Age and gender recognition from speech signals

Age and gender recognition from speech signals

Assim Ara Abdulsatar[1,3], V V Davydov[1, 3], V V Yushkova[4], A P Glinushkin[3] and V Yu Rud[1, 3]
[1]Peter the Great St. Petersburg Polytechnic University, St. Petersburg 195251, Russia
[2]University of Salahaddin, Erbil 44001, Iraq
[3]All Russian Research Institute of Phytopathology, Moscow Region 143050, Russia
[4]Saint Petersburg University of Management Technologies and Economics, 190109, Russia

E-mail: araasimyahya@gmail.com

Abstract. The aim of this research is to identify gender and age from speech, the system consists of two parts. The first part is called pre-processing and future extraction. The second part is called classification. This research investigates an automatic gender and age recognizer from speech. First four formant frequencies and twelve MFCCs are used to extract relevant features to recognize the gender. K-NN has been used as a classifier for the age recognizer model, stimulated using MATLAB. A special selectin of solid feature is used in this work to improve the accuracy of the gender and age classifiers based on the frequency range that the features represent.

1. Introduction

The difficulties that rise during the process of automatic speech estimation based on age are: Firstly, speech is affected by speaker's weight, height and mood, these characteristics may interact with age [1, -3]. Secondly, a huge database is needed for different speakers of different ages. Lastly, most of speeches in our daily life are from noisy environments, filters and high-quality microphones are required. Using different recording tools for recording the same sentence from the same speaker leads to different results.

Age and gender recognition have many practical applications, among them are: Commercial advertisements, the ads can become more relevant and target a specific group of age and gender, leading to an increase in sales. Another application is in forensic science, number of suspects can be reduced if there is an evidence such as a telephone call. This system can also be used for user authentication based on their speech.

There are different approaches to perform automatic gender and age recognition. For example, cepstral features, like Mel Frequency Cepstral Coefficients (MFCC). For age recognition. MFCC is known for producing poor results for gender and age classification with recorded signals [4]. To avoid this problem, the MFCC features are enhanced by analyzing the parameters that affect the process of extracting the features. MFCC has been used in many speech applications from speech recognition to language identification, and there is another acoustic feature that can be extracted, it is format frequency.

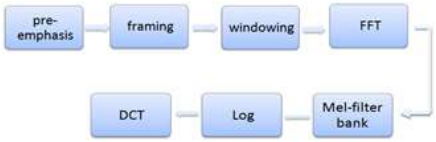

Figure 1. MFCC steps in speech analysis

The first stage of the process is pre-emphasis, this process is needed to increase the energy of the signal at high frequencies because the signal at higher frequencies has less energy compared to low frequency signals, this step is needed because the spectrum for voiced segments has more energy at lower frequencies than higher frequencies, this is called spectral tilt, spectral tilt is caused by the nature of the glottal pulse, therefore boosting high-frequency energy gives more info to acoustic model and consequently improves recognition performance, output response of pre-emphasis filter in n-domain is given below:

$$y(n) = x(n) - 0.95\, x(n-1) \tag{1}$$

Following pre-emphasis (1), the signal is divided into frames, this is known as framing. It is advantageous to frame signals to guarantee stationarity. The width of the frames is generally about 30ms with an overlap of about 20 ms (10 ms shift). Each frame carries N sample points of the speech signal. Overlap rate of frames is between 30 % and 75 % of the length of the frames, M=100 and N=256 that is why we frame the signal into 20-40 ms frames. If the frame is much shorter then we won't have enough samples to get a reliable spectral estimate, if it is longer then the signal will change too much throughout the frame. It is assumed that although the speech signal is non-stationary, but it is stationary for a short duration of time.

