Real-Time Kinect Player Gender Recognition using Speech Analysis

Radford Parker

Figure 1. (a) the input RGB image, (b) the depth map overlayed with detected player genders, (c) the skeletal tracking with detected player genders

Abstract

The objective of this work is to efficiently identify a human's gender by using speech analysis from the streaming audio of the Microsoft Kinect. The current implementation of gender recognition performed by the Kinect involves image processing algorithms that search for gender characteristics through facial features. This paper shifts the classification away from the computationally expensive two-dimensional approach towards one-dimensional speech analysis. By analyzing the characteristics of the frequency spectrum of a player's voice, the algorithm can determine gender accurately and with only milliseconds of speech. This way, the player's gender can be determined almost instantaneously and without significant computational burden.

1. Introduction

The Microsoft Kinect has revolutionized the gaming industry. Instead of requiring a player to use a controller to interact with the gaming console, the player has now become the controller. The device uses an infrared depth sensor and computer vision techniques to figure out where players are located and their three-dimensional pose configuration. This technology has continued the trend of making gaming a more interactive experience.

This interactive experience means less of a burden on players, who now have a more natural interaction with their gaming console. Knowing a player's gender affords a game more information with which it can create a more realistic environment. Traditionally, this information is supplied by the user through a selection process. In an attempt to burden users even less, Microsoft has implemented image processing techniques that can determine a player's gender while they are facing the screen. These techniques perform adequately, but sometimes fail because they rely on an objective separation of the facial features belonging to each gender.

A different measure, which can better disambiguate between sets of gender-dependent features, is vocal pitch. The pitch, or fundamental frequency, of humans ranges from 75 Hz to 275 Hz. Adult males tend to occupy the range from 85 to 180 Hz, while adult females tend to occupy the range from 165 to 255 Hz [1].

Because of this distinct seperation of pitch frequencies, gender recognition by speech analysis can be performed extremely accurately. This change in the type of analysis not only can increase the accuracy of gender recognition, but can also allow for greater computational efficiency because of the switch from two-dimensional analysis to one-dimensional analysis.

2. Related Work

Gender recognition through speech analysis is a well researched topic. There are a variety of features that can be used to classify speech as either belonging to a male or a female. In addition, a variety of different classification techniques have been proposed to levy these features in order to make a more accurate classification.

One proven feature set is to perform spectral coefficient analysis on the speech signal [2, 3, 5, 11]. These approaches transform the frequency spectrum to the Mel Scale in order to obtain a vector of spectral coefficients. Two types of coefficients include the Mel Frequency Spectral Coefficients (MFSC) and the Mel Frequency Cepstral Coefficients (MFCC). It has been experimentally proven that the MFSC perform better for gender classification purposes [3].

Other proposed set of features focus on the acoustic or harmonic structure of the frequency spectrum [4, 12]. These methods are based on the location in the frequency domain of the first two formants. Essentially, these types of algorithms perform an automatic formant extraction technique which detects energy concentrations. The difference between these energy concentrations allows for classification of gender.

The simplest feature of gender classification relies solely on detecting the pitch frequency of the signal [4, 8, 10]. Instead of analyzing multiple energy concentrations in the frequency spectrum, these algorithms attempt to locate the first major concentration in the frequency spectrum. This method usually involves taking the Fast Fourier Transform of the signal and simply making inferrences about the first peak of the spectrum.

Combinations of these proposed features have been utilized in supervised learning-based classifiers [3, 4, 5, 6, 8, 10]. These types of algorithms depend on large training sets and are usually too computationally expensive to run on real-time streaming audio. Only recently has research been done on real-time approaches of gender recognition [9].

This paper details an algorithm that performs real-time gender recognition on streaming audio. It infers the pitch frequency of overlapping windows of speech from each player. The algorithm classifies each window of speech as male, female, or unknown, and then aggregates these decisions over time in order to remove errors caused by misclassification.

