Gender Recognition by Voice

#Define
Haozhe Gu, Mengda Xu, Esther Zhou, Yishen Huang, Erin Mcginnis,
Mat Magno, Aaron Allen, Jaden Zhang, Zhiyi Xu, Jason Vo,
Daniel Burke, Allan Soria, Han Chen, Garland Li, Jeffrey Wang
{hzgu, mdxu, xiezhou, ysshuang,
emmcginnis, mlmagno, aacallen, jidzhang,
yzxu, javo, djburke, aesoria, hhchen, garli, fejwang}@ucdavis.edu

November 29, 2016

1 Abstract

In this paper, we address the issue of determining gender from speech characteristics. We identified the ten best subsets of features for gender prediction by comparing the logistic classification errors for every subset of features. We then found the best five features by choosing the most common features in our top ten subsets. Then, we trained and compared different algorithms - neural networks, support vector machines, logistic regression classifiers, and decision trees - on a data set containing male and female voices using the features that best distinguish between male and female speakers. The resulting models achieved accuracies of over 90% on the testing sets; SVM had the best results, with 98% on the testing set. Thus, our SVM model will allow highly accurate gender classifications of speakers, which if used in speech recognition software, will allow that software to better identify and respond to speakers.

2 Introduction

Speech recognition has many applications such as automatic speaker verification [5], speech controlled user interfaces [16], and speech-to-text language conversion[4]. Studies have shown that prior knowledge of the speaker, such as knowledge of language or gender, can significantly improve the quality of speech recognition results [1][14]. Vocal discrepancies between male and female voices in particular, which include features like pitch and vowel pronunciation [7], can degrade the quality of translations and have been shown to increase error in speaker verification [11]. By identifying the speaker's gender *a priori*, a speech recognition program can adapt its model to account for common differences in vocal qualities between male and female speakers.

Extensive work has been done on techniques to automatically identify gender based on voice. Harb et al. achieved 93 to 98.5 percent accuracy, depending on audio sample length, in gender recognition using an artificial neural network (2005). Other techniques like support vector machines (SVMs), including four SVM formulations for dual age and gender identification[10] and mel frequency cepstral coefficients (MFCC) based SVM [8], have attained similar accuracies. These techniques often

focus on optimal feature selection [6] and comparisons of different models of similar base algorithms [10]. However, a survey of available techniques is necessary to find the most effective algorithm for gender identification by voice. By identifying a simple yet efficacious algorithm for this problem, future research can focus on optimizing and identifying the features of the chosen algorithm to ultimately produce excellent results.

Our paper's contribution will be to apply neural networks, SVMs, logistic regression, and decision trees to the gender identification problem. We will show that SVM and logistic regression performed the best at classifying voices by gender, while ANN and the decision tree were less successful. The discussion that follows will list our theories as to why some algorithms out-performed others, and how this knowledge can be applied to further problems.

The structure of our paper is as follows. First, the methods section will explain how features were selected and how each algorithm was implemented. Next, the results section will summarize the findings from each method. Finally, a discussion section will provide an interpretation for results and discuss future work.

3 Methods

3.1 Feature Selection

In order to rank the importance of each feature in determining the speaker's gender, we computed the logistic classification error of each possible subset of features out of the 20 original features. Sets with the lowest classification error were assumed to contain the most influential features for successful classification.

The features were further refined by finding the ten feature sets with the lowest classification error, then choosing the five most common features within those ten sets.

3.2 Logistic Regression

The logistic regression problem is as follows. Let $g(x^{(i)}; w) = sigm(w^T x) = \frac{1}{1+e^{-w^T x}}$ and $h^{(i)}$ be the i^{th} row of matrix h. Given a data set x, a corresponding label vector y, and a randomly initialized weight vector w, we want to maximize the likelihood of our labels y given x and w, which can be written

$$p(y^{(i)}|x^{(i)}; w) = g(x^{(i)}; w)^{y^{(i)}}(1 - g(x^{(i)}; w))^{1-y^{(i)}}. \tag{1}$$

There is no closed form solution to the above problem. Thus we use the following update rule to iterative move w towards optimal weight set w^*;

$$w_j := w_j + \alpha(y^{(i)} - g(x^{(i)}; w)x_j^{(i)}. \tag{2}$$

We used **R**'s **glm** function carry out the above algorithm.

