# Phonation Detection System - Final Report

You Yuan, Anwen Xu[1] and Junwei Yang[2]

[1] *Electrical Engineering*
[2] *Civil Environmental Engineering*

## I. INTRODUCTION

The analysis of human singing voice brings meaningful insights of people's vocal resonance and assist people to do self vocal training. The resonance condition can be detected through various spectrum techniques, e.g. DFT, MFCCs, which extract dominant frequency, harmonics and other spectral components in the voice that are not easily detectable in time domain waveforms.

We provide a reliable and robust phonation detection system, which helps detecting people's singing resonance position based on machine learning models (SVM, neural networks, decision tree) as well as spectrum analysis techniques (DFT, MFCCs, etc). Models are trained with a labeled supervised learning dataset recorded from both vocalization professionals and non-professionals. And a phonation detection and scoring system is also implemented.

## II. DATA COLLECTION

To ensure good quality of the supervised learning datasets, all voice clips are collected from Stanford music professors teaching vocalization and outstanding students recommended by the professors. The samples are recorded with the same equipment in various but relatively quiet environments in order to make our models' prediction more accurate. Based on vocalization theories, the resonance positions can vary horizontally (from head to chest resonance) and vertically (from backward to forward resonance), thus the supervised learning dataset covers samples of these two dimensions and are labeled into 9 categories. The entire supervised learning dataset contains 700 labeled vowel clips, including male and female voices and over different major scales.

| Labels | Backward | Balanced | Forward |
|---|---|---|---|
| Head | (-1, 1) | (0, 1) | (1, 1) |
| Middle | (-1, 0) | (0, 0) | (1, 0) |
| Chest | (-1, -1) | (0, -1) | (1, -1) |

Table I. Classification of resonance positions

## III. FEATURE EXTRACTION

Feature extraction includes three steps: *audio data preprocessing*, *spectrum feature analysis*, and *training matrix preparation.*

### A. Audio data preprocessing

In the learning dataset, each recorded voice clip is tailored into valid pieces each with only one note, and normalized to have the same magnitude scale. For testing continuous singing clips, the recorded voice clip will be cut into small overlapping clips for the test system which will be explained in Part V.

### B. Spectrum Feature Analysis

Once each voice clip is tailored and normalized, the spectral features are extracted for each voice clip. After careful evaluation of audio energy, perceptual, temporal, spectral features, the most statistically significant DFT features and MFCCs are selected for further model training.

#### *1. DFT Diagram*

Different vocalization methods alter the position of resonance inside the body, and produce distinctly different frequency bandwidths and magnitudes (as shown in Figure 1).

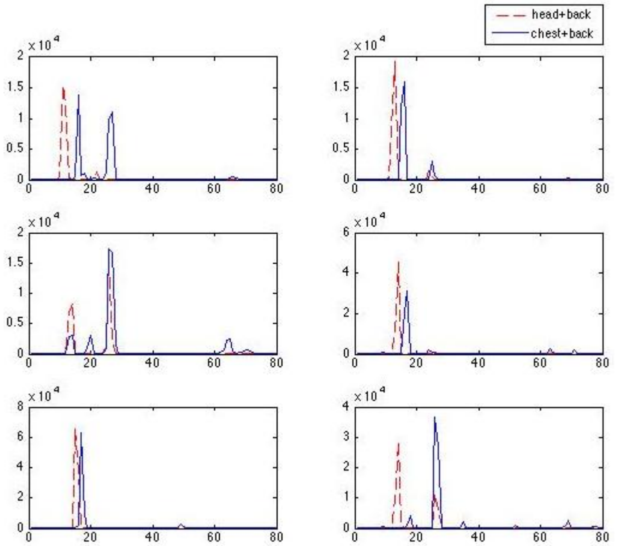


Figure 1. Head Voice's DFT vs Chest Voice's DFT

To exploit such types of differences, we apply Hamming window of size 1024 to sample each note clip with

50% partial overlap. After discrete fourier transform, we choose the median values from above spectrum. In order to reduce the variance in dataset, we averaged the magnitudes of 4 neighboring frequency samples, which reduces the resolution.

