Gender Recognition by Voice and Speech Analysis

Jingyu Shao(shaojy15@ucla.edu)
Department of Statistics, 405 Hilgard Avenue
Los Angeles, CA 90024 USA

Abstract

The physical cues underlying the perception of speech can be used for various purposes. This paper aims at identifying a voice as male or female, based upon acoustic properties of the voice and speech. We designed several feature engineering approaches to first create corresponding prototypes for each category. Then we fed them into several classical supervised learning classification models, together with self-designed recurrent neural networks, to train a powerful classifier for gender recognition, and also conducted large-scale experiments to compare performances between different compulational models and humans.

Keywords: speech perception; gender recognition; supervised learning; recurrent nueral network

Introduction

Discrepancies between male and female voices are related to implicit and multidisciplinary issues, including not only acoustic measurements (mean frequency or magnitude), but also sociology or psychology (behavioral differences across genders). Our work focuses on the acoustic difference because it is more straight-forward and already sufficient for ordinary gender recognition problems.

Mean fundamental frequency is commonly considered to be the main discrepancy between male and female voice. The voiced speech of a typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz. But this may slightly vary through different ages, and even has relations with personal lifestyle. For example, those with a smoking habit have broadly lower voice frequency.

Also, the acoustic discrepancy across genders varies when it comes to different languages. For instance, a study in 2005 by Johnson () reveals that the across-gender variations in acounstic properties are relatively small in Danish but seem to be much more obvious in Russion. The difference even varies when it comes to different dialects. In a Chinses dialect, mean F0 is almost equivalent for male and female speakers[Rose, 1991, 2]; statistical analysis also reveals that Min speakers had a significantly greater maximum range of speaking intensity and a smaller lowest speaking intensity than Mandarin speakers[3].

In our study, we focus on selecting the most appropriate computational model for voice gender discrimination in challenging cirmstances. We also aims at finding suitable feature engineering approaches for feeding the statistical models.

Related Work

Feature Engineering

As a typical binary-classification problem, we need to first create statistical features from inexpressible data formats such as videos and audios with relative domain knowledge. We proposed 2 main approaches for audio data preprocessing.

Dataset Preparation

To distinguish genders from voices, we need raw data of speech recordings. Our dataset consists of 2 parts.

- We downloaded CMU_ARCTIC speech synthesis databases from Language Technologies Institute at Carnegie Mellon University. We selected 4533 utterances speaking out-of-copyright texts from Project Gutenberg[]. Half of them are pronounced by US males, and another half are pronounced by US females. We also downloaded a dataset of 1133 English utterances spoken by a Scottish male, for the purpose of researching into dialect influences. We only selected distributions of 16KHz waveform.
- To create more chanllenging experiment cirmstances, we also collected a small dataset of 30 Chinses utterances with lower signal-to-noise ratio (SNR) from 4 classmates, two of which are males. The average length of these recordings are around 4 seconds. We tested our classifiers on this more challenging dataset, and compared with human performaces.

Acoustic Measurements

There are some important acoustic cues of human voices and speeches, with an analyzed frequency range of 0hz-280hz (human vocal range). Intuitively, we care about the mean frequency and standard deviation of frequency. The minimum and maximum fundamental frequencies might also count. Here we proposed a list of 20 acoustic properties as basic classfication features, which can be acquired in R using the seewave and tuneR packages. Table **??** illustrates the 20 acoustic measurements.

Frequency Spectrogram

Since frequency is the critical indication of human voice properties, we also consider encoding frequency information into spectrums to better reveal across-gender differences. A simple 1-D singal from raw audios is illustrated in Figure **??**. The typical sampling rates for human speeches are 8KHz and 16KHz, as according to NyquistShannon sampling theorem():

If a function x(t) contains no frequencies higher than B hz, a sufficient sample-rate is 2B samples/second, or anything larger.

Human vocal range is from 0 to 280hz, and thus a sampling rate of 16KHz is gauranteed to capture all the information from a continuous-time audio signal of finite bandwidth.