3.3 Classification Decision Tree

We created a classification decision tree using **R**'s rpart library. The rpart library's default measure of impurity is the gini index is $p \cdot (1 - p)$, where p is the fraction of positive samples at a leaf. The

rpart library includes built in cross-validation, which generates a particular error for each node of the tree. The cost parameter used to prune the tree was chosen as the cost parameter of the node with the lowest cross-validation error. The tree was pruned to minimize the cost (complexity) of the tree while maximizing the purity of each leaf. A full description of the rpart library is available in the official CRAN documentation of rpart.

3.4 Support Vector Machine

We used Matlab's Statistics and Machine Learning Toolbox to create, train, and tune our SVM with the top five features found using our feature selection technique in 3.1. We used a linear kernel function to train the SVM on our data. We tuned the SVM parameters using Matlab's built in functions, and found that the optimal maximum penalty imposed on margin-violating observations (the box constraints) for our data is 0.00001.

3.5 Artificial Neural Network (ANN)

We used a feed-forward neural network with backpropogration. The ANN was tuned to use a learning rate $\alpha = 0.07$, one hidden layer, and four nodes within the hidden layer.
Let g be the sigmoid function defined in 3.2; let w be a set of initially random weights; let $a^{(l)}$ mean all nodes in layer 1; and let $a_i^{(l)}$ mean the i^{th} node in layer l. During the feed-forward portion of the training, each activation node $a_i^{(l)}$ was updated with the rule

$$a_i^{(l)} = g(w^{(l-1)^T} a^{(l-1)}). \tag{3}$$

The update rule for backpropagation is

$$w_{kj}^{(l-1)} := w_{kj}^{(l-1)} - \alpha \frac{\partial RSS}{\partial w_{kj}^{(l-1)}} \tag{4}$$

where $\frac{\partial RSS}{\partial w_{kj}^{(l-1)}}$ is equal to $-\delta_k^{(l)} \cdot a_j^{(l-1)}$ and

$$\delta_j^{(l-1)} = \Sigma_k \delta^{(l)} \cdot w_{kj}^{(l-1)} \cdot (1 - a_j^{(l-1)}) \cdot a_j^{(l-1)} \tag{5}$$

for nodes not in final layer l. If a node is in the final layer,

$$\delta_k^{(l)} = -(y_k - a_k^{(l)})(1 - a_k^{(l)})(a_k^{(l)}). \tag{6}$$

We trained our ANN using increasing numbers of epochs, from one to 5000, to find the optimal number of epochs for the data set. Feed-forward and backpropagation steps were executed for each individual sample.

3.6 Outlier Analysis

We will classify outliers in the logistic regression model using Pearson Residuals. The calculation of residuals takes the difference between observed and fitted values and divides by the estimated standard deviation of the observed value; in other words, the Pearson Residual for the i^{th} observation is

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}} \tag{7}$$

where $\hat{\mu}_i$ is the fitted value and the denominator follows from the fact that $\text{var}(y_i) = n_i\pi_i(1-\pi_i)$. The Pearson residuals are approximately normally distributed over all data. Any residual with absolute value exceeding 1.96 under a 95% confidence level will be classified as an outlier.

3.7 Cross Validation

For each of our methods, we use 10-Fold cross validation to evaluate generalization error and robustness of each algorithm. The method splits the data into k partitions.

$$data = \{d_1, d_2, d_3, .., d_k\}$$

Partition i is chosen from the k partitions for testing, and the remaining $k-1$ partitions are used to train the model. This procedure is repeated for each of the k partitions.

$$\text{Set used for testing: } \{d_i\}$$
$$\text{Set used for training: } \{data - d_i\}$$

At each iteration, a new prediction is computed based on the testing set and trained model. For the new prediction, we compute the accuracy based on the proportion of correct classifications. The generalization error is the average of our k prediction errors.

3.8 Data Acquisition and Preprocessing

The raw data was derived from the Harvard-Haskins Database of Regularly-Timed Speech[1], Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University[2] ,VoxForge Speech Corpus[3] and the Festvox CMUARCTIC Speech Database at Carnegie Mellon University[4]. Each .wav file was processed by **R**'s WarrbleR library to produce 20 acoustic properties including mean fundamental frequency (kHz), interquantile range (kHz) and frequency centroid. WarrbleR's audio decomposition process allowed us to filter out frequencies outside human vocal range (0 - 280 Hz) to mitigate the effects of noise in the classification process. There are 3168 records in total; all include labels of "male" or "female".

