Gender Recognition using Power Spectral Density

Robert Irwin; Tyler Olivieri; Devin Trejo

Abstract*—Speech recognition software regarding gender identification usually revolves around finding the fundamental frequency of the signal. In this approach, we find trends in the spacing of fundamental frequency content for human speakers using a windowed approach. A classifier is implemented based on trends discovered from a small set of audio files.

I. Introduction

Speech recognition software regarding gender identification traditionally revolves around finding the fundamental frequency of the signal. In our approach, we aim to find trends in the spacing of the power spectral density over the fundamental frequency range of human speakers.

A. Background

The power spectral density, as it is computed in this paper, can be obtained by calculating the discrete autocorrelation (eq. 1) of the signal, and then the discrete Fourier Transform of the autocorrelation (eq. 2).

$$R_{xx}(l) = \sum_{n \epsilon Z} x(n)\bar{x}(n-l) \quad \ldots(1)$$

In (eq. 1), l represents the number of lags, and $\bar{x}$ is the complex conjugate of x. It is important to note that for real signals: $x = \bar{x}$.

$$S_{xx}(\Omega) = \sum_{n=0}^{N-1} R_{xx}[n]e^{-j\Omega n} \quad \ldots(2)$$

The power spectral density (PSD) is useful for the analysis of speech signals because the autocorrelation acts as a noise filter. Therefore, the resulting frequency spectrum highlights the actual frequencies composing the speech signal, and minimizes the contributions from stochastic noise.

In Figure 1, the effectiveness of the PSD as a noise filter can be observed. The frequency content of a 500 Hz sine wave with a SNR of -5 dB was measured using the FFT and PSD. The PSD showed the same frequency as the FFT but the results are less noisy.

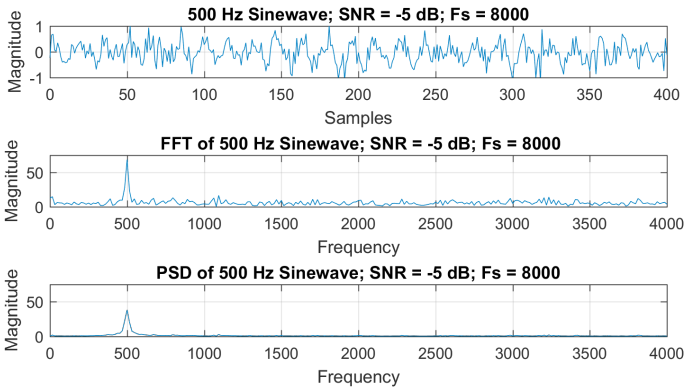

Figure 1: PSD acts as noise filter by looking for periodicity.

B. Characteritics of Human Speech

The average fundamental frequency of males and females is 132 Hz and 223 Hz, respectively (Zhao, O'Shaughnessy, & Minh-Quang, 2007). Males also tend to speak more monotonous than females, and their frequency spectrum tends to span a smaller range of frequencies. In contrast, females speak with more inflection and have a broader spectrum (Hanson & Chuang, 1999).

II. Approach

There are several existing algorithms that estimate the fundamental frequency (f_0), including autocorrelation, PSD, harmonic product spectrum, and cepstrum techniques. Our team focused on autocorrelation and PSD.

The initial approach was to find the fundamental frequency of each audio signal. The audio signal was divided into sections called windows and then each window was processed for f_0. The window was a Hamming window with length 100 ms. A decision was made once a second by averaging 10 windows of data.

We are looking for f_0, therefore, a low pass filter will be applied at 400 Hz before the signal is windowed. We are safe using a 400 Hz low pass filter because we expect f_0 under this range. It allows our classification technique to focus on our region of interest.

A. Autocorrelation

The autocorrelation of a periodic signal is periodic. For a simple sine wave the autocorrelation is periodic at the same period of the sine wave. For a speech signal the autocorrelation will be periodic with the fundamental period because the fundamental frequency dominates most of the wave shape in the time domain.

$$T_0 = \frac{1}{f_0}$$

When the dominant shape repeats itself in a speech signal, the autocorrelation value will be very high at that specific lag. A speech signal is only a quasi-periodic signal, however, if short windows are taken the signal can be assumed to be stationary and periodic. All that is needed to find T_0 is the lag with the first local maxima and the sampling frequency (F_s). The sampling period can then be found.

$$T_s = \frac{1}{F_s}$$

Now, the lag is multiplied by T_s to find the fundamental period and one can obtain an estimate of f_0.

The problem with autocorrelation, especially of speech signals, is that the autocorrelation is very complex and many local maxima occur. The problem here becomes relative peak detection. The lag at the first large peak is wanted for T_0 estimation. It is easy to visually detect this peak but it becomes a complex algorithm to program.

B. PSD

The PSD will show the frequency content of the autocorrelation function. Since the autocorrelation function is periodic with period T_0, the maximum peak of the PSD is taken as the estimate of f_0. To find an estimate of f_0 for the entire signal, the mean was computed of each window's f_0.

