AGE ESTIMATION AND GENDER RECOGNITION BY SPEECH ANALYSIS USING NEURAL NETWORK

[1]Ashish Dongre, [2]Kushbu Undirwade, [3]Monika Kumbhare, [4]Nandeshwar, [5]Prof. Mohammad Ali
[1]B.E. Student, Department of Electronics & Telecommunication , ACET, Nagpur, Maharashtra, India
[2]B.E. Student, Department of Electronics & Telecommunication , ACET, Nagpur, Maharashtra, India
[3] B.E. Student, Department of Electronics & Telecommunication , ACET, Nagpur, Maharashtra, India
[4]B.E. Student, Department of Electronics & Telecommunication , ACET, Nagpur, Maharashtra, India
[5] Senior Professor, Department of Electronics & Telecommunication , ACET, Nagpur, Maharashtra, India

***ABSTRACT:** Automatic age and gender recognition for speech applications is very important for a number of reasons. One of the reasons is that it can improve human-machine interaction. For example, the advertisements can be specialized based on the age and the gender of the person on the phone. It also can help identify suspects in criminal cases or at least it can minimize the number of suspects. Some other uses of this system can be applied for adaptation of waiting queue music where a different type of music can be played according to the person's age and gender. And also using this age and gender recognition system, the statistics about age and gender information for a specific population can be learned. To remove the noise and to get the features of speech examples, some digital signal processing techniques were used. Useful speech features that were used in this work were: pitch frequency and cepstral representations. The performance of the age and gender recognition system depends on the speech features used. As the first speech feature, the fundamental frequency was selected. Fundamental frequency is the main differentiating factor between male and female speakers. Also, fundamental frequency for each age group is different. So in order to build age and gender recognition system, fundamental frequency was used. To get the fundamental frequency of speakers, harmonic to sub harmonic ratio method was used. The speech was divided into frames and fundamental frequency for each frame was calculated. In order to get the fundamental frequency of the speaker, the mean value of all the speech frames were taken. It turns out that, fundamental frequency is not only a good discriminator gender, but also it is a good discriminator of age groups simply because there is a distinction between age groups and the fundamental frequencies.*

***Keywords:** short-time average magnitude, short-time energy, short-time zero crossing rate, short-time auto-correlation*

I. INTRODUCTION

Voice is one of the most common means of communication in the world. The vibration of an object called the sound source causes surrounding air molecules to vibrate and spread. Such continuous sound vibration in the air or other media is what gives rise to sound. Human ears acquire and perceive the voice signal depending on the frequency of the sound, which allows us to distinguish the voice from different speakers or sound sources. Generally, voice contains a large number of sound waves with different frequencies, by which humans can recognize the attributes of each individual voice. In real life, human ears could usually verify a speaker by listening to the linguistic information. A voice contains a lot of linguistic information of the speaker. Human ears can identify some natural attributes (gender, age, origin, etc.) even when the voice comes from unfamiliar speakers. Therefore, voice features could be regarded as voiceprints capable of identifying the genders, ages, origins and emotional states of the speakers. Biometrics is a branch of computer science that studies the characteristics and traits of humans for identification and access control or surveillance purposes. There are two characteristics in biometric identifiers, physiological characteristics such as face recognition, DNA, retina or fingerprint and behavioural characteristics such as typing rhythm, voice or gait. In this project, our focus is on voice which is one of the behavioral characteristics in biometric identifiers Age and gender recognition for speech applications has many practical applications and it can be useful in many applications such as human-computer interaction or information retrieval. It can also improve the intelligibility of systems and can be helpful in speaker recognition and surveillance systems.