In Figure 1, the commonness in the same group and the difference between different group is conspicuous. Head voice often has a powerful and centralized peak, while the energy of chest voice is distributed in a much flatter way. Although we implement a simple SVM model only based on DFT features, its accuracy of head/chest resonance prediction is only 64%. Therefore, we investigated other features to improve the overall performance. Based on P-value evaluation results, we select the frequencies and normalized magnitudes values of five toppest peaks from DFT diagrams as dominant features.

#### *2. MFCCs*

MFCCs are used extensively in speech and speaker recognition. Essentially, they represent the Discrete Cosine Transform of the log spectrum of a signal analyzed on an auditory frequency scale (the Mel scale). The process creates a 13-dimensional vector that summarizes the signal's spectrum. We included MFCCs to represent the differences in the shape of the spectrum for different signals.

### C. Training Matrix Preparation

As mentioned above, 22 features (13 MFCCs + 9 DFTs) are extracted for each voice clip. The feature matrix is constructed with each row containing the 22 features of one voice clip, which is of size (700, 22) in total. The target matrix is constructed with vertical and horizontal resonance position labels for each voice clip, which is of size (700, 2). Additionally, the feature and target matrices of male and female are separated for independant model training, given the fact that male and female demonstrate completely different vocalization characteristics. Table II below shows the actual feature and target matrices utilized by model training.

| | Horizontal Target | Vertical Target |
|---|---|---|
| Male | Feature Matrix(350,22)<br>Target Value(350,1) | Feature Matrix(350,22)<br>Target Value(350,1) |
| Female | Feature Matrix(350,22)<br>Target Value(350,1) | Feature Matrix(350, 22)<br>Target Value(350, 1) |

Table II. Four feature matrices and target matrices size

## IV. MACHINE LEARNING MODEL

For each of feature matrix and target matrix with different gender and resonance dimension, the following machine learning models are trained and used to predict the singer's phonation condition.

1. Logistic Regression Model
2. Support Vector Machine
3. Neural Network
4. Decision Tree

#### *1. Logistic Regression Model*

Logistic regression is implemented using standard gradient descent. This model is based on the fact we learn from DFT diagrams that generally head voice contains purer higher frequency components while chest voice contains wider range of lower frequency and the assumption of linear relations between the logit of the explanatory feature inputs and the classification results. The performance of logistic regression acts as our performance baseline for other models evaluations.

#### *2. SVM*

SVMs with different kernels are implemented using Libsvm package. For each of the four groups, a separate model selection is performed separately using different SVM parameters.

Parameter Selection - We adopt C-SVC (standard multi-class classification) for all groups of dataset. In order to discover the locally optimum value of each parameter, an automatic SVM options selected program is created and applied. For each model, common kernel types, number 1 to 10 degrees and 10 logarithmically spaced values of $\gamma$ were evaluated based on cross validation results. Models with maximum CV values are chosen as final SVM models.

| | Male Vertical | Female Vertical |
|---|---|---|
| Kernal Type | polynomial | polynomial |
| Degree | 2 | 2 |
| $\gamma$ | 1e-04 | 1e-03 |
| Coeff0 | 6 | 1 |
| Cross Validation | 80.60% | 81.37% |
| | Male Horizontal | Female Horizontal |
| Kernal Type | polynomial | polynomial |
| Degree | 4 | 3 |
| $\gamma$ | 1e-06 | 1e-03 |
| Coeff0 | 15 | 17 |
| Cross Validation | 78.36% | 64.71% |

Table III. SVM Parameters

Performance Analysis - As shown in Table III, SVM models for vertical resonance groups perform far better