After framing, the window function is used to smooth the signal for the computation of DFT (discrete Fourier transform). There are different windowing functions. Commonly used ones are: Hamming, Blackman, Gauss, rectangular, and triangular. The hamming window is usually used in speech analysis because its spectrum falls off rather quickly, so the resulting frequency resolution is better, which is suitable for detecting formants. The equation is given below:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{M}\right) \tag{2}$$

Where *M* is the order of the filter, which is equal to the filter length – 1 and *n* is the index, the result of this process (2) is shown in the figure below:

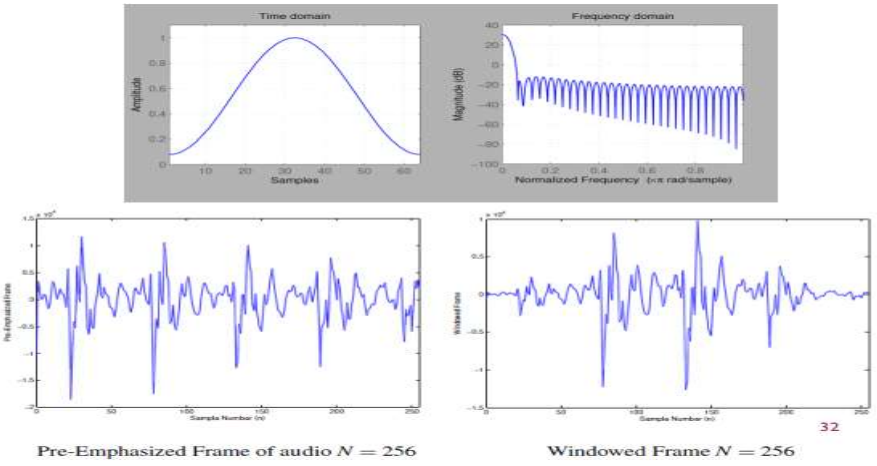

Figure 2. Original pre-emphasized frame (on left) vs. windowed frame (on right)

Fast Fourier Transform is used to convert each frame of N samples from time domain into frequency domain. This statement is supported the equation below:

$$F(x) = \sum_{n=0}^{N-1} f(n).\, e^{-j2\pi(x\frac{n}{N})} \tag{3}$$

Where *N* is the length of the discrete signal, *F(x)* is the signal in frequency domain and *f(n)* is the signal in time domain (3), FFT size can be 512, 1024 or 2048.

Human hearing is not sensitive to all frequency bands. Humans are less sensitive at higher frequencies, roughly > 1000 Hz. In other words, our hearing perception is non-linear. Human ear is like a filter that only concentrates on certain frequency components thus Mel-filter bank is implemented, this statement is shown in the figure below:

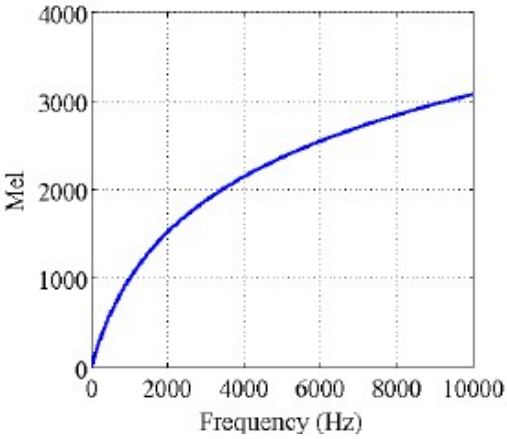

Figure 3. Mapping between the linear frequency scale and the Mel scale

Next step, the logarithm of the square magnitude of the output of Mel-filter bank is computed, dynamic range of values is compressed through logarithm, human response to signals is logarithmic. Human hearing is less sensitive to small differences in amplitude at high amplitudes than at lower amplitudes, this makes frequency estimates less sensitive to slight variations in the input, the picture below shows the signal before and after taking the logarithm:

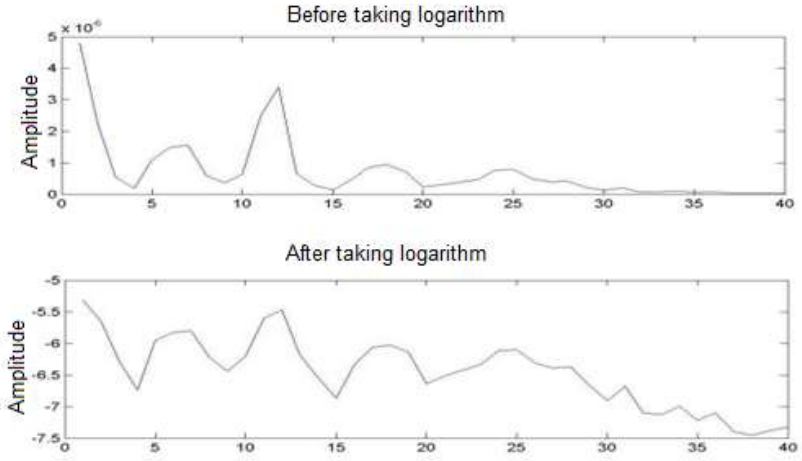

Figure 4. Amplitude of the signal before and after taking its logarithm

The last stage of MFCC is DCT, it's the process of converting the log Mel spectrum into time domain using DCT [2,6]. The result of this conversion is called Mel Frequency Sepstrum coefficients, this set of coefficients is called acoustic vectors. In fact, DFT should be applied because we are transforming the signal from frequency domain back to time domain using the following equation:

$$yt(k) = \sum_{m=1}^{M} \log(|Yt(m)|) \cos\left(k(m-0.5)\right)\frac{\pi}{M} \ , k = 0, \ldots . J \qquad (4)$$

Once the main features are extracted from the speech signal (4), classifiers are used. In this research, KNN (K-Nearest Neighbor) is used. KNN algorithm is one of the simplest machine learning algorithms. Data processing is carried out in MATLAB, which is actively used in various scientific fields [5-7]. In addition, a developed signals processing method can be used to obtain information in direction finding systems, radio monitoring and different communication systems [8-14].

2. Experimental results

The experimental results obtained for the prediction of age and gender using 40 clean speech samples in English. The model classes (seniors, adults and children from both genders), the gender recognition system reached 66 % recognition accuracy. While age recognition reached 55 % recognition accuracy.

Table 1. Rates of age recognition and rates of gender recognition using MFCCs only

MFCCs	Gender recognition rate	Age recognition rate
MFCC 1 2 3	55.55%	44.44%
MFCC 4 5 6 7 8 9	66.66%	44.44%
MFCC 10 11 12	22.22%	33.33%
All MFCCs	66.66%	44.44%

Table 2. Rates of recognition (age and gender merged) using MFCCs only

MFCCs	Recognition rate (age and gender together)
MFCC 1 2 3	33.33%
MFCC 4 5 6 7 8 9	44.44%
MFCC 10 11 12	11.11%
All MFCCs	44.44%

When using MFCC feature the best result was obtained using MFCC 4 to 9. The following tables are the results obtained using formants:

Table 3. Rates of age recognition and rates of gender recognition using formant frequencies only

Formants	Age recognition rate	Gender recognition rate
f3,f4	22.22%	22.22%
f3	38.88%	27.77%
f4	27.77%	44.44%
f1,f4	38.88%	55.55%
f1,f2,f3	50%	55.55%
f1,f2,f4	44.44%	50%
f1,f2	55.55%	61.11%
f1,f3,f4	33.33%	44.44%
f1,f3	50%	55.55%
f1	38.88%	38.88%
f2,f3	55.55%	50%

f2,f3,f4	38.88%	50%
f2,f4	50%	38.88%
f2	38.88%	38.88%
f1,f2,f3,f4	50%	66.66%

Table 4. Rates of recognition (age and gender merged) using formant frequencies only

Formants	Recognition rates (for age and gender together)
f1,f2,f3,f4	50%
f1,f4	33%
f1,f3	39%
f1,f2	55%
f1	28%
f1,f2,f3	39%
f1,f2,f4	39%
f1,f3,f4	28%
f2,f3,f4	33%
f2,f4	33%
f2	33%
f4	17%
f3	22%
f2,f3	39%
f3,f4	22%

Table 5. Rates of recognition using formant frequencies and MFCCs only

Formant frequencies	MFCCs	Age & gender	Gender	Age
All	All	50%	66.66%	50%
f1 f2	All	50%	55.55%	50%
f1 f2	MFCC 1 2 3	50%	55.55%	50%
f1 f2	MFCC 10 11 12	50%	55.55%	50%
f1 f2	MFCC 4 5 6 7 8 9	50%	55.55%	50%

When using formant frequencies feature the best result was obtained by using formant 1 and 2, best result for gender recognition obtained by using all formants together.
Gender recognition rate for different formants is shown in the table, excluding children [15, 16], we achieved a much higher rate (91%).