There are a number of reasons that show that automatic age and gender recognition is not an easy task. One of the first reasons is that each person's speech characteristic is unique so that makes classification a hard task. Also another challenge is the noise factor. Noise can be anything other than the speaker's voice. These problems are described in more detail below:

- Each speaker of the language is different. The difference comes from the vocal anatomy of speaker. One male and one female's speech characteristics can be very similar in terms of gender and also people from different age groups can have similar speech characteristics in terms of age classification. So in order to get good recognition results, the system must be trained on lots of data in order for the system to be accurate.
- The biggest problem is the noise factor. The noise can interfere with the actual speech and this can lead to wrong classification. Noise can be anything like, crowd of people noise, street noise, suburban train noise, car noise, restaurant noise or similar kind. So in order to have a reliable age and gender recognition system, some pre-processing techniques need to be applied to raw speech data and this noise needs to be eliminated. In this thesis, the focus is mainly on two speech features: pitch

II. LITERATURE SURVEY

The several works have been dedicated to DNNHMMs based large vocabulary continuous speech recognition. However, to knowledge only few works on the application of DNN-HMMs in age & gender recognition, have been reported. In [8], a Generalized Discriminant Analysis (GerDA) based on DNNs, is to learn the discriminative features for classifying high or low of arousal and positive or negative valence. Recently, most researchers have seen increased attention being given to decision level and model-level fusion in data fusion approaches. Accordingly, two popular data fusion approaches at decision and model levels: error weighted classifier combination and the coupled hidden

Markov model (C-HMM). The former used an empirical weighting scheme for recognition decision, and the latter modeled the asynchronous (e.g., audio and visual) nature of the multistream features for different applications. These models were successfully used in different fields such as age & gender recognition, interest detection, human identification, hand gesture recognition, 3-D surface inspection, speech prosody recognition, audio-visual speech recognition, and speech animation. Visual information has been shown to be useful for improving the accuracy of speech recognition in both humans and machines [4]. These improvements are the complementary nature of the audio and visual modalities. For example, many sounds that are confusable by ear are easily distinguishable by eye. The improvements from adding the visual modality are often more pronounced in noisy conditions where the audio signal-to-noise ratio (SNR) is reduced [5]. When developing a speech recognition system that incorporates both the audio and visual modalities, a principled method for integrating the two streams of information must be designed. Because of the success of hidden Markov model (HMMs) in audio speech recognition, most audio-visual speech recognition (AVSR) systems extend HMM techniques to incorporate both modalities. This is describe efforts in developing an AVSR system which is built upon existing segment-based speech recognizer [7]. This AVSR system incorporates information collected from visual measurements of the speaker's lip region using an audio-visual integration mechanism that we call a segment-constrained HMM [8].They are a new unified training algorithm for both the feature extractor and HMM classifiers. We interpret the feature extractor as a multilayer perceptron (MLP) with four layers, i.e., one for the filter banks, one for the feature transformation, and two for the delta and acceleration calculations. It enables us to derive efficient expressions of weight update formulas systematically by back propagation for all of the feature extractor modules. The back propagation starts with the output of HMM classifiers through an efficient inversion algorithm. Determining both the age and gender of speakers is a complicated task and has received considerable attention in recent years. The achieved are encouraging and are beginning to make it feasible to use this technology as a viable alternative to existing methods of providing user demographics. Age and gender classification systems are generally implemented as a fusion of several subsystems [6], with each subsystem operating using a form of Gaussian mixture model, multilayer perceptron, hidden Markov models and/or support vector machines [8]. If the phone is aware of its owner mood can offer more personal interaction and services. Mobile sensing, in recent years, has gone beyond the mere measure of physically observable events. Scientist studying affective computing [2], have published techniques able to detect the age & gender state of the user [2], allowing the development of age & gender-aware mobile applications [7]. Existing work focused on detecting age & genders rely on the use of invasive means such as microphones and cameras [5], and body sensors worn by the user [7]. There is method based on the employment of audio signals represents an efficient alternative to the mentioned approaches. The general influence of speaker age on voice characteristics is being studied since the late 1950s [1] and sustained continuous attention since then (see e.g. [2]), the first actual systems estimating the age and the gender of the speaker were developed only recently [6]. The quality of these systems is difficult to compare, as they vary considerably regarding the number and distribution of speaker age as well as the types of speech material. The variability of IVR system use patterns across age and gender is investigated in [7], indicating that dialog strategies tailored to specific age and gender groups can be very useful in improving overall service quality. In this context recognizing people age & gender state and giving a suitable feedback may play a crucial role. As a consequence, age & gender recognition represents a hot research area in both industry and academic field. There is much research in this area and there have been some successful products