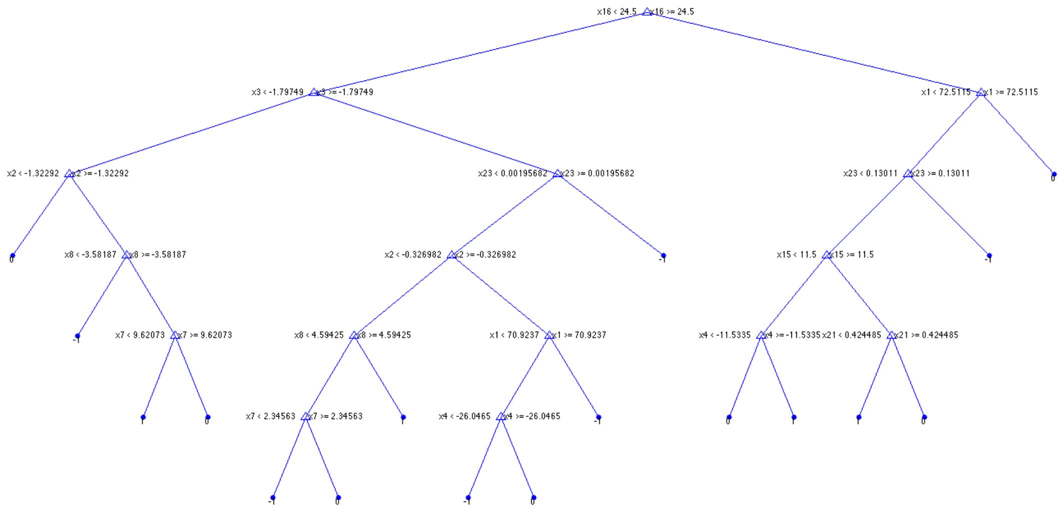


Figure 2. Decision Tree: Female Vertical Resonance Position

than horizontal resonance groups. Intuitively it is much easier for human ears to detect head voice from chest voice than to distinguish forward resonance from backward resonance, because vertical resonance could partially be differentiated by pitch while the difference in horizontal resonance is somehow subtle.

### *3. Neural Network*

Given the complicated nature of vocal resonance, a (22,23,23,1) Neural Network model is also considered and implemented, which includes one input layer, two hidden layers with 23 neurons and transfer function ”tansig” and one output layer with transfer function ”purelin” from MATLAB Tool Box. The number of neurons are selected based on iterative tests with different parameter combinations. In total four Neural Network models are trained separately for different combinations of gender of the singer and resonance classification directions: male-horizontal, male-vertical, female-horizontal, female-vertical.

### *4. Decision Tree*

Decision tree is trained via the classification tree functions inside Matlab machine learning package. Each combination of the feature matrices and target matrices (male-horizontal, male-vertical, female-horizontal, female-vertical) is trained separately to generate decision tree models. Through analysis of features used to determine the decision tree, it is shown that the first 13 features, the MFCCs, and the final 4 features, the magnitude of the peak DFTs, have more importance in the decision trees. However, the features, which represent the locations of different DFTs peak, are not used in the decision tree. One of the tree model is shown in Figure 2.

### *5. Model Selection/Evaluation*

After running all the models above, based on the performance accuracy (as shown in figure 3 and 4), the best models are selected for prediction.

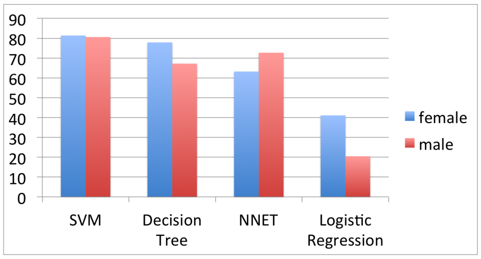


Figure 3. Vertical Resonance: Model Accuracy Comparison

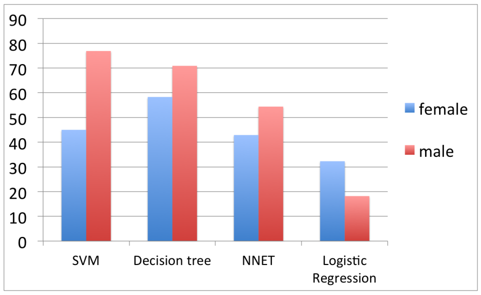


Figure 4. Horizontal Resonance: Model Accuracy Comparison

As mentioned, the logistic regression model acts as the baseline for our model evaluation. The SVM model works best in determining the vertical resonance position, with approximately 81% accuracy for female and 80% for males, which is a significant improvement from the baseline (40% for female and 20% for male). But in horizontal direction, decision tree works best of the 3 models with accuracy of 59% for females and 71% for males, which are relatively higher compared to other models. This phenomenon can also be explained by our feature choices: MFCCs and DFT features are more related to the vertical resonance position of singing than to horizontal ones. Based on the best models chosen above, the misclassification rate for each of the class labels are also calculated to evaluate and analyze the model performance (as shown in figure 5, 6, 7 and 8).