Table 6. Gender recognition rates using formant frequencies (without considering children)

Gender recognition (without children)	
Formants	**Gender**
f1 f2 f3 f4	83.33%

f1 f2	83.33%
f3 f4	91.66%
f3	66.66%
f4	66.66%
f1 f4	66.66%

3. Conclusion

In conclusion, gender and age can be extracted from speech signals. Using formant frequencies, the best result was obtained using formant frequency 1 & 2 because the main information of speech signal is concentrated at the low frequency part, the same is true for MFCC but the results were not improved when using MFCC because MFCC is not active with data that has been recorded at different places, with different computers and microphones, in order to improve the results, MFCC should be combined with another acoustic feature especially the fundamental frequency. In future the results can be improved further by using a smart classifier instead of the KNN, for example SVM or ANN.

Acknowledgments

I would like to express my deep gratitude to my lecturer (Fatima K. Faek) in the University of Salahaddin as well as my current professor (V. V. Krasnoshchekov) in Peter the Great Saint Petersburg Polytechnic University who helped me and motivated me in writing this paper which also helped me in doing a lot of research. I would also like to express my deep gratitude to my parents for their encouragement and support. This work partly supported by Grant of RFBR No. 18-29-25071.

References

[1] Bahari M H and Van hamme H 2011 *Proceedings of the 2018 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications* (Italy) 123-127
[2] Faek F K and Al-Talabani A K 2013 *International Journal of Computer Applications* **70(20)** 11-14
[3] G Dobry, M Hetch, M Avegal and Y Zigel 2011 *IEEE Trans. Audio, Speech and Language Processing* **19(7)** 1975–1985
[4] M Li, K J Han and S Narayanan 2012 *Computer Speech and Language* **27** 151-167
[5] R Davydov, V Antonov and N Kalinin 2015 *Journal of Physics: Conference Series* **643(1)** 012107
[6] J Stenis, W Hogland, M Sokolov, V Rud' and R Davydov R. 2019 *IOP Conference Series: Materials Science and Engineering* **497(1)** 012061
[7] R V Davydov, V I Antonov and D V Molodtsov 2018 *Journal of Physics: Conference Series* **1135(1)** 012088bnh
[8] Podstrigaev A S, Davydov R V, Rud' V Yu and Davydov V V 2018 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11118 LNCS** 624-630
[9] Moroz A V, Davydov V V, Rud V Yu, Rud Yu V, Shpunt V Ch and Glinushkin A P 2018 *Journal Physics: Conference Series* **1135(1)** 012060
[10] Davydov R V, Saveliev I V, Lenets V A, Tarasenko M Yu, Yalunina T R, Davydov V V and Rud' V Yu 2017 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10531 LNCS** 177-183
[11] Petrov A A, Davydov V V and Grebenikova N M 2018 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11118 LNCS** 641-648
[12] Logunov S E, Davydov V V, Vysoczky M G, Koshkin A Y and Rud' V Yu 2017 *Journal of Physics: Conference Series* **917(5)** 052058
[13] Davydov V V, Sharova N V, Fedorova E N, Gilshteyn E P, Malanin K Y, Fedotov I V, Vologdin V A and Karseev A Yu 2015 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9247** 712-721

[14] Myazin N S, Logunov S E, Davydov V V, Rud' V Yu, Grebenikova N M and Yushkova V V 2017 *Journal of Physics: Conference Series* **929 (1)** 012064
[15] Hugo M and Isabel T 2011 Project ACM: Transactions on Speech and Language Processing **7(4)** p. 16
[16] V Tiwari, G Ganga, J Singhai and M Azad 2011 *An International Journal (SPIJ)* **5(2)** 52