We use simple spectrogram to preprocess the raw data. We split the audio into small windows of unit length (100ms), compute Fast Fourier Transforamtion (FFT) on the small audio clip and take magnitude. In this way, we can fully capture frequency contents in local window, as illustrated in Figure **??**. Finally frames are concatenated from adjacent windows to form "spectrogram" for each raw audio. Each pair of {spectrogream, label} is considered a training data point for our supervised learning model.

However, a critical problem occurs when we consider about the dimension of created features:

$$\begin{aligned} dim(feature) =& (\#audioClips, \#frequencyBins) \\ =& (\frac{len(audio)}{len(audioClip)}, \frac{max(frequency)}{2}) \end{aligned}$$

When the unit length of each audio clip and the maximum of frequency spectrum are fixed, the first dimension of feature matrices varies with respect to the length of input audio data. This would cause a problem for most supervised learning models.

Dimension Reduction

Computational Models

Classification is a typical problem in machine learning, which aims at developing an algorithm (classifier) that maps input features to a specific category. A large number of algorithms for classification can be phrased in terms of a linear function that assigns a score to each possible category *k* by combining the feature vectors of an instance with a vector of weights, such as logistic regression or SVM. We can concert them into non-linear classifiers by applying non-linear kernels. There are also methods that model conditional density functions for discrimination, such as naive Bayes classifier, which belong to the generative model category.

Classic Supervised Learning Models

SVM and Logistic

Naive Bayesian We use Bayesian inference to learn parametric models from our training data. Because Bayes classifiers are highly scalable and requiring a number of parameters, we need to train with large number of predictors in the learning process. The basic idea is to learn a parametric likelihood probability model from existing data. Since the overall proportion of male and female is close to 1:1, the prior probabilty is assumed to be $P(male) = 0.5$ and $P(female) = 0.5$. We have two basic assumptions for naive Bayes classifier: 1) independence of the predictor variables, and 2) Gaussian distribution (given the target class) of metric predictors.

Recurrent Nueral Network

Experiments

Conclusion

Discussion

Formalities, Footnotes, and Floats

Use standard APA citation format. Citations within the text should include the author's last name and year. If the authors' names are included in the sentence, place only the year in parentheses, as in ? (?), but otherwise place the entire reference in parentheses with the authors and year separated by a comma (?, ?). List multiple references alphabetically and separate them by semicolons (?, ?, ?). Use the "et al." construction only after listing all the authors to a publication in an earlier reference and for citations with four or more authors.

Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes in 9 point type at the bottom of the column on which they appear. Precede the footnote block with a horizontal rule.[2]

Tables

Number tables consecutively. Place the table number and title (in 10 point) above the table with one line space above the caption and one line space below it, as in Table 1. You may float tables to the top or bottom of a column, or set wide tables across both columns.

Table 1: Sample table title.

Error type	Example
Take smaller	63 - 44 = 21
Always borrow	96 - 42 = 34
0 - N = N	70 - 47 = 37
0 - N = 0	70 - 47 = 30

Figures

All artwork must be very dark for purposes of reproduction and should not be hand drawn. Number figures sequentially, placing the figure number and caption, in 10 point, after the figure with one line space above the caption and one line space below it, as in Figure 1. If necessary, leave extra white space at the bottom of the page to avoid splitting the figure and figure caption. You may float figures to the top or bottom of a column, or set wide figures across both columns.

Figure 1: This is a figure.

[1] Sample of the first footnote.

[2] Sample of the second footnote.

Acknowledgments

Place acknowledgments (including funding information) in a section at the end of the paper.

References Instructions

Follow the APA Publication Manual for citation format, both within the text and in the reference list, with the following exceptions: (a) do not cite the page numbers of any book, including chapters in edited volumes; (b) use the same format for unpublished references as for published ones. Alphabetize references by the surnames of the authors, with single author entries preceding multiple author entries. Order references by the same authors by the year of publication, with the earliest first.

Use a first level section heading, "**References**", as shown below. Use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 1/8 inch. Below are example references for a conference paper, book chapter, journal article, dissertation, book, technical report, and edited volume, respectively.