music/ Silence and Gender Detection Algorithm

[2] H. Harb, L. Chen, Gender Identification Using a General Audio Classifier

[3] H. Harb, L. Chen, Voice-Base Gender Identification in Multimedia Application

[4] Z. Rafii and B. Pardo, Repeating pattern extraction technique (REPET): A simple method for music/voice separation, Audio, Speech, and Language Processing, IEEE Transactions on 21.1 (2013), 73-84.

[5] Z. Rafii and B. Pardo, ONLINE REPET-SIM FOR REAL-TIME SPEECH ENHANCEMENT

[6] Z. Rafii and B. Pardo, A Simple Music-Voice Separation Method based on the Extraction of the Repeating Musical Structure - ICASSP 2011

[7] http://www.zafarrafii.com/repet.html