4 Results

4.1 Feature Selection

We used the top five features identified by our feature selection algorithm in 3.1 to train each algorithm and predict the gender of the speakers. The resulting accuracy of our algorithms improved significantly over the classification accuracy of algorithms using all 20 features.

The graph below depicts the frequency of features within the top ten feature sets. We identified mean fundamental frequency, minimum fundamental frequency, spectral flatness, modulation index and interquantile range as our top five features.

[1] http://www.nsi.edu/ ani/download.html
[2] http://www-mmsp.ece.mcgill.ca/Documents../Data/index.html
[3] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/8kHz_16bit/
[4] http://festvox.org/cmu_arctic/

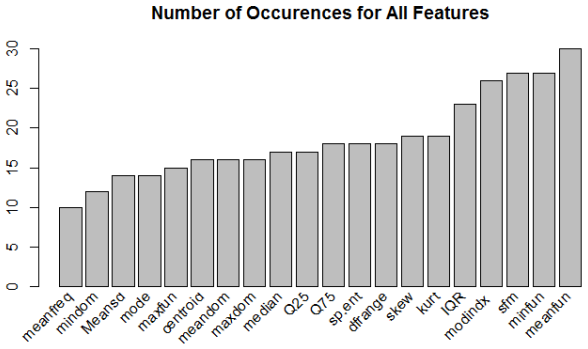

Figure 1: Frequency of audio features within the top ten subsets of all features

4.2 Logistic Regression

The generalization accuracy for 10-fold cross validation with the full 21 features was 97% with an average AUC of about 0.99. We obtained classification accuracy of greater than 98% using logistic regression with the top five identified features. The following figure is the ROC curve for logistic regression.

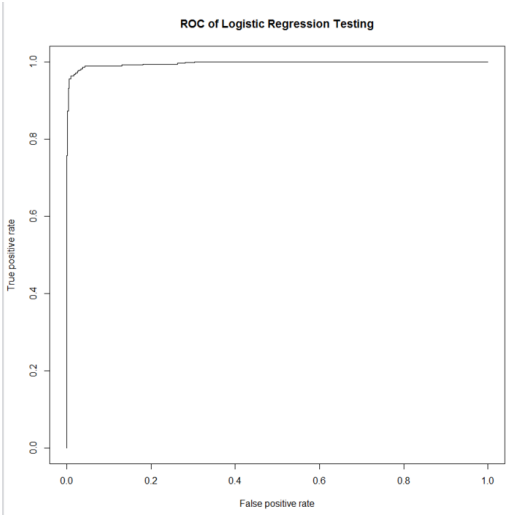

Figure 2: The ROC curve for logistic regression

4.3 Classification Decision Tree

The figure below contains the decision tree obtained from the our data.

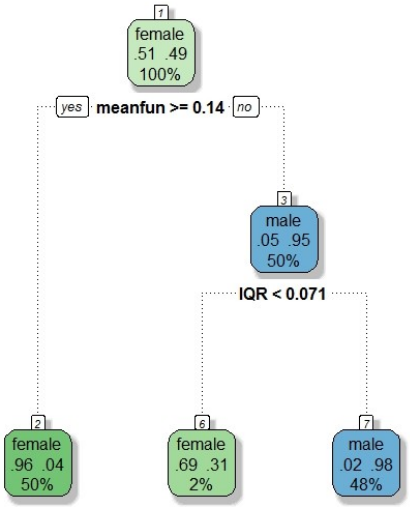

Figure 3: Graph of Classification Tree

In each node, the first line represents the predicted class (gender), the second line gives conditional probability for each class of gender at node (female on the left and male on the right), and the last line represents percent of the entire population observed in the sample.

The topmost node of the tree plot splits the data depending on whether mean fundamental frequency is greater or equal to 140 Hz, resulting in about 50% of the population in each child node. The lower left leaf is categorized as female and has 96% accuracy.

The node near the middle of the graph divides its data by deciding whether or not a sample has interquantile range smaller than 71 Hz. The resulting split correctly predicts the value of 98% of samples in its right child node, which contains 48% of the original population, and 69% of samples in its left child node, which contains 2% of the original population.

4.4 Support Vector Machine

The generalization accuracy for 10-fold validation for our SVM was greater than 97%, and the average area under the curve was about 99%.