3. Methodology

There are many algorithms that will find the frequency spectrum of an audio singal. This work makes use of the Cooley-Tukey FFT algorithm. This implementation is very fast because it takes advantage of the Danielson-Lanczos Lemma. Esssentially, this method divides the signal into two seperate sub-signals composed of the odd and even samples, respectively. The FFT is computed on each segment and the output spectrums are combined. The equation for this process is

$$X_k = \sum_{m=0}^{N/2-1} x_{2m} e^{-\frac{2\pi i}{N/2} mk} + e^{-\frac{2\pi i}{N} k} \sum_{m=0}^{N/2-1} x_{2m+1} e^{-\frac{2\pi i}{N/2} mk}, \tag{1}$$

where k is the frequency and N is the number of samples.

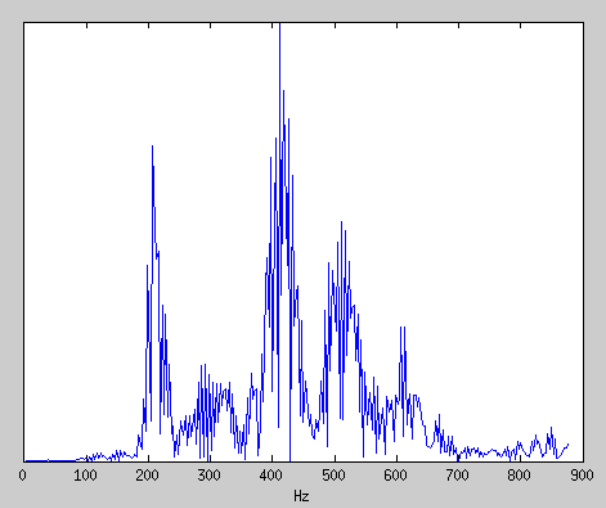

Figure 2. Spectrum output of FFT

The output spectrum from the Cooley-Tukey FFT is very noisy. An example spectrum of a female utterance is shown in Figure 2. A guassian smoothing kernal is applied to the signal according to

$$f(k) = \frac{1}{11} \sum_{i=-5}^{5} f(k+i) e^{-\frac{(k+i)^2}{2\sigma^2}}, \tag{2}$$

where $f(x)$ is the frequency at x. The output of this operation can be seen in Figure 3.

After the signal is smoothed, the algorithm detects the frequency with the most energy that lies in the range of human speech. While this frequency, p_l, represents the peak of the spectrum, it could still be a harmonic of fundamental frequency. In order to determine if this is the case, the algorithm searches the neighborhoods around $p_l/2$ and $p_l/3$ for smaller peaks.

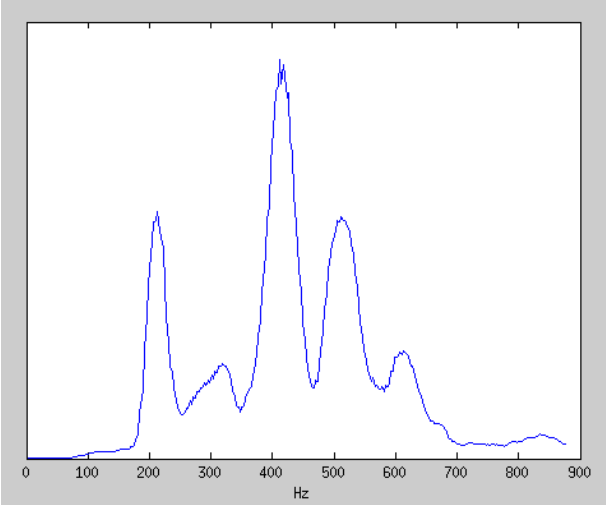

Figure 3. Spectrum after smoothing operation

These smaller peaks, p_s, occur at local maxima and must follow two conditions. First, the frequencies at a certain offset, o, must satisfy

$$\frac{((1-\frac{f(p_s-o)}{f(p_s)})+(1-\frac{f(p_s+o)}{f(p_s)}))}{2} > \tau_{off}, \quad (3)$$

where $f(x)$ is the magnitude of the frequency response at x Hz.