1.1 Speech Analysis

The techniques used to process speech signals that can be broadly classified as either time-domain or frequency-domain analysis. In time-domain analysis, the measurements are performed directly on the speech signal to extract information. In frequency-domain analysis, the information is extracted after the frequency content of the speech signal computed to form the spectrum.

III. METHODOLOGY

Our goal is this thesis is to create a robust age and gender recognition system for speech applications which also gives good recognition rates under real world conditions. Some of the areas that this system can be used in are explained in more detail below: Phone Ads: A good use of Age and Gender Recognition System is phone ads. Many big• companies play phone ads while a customer waits on the line. So in order to play the same ad for every customer, it can be customized. And the end result is that, the ads become more efficient and possibly the sales increase because the ads are more relevant to the specific gender and age. Criminal cases: This is also a good example to the usage of this system. It turns out that, a• lot of times, in criminal cases the evidence is in the form of telephone speech. And by analyzing age and gender of the suspects, number of suspects can be narrowed down. Waiting queue music: This is also another example to the usage of the system. Waiting• queue music on phone lines can be customized according to the age and gender of the caller. This can help increase customer satisfaction of the companies. Statistics of a certain population: Age and gender recognition system can be handy when• researchers or companies collect age and gender information of a certain group of people. That information can help understand the experiment better and make better analysis.

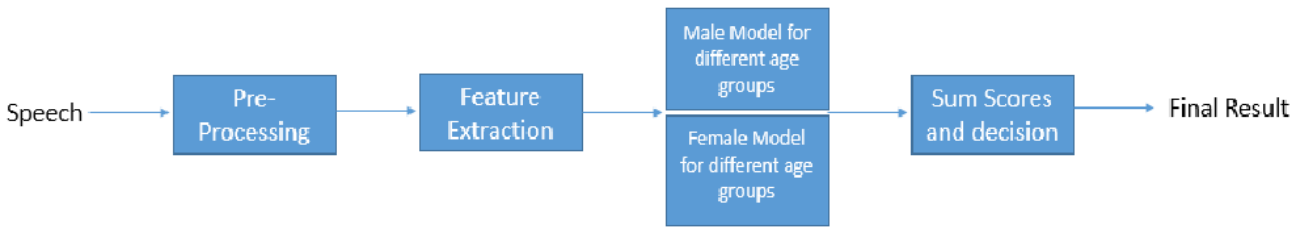

Figure 1. Age and Gender Recognition Model Trained Using Pitch as Features

3.1 Speech Preprocessing

Also, speech signals in time domain can be classified into three categories as voiced, unvoiced and silent speech as shown in Figure 1. Voiced sounds are periodic in nature and have higher energy than unvoiced sounds which are aperiodic and noise-like. The silence is when there is no speech and may have energy level related to the background noise.