Finding f_0 using PSD provided an easier algorithm compared to autocorrelation as finding the highest value in an array is easier than relative peak detection.

C. Concentration of PSD

The f_0 we found was consistently in the female range, possibly due to a significant amount of background noise, so a different feature had to be found. We decided to approach the problem with the knowledge that females tend to have a broader spectrum in comparison to males, while males tend to speak more monotonous.

To obtain these features, a Hamming window of 100 ms was used to analyze portions of the signal. In each window, we computed the PSD, and found all prominent peaks. Our team considered a prominent peak as any local maxima in the spectrum whose magnitude was at least 40% of the absolute maxima. All prominent peaks were stored from 10 windows (1s) before the classification was made. The absolute peaks from each window were stored in one vector, and all prominent peaks were stored in another. The mean and variance of each vector was computed.

To obtain a measure of the monotony of the speaker, the difference between the variance of the fundamental frequency estimates, and the prominent peaks was computed. A low variance would signify similar frequency content in each window, which means that the speaker is monotonous, and more likely to be male. A large variance would signify a change in inflection and would most likely be a female speaker.

As mentioned previously, we were also interested in the concentration of the spectrum about the estimated f_0. The concentration of the spectrum was measured by computing the difference between the mean of the f_0 estimates, and the mean of the prominent peaks. If the prominent frequency content of the window was highly concentrated, the difference of the means would be small, which would signify a male speaker.

We also accounted for the possibility of no audio in the signal. To reduce the number of false classifications, we included a silence detection algorithm. Silence was found by looking at the total energy in the decision region. If the total energy of the decision region was less than 0.25% of the total signal energy then the window was classified as silence.

D. Building a Classifier

In order to use the approach above, in which we obtain two features, a classification algorithm was necessary. The classification vector was made by finding the monotony and concentration measure of the entire signal for 15 different audio files. The male and female results were averaged seperately. A linear function was implented in which the male average variance corresponded to -50 and the female average variance corresponded to 50, while the midpoint of the male and female variances correspond to 0. The same thing was done for the average means. The process is illustrated below.

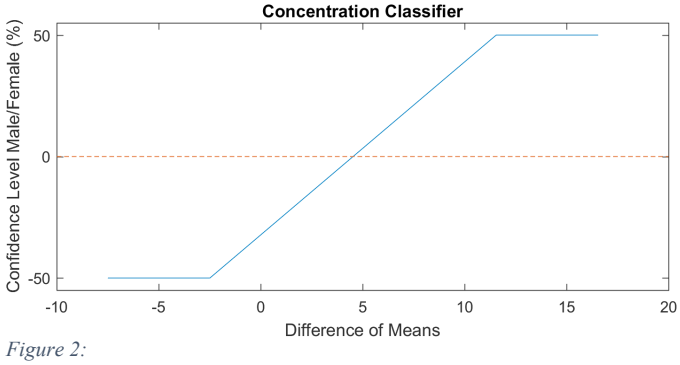

Figure 2:

The equation that governs the classifier above is,

$$class(x_i) = \begin{cases} -50, & x_i < meanM \\ \frac{100}{meanF - meanM}(x_i - midpt), & meanM \le x_i \le meanF \\ 50, & x_i > meanF \end{cases}$$

where, *meanF* and *meanM* represent the males' and females' difference between the means described in the *Concentration of PSD* section of the Approach, *midpt* is the midpoint of *meanF* and *meanM*, and x_i is the feature extracted at each decision.

Each time we make a decision, we plug the concentration and monotony measure into its corresponding classifier equation. The results are added, and a positive result corresponds to a female classification while a negative result corresponds to a male classification. The absolute value of the sum is given as the confidence of our decision.

III. Results

In our results, a confidence level is displayed as a percentage and color concentration overlaid on the portions of the time domain signal being classified. The darker blue the area, the more confident our algorithm is that the speaker is male. Conversely, the more pink concentration that exists shows that our algorithm believes the speaker is more likely to

be female. If our classification method could not determine if the speaker was male or female, it defaulted to predicting male with a confidence of 0%. Also, any portions of dark grey indicate that there was silence present in the signal and therefore did not aid or effect our classification metric.

We correctly classified 89 out of 130 (68.47%) windows in our training files. The statistics were obtained by summing all correct window classifications in a given signal with a speaker of a known gender and dividing by the total number of classifications. A sample of two classifications can be seen in Figure 3 and Figure 4. Figure 3 is a male speaker which can be observed to be correctly classified 10 out of 12 (83.33%). Figure 4 is a female speaker which was classified correctly 7 out of 9 (77.78%). All of our speech signals consisted of only one speaker. If a decision on the entire speech signal had to be made knowing that each signal contained only one speaker, our team classified 10 out of 15 (66.67%) speakers correctly.