Second, this local maxima must represent a significant portion of the entire frequency spectrum given as

$$f(p_s) > \tau_{sig} \sum_{i=0}^{300} f(i), \quad (4)$$

where τ_{sig} is the threshold for signifigance.

If both of these conditions are met for some frequency, it is inferred that this smaller peak represents the fundamental frequency of the input signal. If no smaller peaks are found in the signal, then p_l is the chosen as the fundamental frequency.

This entire process is computed on overlapping hamming windows, which are weighted according to

$$w_i = 0.54 - 0.46\cos\frac{2\pi i}{f_s - 1}. \quad (5)$$

For each window, a fundamental frequency is computed. If this frequency falls within the range (75, 150) Hz, it is determined to be male speech. If instead, the frequency falls within the range (165, 275) Hz, then it is determined to be female speech.

Scores for each gender are aggregated over all windows. The gender with the larger score is chosen to be the gender of the speaker. If however, the score is tied, the gender is chosen according to the average pitch over all previous windows. If this average pitch is greater than 157.5 Hz, it is determine to be female, while anything else is classified as male.

4. Results

In order to make an accurate assesment of the this method, this exact algorithm was also implemented in an offline fashion. This way, it could be tested on large data sets of speech files without requiring the creation of an entirely new Kinect-based data set.

Results were obtained using the TSP speech corpus. This data set contains over 1,400 speech samples from 25 different speakers. This data was recorded at 48 kHz and all speech clips are one to four seconds in length. For all results listed here, τ_{sig} was 0.001, τ_{off} was 0.1, σ was 5, and o was $\frac{15}{binsize}$. Here, $binsize$ is defined as $\frac{f_s}{2} \times \frac{1}{spectrumsize}$ and $spectrumsize$ is 8192.

4.1. Classifcation

For the method outlined above, the contingency table for the 48 kHz data set is given in Table 1. The overall accuracy of the classification is 99%.

One of the problems with pitch detection techniques is that it can be susceptible to noisy signals. The Kinect device only samples at 16 kHz, so it is important that this algorithm also work for this type of signal. In order to address this concern, the TSP 16 kHz data set was also tested and the results are shown in Table 2. The overall accuracy of the classification for this down-sampled data set is 97%.

	Actual Male	Actual Female
Predicted Male	659	8
Predicted Female	1	776

Table 1. Contingency table for 48 kHz data

	Actual Male	Actual Female
Predicted Male	653	30
Predicted Female	7	754

Table 2. Contingency table for 16 kHz data

It is shown here that although 16 kHz data does cause the misclassification rate to rise, the algorithm is still very accurate.

4.2. Timing

The TSP 48 kHz data set was timed for how long it took each utterance to be classified by this approach. The algorithm was executed on an Intel i5 3.3 GHz processor with 8 GB of RAM and in a 64-bit architecture. On average, the decisions were made for a second's worth of windows in about 13 milliseconds. No utterance took more than 50 milliseconds to be completely classified over all windows. When running the entire dataset and adding the additional overhead and the scoring script, all utterances were classified in 58 seconds. This means that, on average, the classification was completed for a second's worth of windows in about 16 milliseconds.

Author	Year	Method	Data Set	Sampling Rate	Accuracy
Konig and Morgan [6]	1992	ANN on LPC	DARPA RMD	16 kHz	84%
Vergin et al. [12]	1996	Harmonic Relationships	ATIS	16 kHz	85%
Parris and Carrey [8]	1996	HMM on Pitch	OGI	8 kHz	97%
Slomka and Sridharan [10]	1997	GMM on Pitch	OGI	8 kHz	94%
Neti and Roukos [7]	1997	GMM on Acoustics	ATIS	16 kHz	95%
Harb and Chen [3]	2003	ANN on MFSC	-	-	92%
Harb and Chen [4]	2005	ANN on Acoustics and Pitch	-	22 kHz	93%
Ting at al. [11]	2005	GMM on MFCC	-	22 kHz	98%
Keyvanrad and Homayounpour [5]	2009	ANN on MFCC	OGI	8 kHz	96%
DeMarco and Cox [2]	2011	GMM on MFCC	TIMIT/ABI	16/22 kHz	95%
Proposed Method	2012	Aggregated Pitch Detection	TSP	16 kHz	97%
Proposed Method	2012	Aggregated Pitch Detection	TSP	48 kHz	99%