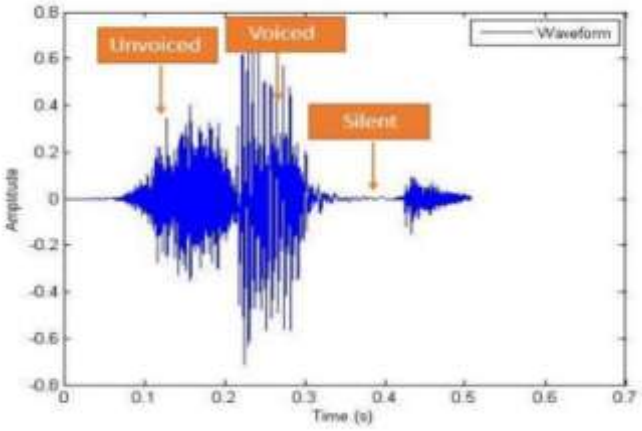

Figure 2 Time Domain Classification of the Speech

In time domain of speech signals, it can be seen that voiced sounds have a repeating periodic pattern. Each of these identifiable patterns is called a cycle. The duration of a cycle is called the pitch period and the fundamental frequency ($F0$) or pitch frequency is the inverse of pitch period ($T0$). Fundamental frequency shows how high or high low, a person's voice sounds.It is the frequency of his or her vocal cord vibration. Adult males typically have a fundamental frequency between 85 Hz to 155 Hz. Adult females, on the other hand have higher fundamental frequencies. An adult female fundamental frequency is in the range of 165 Hz to 255 Hz. Infants have much higher fundamental frequencies when they speak. It is generally between 250 Hz to 650 Hz. A ten year old boy or girl has a fundamental frequency of 400 Hz. When a person speaks, his/her fundamental frequency changes because of the structure of the language such as intonation and rhythm. So it is not easy to say that there is just one fundamental frequency of a person. However, when the person speaks in a natural voice, it is considered as his/her fundamental frequency.

3.2 Speech Framing and Windowing

Speech signal is a time-varying signal. It is stationary and changes over time. So in order for speech to be processed, it must be divided into non-stationary frames. The general size of speech frames varies between 10ms to 40ms where speech is said to be not changing. Once the speech signal is cut into frames, the next step is in many cases are windowing. Basically, the speech frame is multiplied by a window function. The most basic windowing function is the rectangular window. When this windowing is applied to the frame, none of thevalues of the frame changes.

3.3 Pitch Extraction Method

One of the features that can be used for age and gender recognition is pitch or fundamental frequency. As mentioned earlier, the male fundamental frequency changes between 85 Hz to 180 Hz, whereas a typical female fundamental frequency is between 165 Hz to 225 Hz. And also these fundamental frequencies change with age.

IV. RESULTS AND SIMULATION

This project was developed in Matlab. Some of the Matlab's libraries for signal processing and speech processing which will also give good recognition results under some kind of background noise or silence. So for real world applications and for a variety of speakers, the system will be able to recognize the age and the gender of the speaker. As a first approach of recognizing the gender of the speaker, the pitch information is used. Pitch is a fundamental difference between males and females. The library provided by] was used to extract the pitch information of the training examples. Each training example was
windowed at 25 ms. After extracting all the pitch information of all the frames, the mean value of all this windows was taken and it was considered as the pitch of the training example. Human speech normally varies between 100 Hz to 300 Hz so all the frequencies below 100 Hz and above 300 Hz were neglected