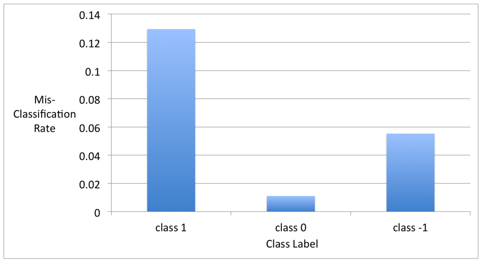


Figure 5. Misclassification Rate of SVM Model for Female Vertical

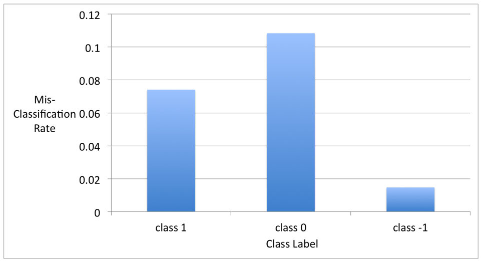


Figure 6. Misclassification Rate of SVM Model for Male Vertical

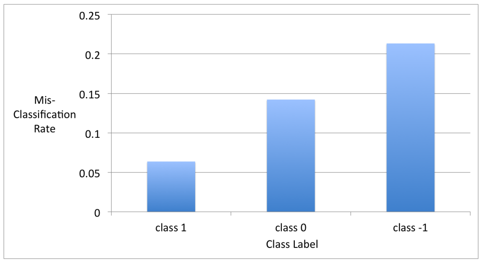


Figure 7. Misclassification Rate of Decision Tree Model for Female Horizontal

From the figures, it is illustrated that for SVM model to detect the female vertical resonance position, the class label 0 (middle voice) prediction is very accurate with only 1% error rate. However, the class label 1 (head voice) has an error rate of 12.5%. For the SVM model to detect the male's vertical resonance position, the prediction of class label 1 becomes very accurate (1.5% error rate) while class label 0 has an error rate of 11.5%. For the decision tree model, which is to detect the horizontal resonance position of male and females, class label 1 (forward voice) is minimally classified with 6% error rate to predict female's horizontal position and 7.5% to predict the male's horizontal resonance position.

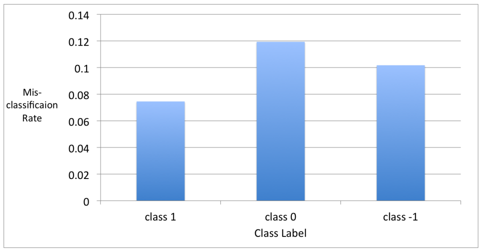


Figure 8. Misclassification Rate of Decision Tree Model for Male Horizontal

## V. PHONATION DETECTION SYSTEM

A detailed flow graph of the phonation detection system and how user interface interacts with the backend is displayed in the Figure 9 below.

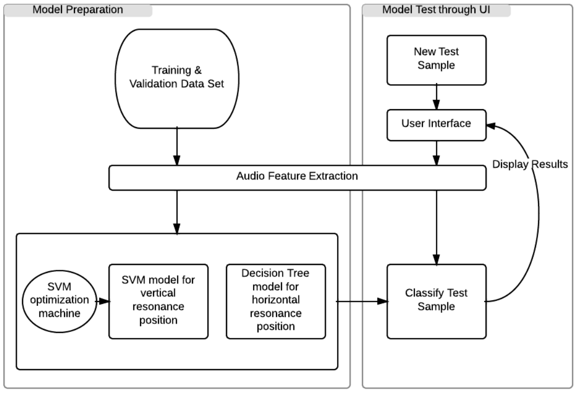
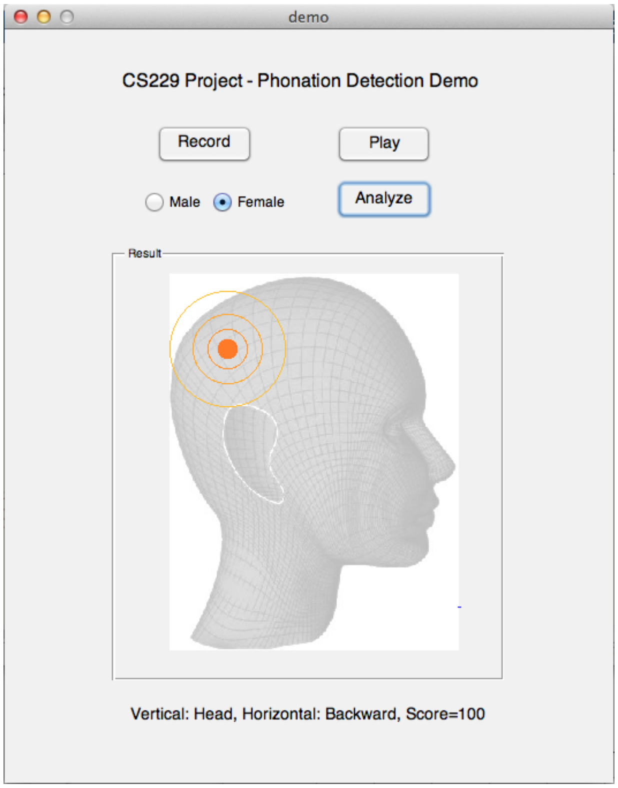