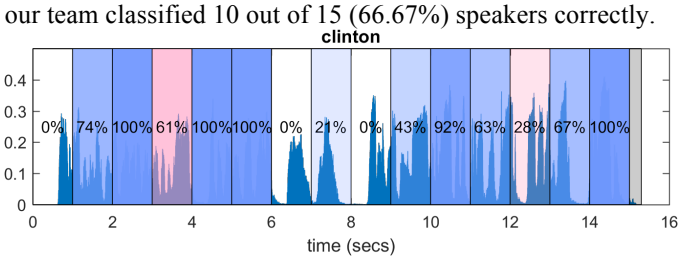

Figure 3: Bill Clinton Gender Classification

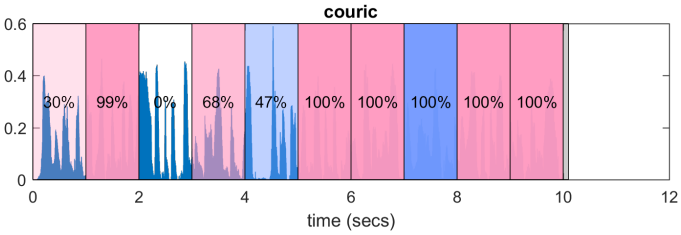

Figure 4: Katie Couric Gender Classification

The two signals shown above show good predictions while for other signals our predictions were not so great; in particular Condoleezza Rice. To compare, we plot their signal's spectrum in Audacity to see what may be causing the incorrect classification.

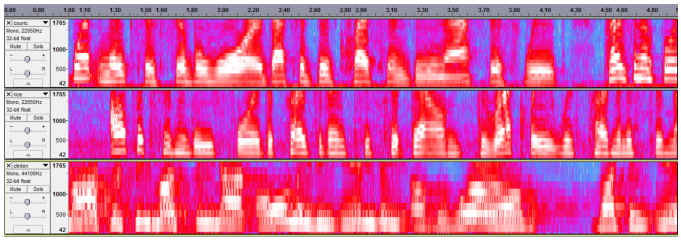

Figure 5: Spectrogram top to bottom: 'couric.mp3'; 'rice.mp3'; 'clinton.mp3'

The first spectrogram shown is the correctly predicted 'couric.mp3' signal. Overall, her spectrum is what we expect from a female; broad frequency spectrum with inconsistent frequency content across windows. The middle spectrogram represents our 'rice.mp3' speech signal. Her spectrum resembles that of a male. Notice the narrow bands of prominent frequency content, and consistency across windows. Compare 'rice.mp3' to 'clinton.mp3', the last spectrogram in Figure 4. The concentration of frequency harmonics is why we expect our prediction was poor on Rice's audio signal.

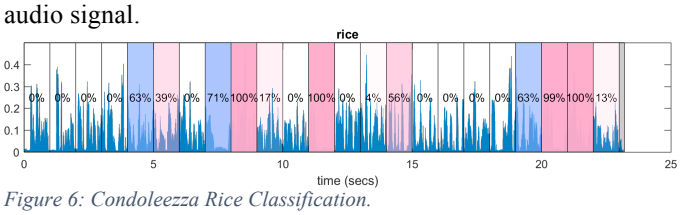

Figure 6: Condoleezza Rice Classification.

IV. Conclusion

In summary, our method of using inflection for gender classification was successful in predicting a majority of our speech signal's gender. Shorter signals (<5 seconds) were harder to classify since there was not enough data to make a confident prediction ~~on~~. For longer signals we had better classification with the one exception being 'rice.mp3'. Overall, our team was satisfied with the results and predict where a few errors may have occurred.

In our analysis we used a window size of 100 ms with no overlap between windows and a decision interval of one second. More experimentation needs to be done in finding what the ideal size is for these regions. Perhaps a window should span the time it takes a person to say a complete word and decision region should encompass the average time it takes a person to say an entire sentence. Our analysis did not focus on these features and instead showed proof of concept for using inflection as a feature in our classification method.

Another area for poor classification was the limited data set we had to work with. We trained our algorithm on a small data set of 15 speech files, where in ideal machine learning applications you train over "big data" sets. In our data set some of the files were provided with the project guidelines and others we recorded using a Mac Book Pro's built in microphone.

We predict that if we had a data corpus where all the signals were recorded using the same microphone and in the same location our results would improve. The added variables of different noise floors from different recording setups introduces more classification errors.

Further improvements can be added by looking at inflection of speakers of different languages and comparing the results to the English speakers analyzed in this paper.

V. References

Hanson, H. M., & Chuang, E. S. (1999). Glottal Characteristics of Male Speakers: Acoustic Correlates and Comparison with Female Data. *The Journal of the Acoustical Society of America*.

Zhao, Z., O'Shaughnessy, D., & Minh-Quang, N. (2007, Aug). A Processing Method for Pith Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches. *Signals, Systems and Electronics*, 59-62.