4.5 Artificial Neural Network

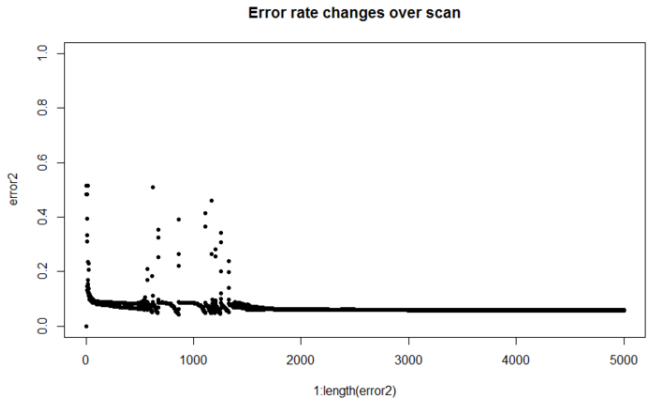

Figure 4: ANN Accurancy vs Epochs

In the above figure, we see that the ANN's classification error initially drops with an increased number of epochs. After about 200 epochs, the error remains relatively constant except for some outliers from epoch 700 to around 2000.

Excluding the outliers and initial drop in error, the classification error remains at about 10%, which means our ANN performs with 90% accuracy.

4.6 Outlier Analysis

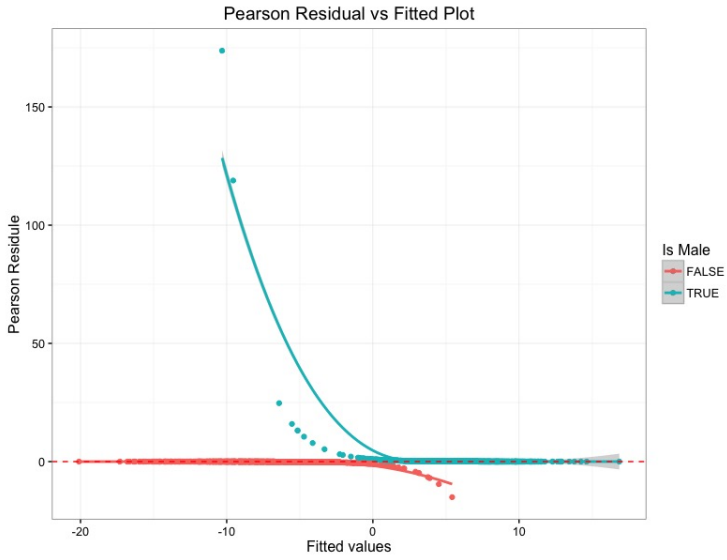

Figure 5: Residual Plot

The above figure depicts the Pearson residual versus the fitted value plot for our logistic regression model.

We predict whether a speaker is male or not; a probability of 1 corresponds to absolute certainty that the speaker is male, while that of 0 means we're certain the speaker is female. Thus, if the difference of expected and observed (the Pearson residual) is positive, the speaker was actually female, and if that difference is negative, the speaker was male. If the difference is close to zero, our prediction was very close to the actual value.

Based on our residual calculations, there exist 24 points with absolute values greater than 1.96 for both genders out of the training set of 2218 observation. Thus, these 24 points will be classified as outliers.

5 Discussion

Our top five predictors are consistent with those identified in other studies, and overall, SVM was the most robust and accurate of our methods. Previous studies have also distinguished mean fundamental frequency[13], minimum fundamental frequency[13], and spectral flatness[12] as excellent markers for each gender, and attained similar results by removing less important features. Furthermore, we can extrapolate from the average weights over all logistic regression runs that of the top five features, mean fundamental frequency (f0) is best marker for gender.

The importance of f0 in distinguishing between male and female voices is also supported by our decision tree; the high purity of the leaf containing voices with f0 less than or equal to 0.014KHz indicates that the majority of our sample population can be predicted using f0 alone, a conclusion supported by previous studies [9]. In explaining the few misclassified voices, we note that our outlier analysis shows that some voices within our data are very different from the mean voice. We hypothesize that these voices have markedly different values for f0, which would be sufficient to produce misclassifications under our models.

Logistic regression and SVM in particular are known to be sensitive to outliers [3][15]. Despite this documented sensitivity, the high accuracies of logistic regression and SVM indicate that there is a distinct split between data points with male and female labels, even when faced with outliers. The slight advantage of SVM over logistic regression implies that this decision boundary is linear or close to linear. The near optimal ROC curves for both logistic regression and SVM indicate that the models themselves are very robust.