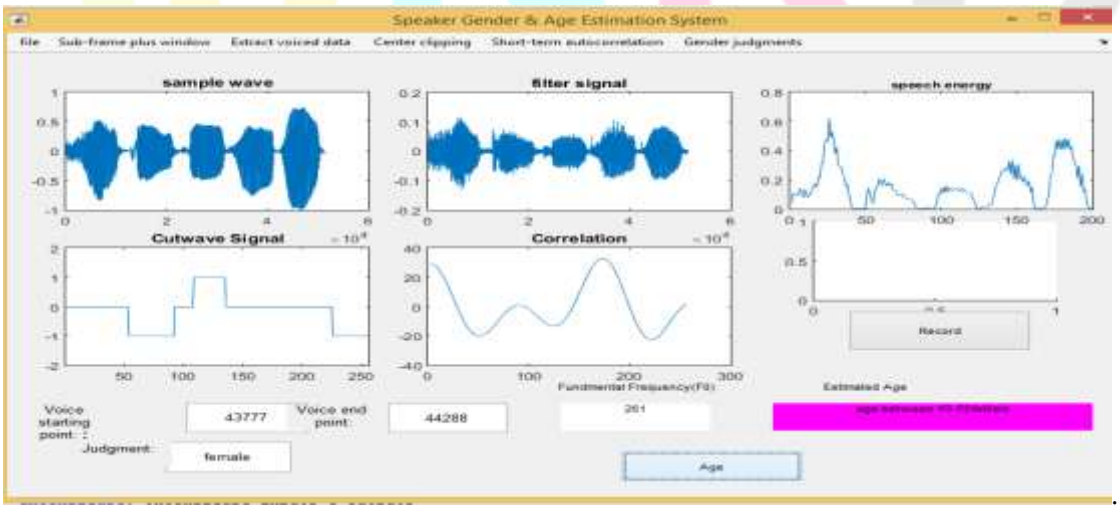

Figure 3 . Gui shows all the result that we processing

This approach is very different from the first approach. In this approach, a robust age and gender recognition system will be created. When finding out age and gender of the speaker, pitch information might be useful. A Graphical User Interface was developed for the real time tests. Anyone can record his/her voice by speaking into the microphone and using Matlab's internal recording commands the voice is recorded and then it is plotted in the time domain and in the frequency domain as a spectrogram. As soon as the speech is plotted, pre-signal processing techniques and feature extraction is applied to the speech and then the classification algorithms are run

V. CONCLUSION

By comparing the parameters obtained by short-time analysis of the male and female voice samples, it is observed that there is sufficient difference between the parameters. This difference in parameters can be used as the working principle of a Gender Classifier which predicts the gender of the speaker in a voice signal by analyzing it.

Our long term goal is to implement a gender classifier that can automatically predict the gender of the speaker based on the above investigation. The pitch data is applied to identify sex and age of speaker

VI. REFERENCES

[1] Parwinder Pal Singh and Pushpa Rani, An Approach to Extract Feature using MFCC, IOSR Journal of Engineering,Vol. 04, August. 2014.

[2] D.Shakina Deiv, Gaurav, Mahua Bhattacharya, Automatic Gender Identification for Hindi Speech Recognition ,International Journal of Computer Applications (0975 –8887) Volume 31– No.5, October 2011.

[3] M.A Anusuya and S.K. Katti, Front End Analysis ofSpeech Recognition- A review, Int J Speech Technol,Springer, DOI 10.1007/s10772-010-9088-7.

[4] M.Li, K. Han and S. Narayanan, automatic Speaker Ageand Gender Recognition Using Acoustic and ProsodicLevel Information Fusion, Computer Speech and Language,Jan 2013.

[5] H. Kim, K. Bae, H. Yoon, Age and Gender Classification for a Home-Robot Service, Proc. 16th IEEE International Symposium on Robot and Human Interactive Communication.

[6] Qiyue Liu, Mingqiu Yao, Han Xu, Fang Wang, Different Feature Parameters in Speaker Recognition, Journal of Signal and Information Processing 2013.

[7] Parwinder Pal Singh and Pushpa Rani, An Approach toExtract Feature using MFC, IOSR Journal of Engineering(IOSRJEN) Vol. 4, Issue 08(August 2014).

[8] Jamil Ahma, Mustansar, Fiaz, Soon-il Kwon, Maleera tSoanil, Bay Vo and Sung Wook Baik, Gender Identification using the MFCC for Telephone Applications- A Comparitive Study, International Journal of Computer Science and Electronics Engineering (IJCSEE) 2015.

[9] Jerzy SAS, Aleksaner SAS, Gender Recognition UsingNeural Networks and ASR Techniques, Journal of Medical Informatics and Technology 2013.

[10] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi,Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal Of Computing, Volume 2, Issue 3, March 2010.