Table 3. Comparison to other works

4.3. Kinect Implementation

The same algorithm was implemented using the real-time streaming audio from the Kinect microphone array. Making use of the skeletal tracker demo that comes from the Microsoft SDK, the algorithm was able to tell when a new player has entered the scene. Once the audio signal from the microphrone was significant enough to be speech, the aforementioned pitch detector algorithm was run. This was completed in real-time and the resulting classification was used to color the player's skeleton in the GUI accordingly. An example output is shown in Figure 1.

4.4. Comparison

Although the data set used here is large enough to make assumptions about the algorithm's accuracy, it is difficult to compare to other works because every researcher does not use the same dataset. Presented in Table 3 are some results obtained for several different datasets. It is also important to note than most of these methods do not claim to work in real-time.

As seen in the table, despite the simplicity of the proposed algorithm, it can classify with similar accuracy to the most modern implementations. No absolute judgement should be made on which algorithm works the best until all methods have been run on the same data set.

5. Conclusion

This method does an excellent job of classification. For the application of real-time gender recognition for streaming audio from the Kinect device, the computational efficiency is maximized as desired.

Because we know that this classification has a margin of freedom, future work should take advantage of the extra time for a trade-off in increased accuracy. Certainly adding more features and training a supervised classifier could help alleviate some misclassifications.

Another interesting addition to this algorithm would be some form of sound source localization by using beam forming. Instead of classifying players as they enter the scene, the Kinect could localize where the speech is coming from and make an assumption about which player might be speaking.

References

[1] R. Baken. *Clinical Measurement of Speech and Voice*. 1987.

[2] A. DeMarco and S. Cox. An accurate and robust gender identification algorithm. *ICSA-11 Conference Proceedings*, 20011.

[3] H. Harb and L. Chen. Gender identification using a general audio classifier. *ICME-03 Conference Proceedings*, 2:733–736, 2003.

[4] H. Harb and L. Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 24:179–196, 2005.

[5] M. Keyvanrad and M. Homayounpour. Feature selection and dimension reduction for automatic gender identification. *CSICC-09 Conference Proceedings*, 14:613–618, 2009.

[6] Y. Konig and N. Morgan. GDNN a gender dependent neural network for continuous speech recognition. *IJCNN-92 Conference Proceedings*, 2:332–337, 1992.

[7] C. Neti and S. Roukos. Phone-context specific gender-dependent acoustic-models for continuous speech recognition. *ASRU-97 Workshop Proceedings*, pages 192–198, 1997.

[8] E. Parris and M. Carey. Language dependent gender identification: Acoustics, speech, and signal processing. *ICASSP-96 Conference Proceedings*, 2:685–688, 1996.

[9] E. Scheme, E. Castillo-Guerra, K. Englehart, and A. Kizhanatham. Practical considerations for real-time implementation of speech-based gender detection. *CIARP-06 Conference Proceedings*, 2006.

[10] S. Slomka and S. Sridharan. Automatic gender identification optimised for language independence. *TENCON-97 Conference Proceedings*, 1:685–688, 1997.

[11] H. Ting, Y. Yingchun, and W. Zhaohui. Combining MFCC and pitch to enhance the performance of the gender recognition. *ICSP-06 Conference Proceedings*, 2006.

[12] R. Vergin, A. Farhat, and D. O'Shaughnessy. Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female flassification. *ICSLP-96 Conference Proceedings*, 2:1081–1084, 1996.