Our other two models did not perform as well as logistic regression or SVM. Results for ANN were not as accurate as those of SVM or logistic regression. We hypothesize that either the small amount of data was insufficient to train the neural network, or, given the simplicity of vocal gender markers, our network may have been too complex. Increasing the number of epochs was initially beneficial, but after about 50 rounds, the accuracy stagnated at 90%. The decision tree is highly reliant on the initial data, as that is the data it constructs its splits from, and thus is unlikely to provide accurate results from any data set that deviates too much from the original.

Thus, we conclude that of the models presented, SVM with a linear kernel function is the most accurate and robust algorithm for classifying voices. The high accuracy of all our models also indicates the importance of features themselves in the final results; in essence, each model only needed to weight f0 highly to achieve good results. This strong reliance on particular features means that our models are near guaranteed to misclassify female and male voices deviating from mean female f0 and mean male f0, respectively. Further work could include methods of accounting for androgynous voices and accounting for our chosen model's sensitivity to outliers.

References

[1] Acero, A., and Xuedong Huang. "Speaker and Gender Normalization for Continuous-density Hidden Markov Models." *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (1996): n. pag. Web.

[2] Chatterjee S. and A. S. Hadi. "Sensitivity Analysis in Linear Regression", 1988, Wiley, New York.

[3] D. Pregibon, "Logistic regression diagnostics", Ann. Statist., 1981, 9, pp. 977-986

[4] Dietz, Timothy Alan. Speech Recognition Text-based Language Conversion and Text-to-speech in a Client-server Configuration to Enable Language Translation Devices. International Business Machines Corporation, assignee. Patent US6385586 B1. 7 May 2002. Print.

[5] Furui, S. "Cepstral Analysis Technique for Automatic Speaker Verification." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2 (1981): 254-72. Web.

[6] Harb, Hadi, Liming Chen. "Voice-Based Gender Identification in Multimedia Applications." *Journal of Intelligent Information Systems* 24.2-3 (2005): 179-98. Web.

[7] Klatt, Dennis H. "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers." *The Journal of the Acoustical Society of America* 87.2 (1990): 820. Web.

[8] Lee, K.H., S.I. Kang, D.H. Kim, and J.H. Chang. "A Support Vector Machine-Based Gender Identification Using Speech Signal." *IEICE Transactions on Communications* E91-B.10 (2008): 3326-329. Web.

[9] Levitan, Sarah Ita, Taniya Mishra, and Srinivas Bangalore. "Automatic Identification of Gender from Speech." *Speech Prosody 2016* (2016): n. pag. Web.

[10] Li, Ming, Kyu J. Han, and Shrikanth Narayanan. "Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion." *Computer Speech & Language* 27.1 (2013): 151-67. Web.

[11] M. J. Carey and E. S. Parris. 'A Speaker Verification System Using Alphanets', Proc ICASSP 1991, Toronto.

[12] Přibil, Jiří, and Anna Přibilová. "Spectral Flatness Analysis for Emotional Speech Synthesis and Transformation." *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions Lecture Notes in Computer Science* (2009): 106-15. Web.

[13] Traunmuller, H., Eriksson A. "The frequency range of the voice fundamental in the speech of male and female adults" (1995). Ms.

[14] Xu, Dongxin, Peiji Zhu, Taiyi Huang, and Daowen Chen. "Using High-level Linguistic Knowledge for Chinese Speech Recognition." *[1988 Proceedings] 9th International Conference on Pattern Recognition* (1988): n. pag. Web.

[15] Yang, Xulei, Qing Song, and A. Cao. "Weighted Support Vector Machine for Data Classification." *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (2005): n. pag. Web.

[16] Zuberec, Sarah E., Cynthia DuVal, and Benjamin N. Rabelos. Speech Recognition User Interface. Microsoft Corporation, assignee. Patent US6965863 B1. 15 Nov. 2005. Print.

6 Author contributions

Our contributions are as follows. The algorithms team wrote the initial code for each method and measured accuracy. This team consisted of Mengda Xu and Xie Zhou for SVM, Allan Soria and Yishen Huang for logistic regression, Yishen Huang for ANN, and Jaden Zhang, Haozhe Gu, Mat Magno for CART analysis.
The prediction team conducted research on how to best address the original problem; this team consisted of Haozhe Gu and Jason Vo.
Jason Vo and Daniel Burke analyzed the original audio data, created our feature extraction method, and computed the top features.
Haozhe Gu and Zhiyi Xu identified outliers within the data set and theorized about their effects.
Erin McGinnis did the report writeup.
Aaron Allen will create the presentation.