Figure 9. Vertical Phonation Detection System

As discussed in part IV 5, SVM and Decision Tree models are trained for vertical and horizontal classification uses, resulting four models of different resonance dimension and gender combinations. When a test sample is recorded through UI, the voice clip is tailored to unit test clips. Such test unit clips will be classified based on SVM and Decision Tree models with the corresponding gender, and the prediction with highest confidence will be selected as Vocalization Detection System output. Also, the Phonation Detection System will perform



CS229 Project - Phonation Detection Demo

○ Male ● Female

Vertical: Head, Horizontal: Backward, Score=100

Figure 10. User Interface

a weighted Euclidean distance calculation between testing voice clip and the most similar classified centroid, and the distance will be projected to [0,100] range as the performance score.

## VI. FUTURE WORK

The phonation detection system works well with most testers on the poster day, but we would like to extend the work continuously to make the system more comprehensive and accurate.

1. In this Phonation Detection System, we didn't distinguish vowel as all testers are required to sing with vowel "Ah" only. However lots of relevant research have found out, different vowels represent distinguishable harmonic content, this may raise challenges to classify resonance position and distinguish vowel with spectrum analysis at the same time.

2. There is still lots of other meaningful phonation classification that will be helpful in vocal-training. For example, classical vocal professionals distinguish themselves from contemporary singers by lifting their soft palate to have a more "operatic" sound. Such differences might be extracted through both spectrum analysis and sound energy analysis.

## ACKNOWLEDGMENTS

We would like to extend our gratitude to all those people who helped and supported us in the project. We would like to thank Professor Claire Giovannetti, Professor Wendy Hillhouse, Professor Gregory A. Wait, Yuanyuan Wen and Chenjie Luo for kindly sharing their music knowledge and providing their voice clips.

[1] Zhu Li and Yao Wang Audio feature Extraction and Analysis for Scene Segmentation and Classification 2001; Volumn 25, Issue 14:48–50.

[2] Ingo Mierswa and Katharina Audio Feature Extraction for Classifying Audio Data 2004

[3] Bryan Huh and Arun Miduthuri Vocal-Based Musical Genre Classification 2005

[4] Pat Taweewat Detection of a Specific Musical Instrument Note Playing in Polyphonic Mixtures by Extreme Learning Machine and Partical Swarm Optimization 2012; Volumn 2, Issue 5

[5] Greg Sell, Gautham J.Mysore and Song Hui Chon Musical Instrument Detection 2006

[6] Francesco Camastra and Alessandro Vinciarelli Machine Learning for Audio, Image and Video Analysis 2007

[7] Polina Proutskova, Christophe Rhodes, Tim Crawford and Geraint Wiggins Breathy, Resonant, Pressed - Automatic Detection of Phonation Mode from Audio Recordings of Singing 2013; Volumn 42, Issue 2:171-186

[8] Stefan Steidl Vocal Emotion Recognition: State-of-the-Art in Classification of Real-Life Emotions 2010

[9] Nathalie Henrich, John Smith and Joe Wolfe Vocal Tract Resonances in Singing: Strategies Used by Sopranos, Altos, Tenors and Baritones 2010

[10] Giovanni De Poli and Paolo Prandoni Sonological Models for Timbre Characterization 2010