# Feature Extraction from Children's Speech for Gender Classification

[1] Khin Aye Chan, [1] Su Su Maung [*], [2] Khine Thin Zar
[1, 1*, 2] Department of Computer Engineering and Information Technology
[1, 1*, 2] Yangon Technological University
[*]Corresponding Author: susuela@gmail.com

***Abstract*— Today classification of gender is one of the most important procedures in speech processing. A successful gender classification approach can boost the performance of many different applications as well as face recognition, smart human-computer interface and computer-aided physiological or psychological analysis. Gender identification task from children's speech is a challenging problem as there's no significant difference in the acoustic properties of male and female children. This paper is about investigation on the efficient features to discriminate the gender from children's speech. The Mel Frequency Cepstral Coefficient (MFCC) method is used for extracting features from speech signals. This is one of the most popular feature extraction techniques used in speech recognition. Voice samples for feature dataset are collected from children of age range 6 to 11 years, both male and female. In present system ACID pro voices editing software is used at the stage of preprocessing audio files and then first 12 MFCCs are extracted from the preprocessed signal. Features are evaluated using a nonlinear classifier namely Random Forest (RF) for gender classification from children speech. Experimental result represents that proposed system of using MFCC for gender prediction have good accuracy.**

*Keywords*: ***Gender Classification, Feature Extraction, Speech Recognition, MFCC, RF***

## I. Introduction

Speech recognition is the knowledge base subfield of linguistics that develops methodologies and technologies that permit the recognition and translation of spoken communication into text by computers. It's conjointly called automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates data and analysis within the linguistics, computer science, and electrical engineering fields. Some speech recognition systems need "training" (also referred to as "enrollment") wherever a private speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the identification of that person's speech, leading to inflated accuracy. Systems that don't use training are referred to as "speaker independent" systems. Systems that use training are referred to as "speaker dependent" [9].

Automatic gender classification from speech is widely applied in speech recognition. Many applications including speaker identification, speaker segmentation, smart human computer interaction, biometrics social robots, audio or video content indexing, etc use gender classification. Gender identification can improve the prediction of other speaker traits such as age and emotion, by jointly modeling gender with age (or emotion) or either by together modeling gender with age (or emotion) or during a pipelined manner. Speaker verification systems additionally implicitly or expressly use gender information.

Generally humans can easily identify a person's age, gender and emotion by hearing to his/her voice. In some circumstances like conversations over the telephone, the genders of adults are easy to identify, but the genders of children are difficult to identify. The acoustic and linguistic characteristics of child speech are particularly different from those of adult speech. Children generally have higher elementary and formant frequencies than those of adults, thanks to a shorter vocal tract, smaller vocal folds, developing articulators (e.g. tongue size and movement) [8][10].Therefore, there is no significant difference in their acoustic-phonetic properties in both male and female child. In this system, MFCC features are extracted from children speech and prediction is done basically on the feature dataset using RF classifier.

Rest of the paper is categorized as follows. Section II discusses previous works in the field of children gender classification. Section III describes the methodology employed for feature extraction. Proposed system overview is shown in Section IV. Results are given in Section V. Paper is concluded by declaring some future research directions in Section VI.

## II. Literature Review

Classifying the gender of a person based on their voice is a challenging problem in speech recognition. There are numerous machine learning, deep learning models to classify the person is male or female based on speech. Deep learning models are more appropriate for unstructured data such as audio, video and images. Deep learning models perform better results when the data is large. Many researches have made for gender identification in adults using various classification approaches and feature combinations. Feature extraction from a given signal is the most significant phase in gender identification. For efficacy of recognition mechanism robust features are required. In paper [7] the main part of study is the feature extraction mechanism with the detailed process for the MFCC features. There are many techniques for the feature extraction but the advantage of using MFCC technique as a method for extracting features is coming our robust and concise features that produces with high accuracy in the recognition process and provide with effective results during classification mechanism. The accuracy computed with MFCC technique results to be 95% which is best result obtained compared to other feature extraction methods.

For classification tasks, distance measures, Gaussian mixture model (GMMs) and hidden Markov model (HMMs) is employed as earlier approaches [5]. Recently, GMM-Universal Background Model (GMM-UBM) and GMM-Support Vector Machine (GMM-SVM) have shown significant improvement in gender identification. GMMs and SVMs are most commonly used in gender identification with MFCCs, formants, pitch and its statistical variations. However, there can be extent for increasing the accuracy using different combinations of features and implementing suitable classifiers based on the nature of the data. As of today, Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs) and RF are generally used for classification [3].

These analyses mainly focus on gender classification for adults and do not examine the gender classification for children. Gender classification for children is more difficult than adults as there is no obvious difference in the acoustic properties of male and female children. As of today, very few approaches have focused on gender identification from children's speech. Different combinations of spectral, prosodic and excitation source features are explored for the task. Spectral features namely MFCCs; prosodic features such as pitch are mostly used in many approaches for this task. As the features are already known the classification task requires the implementation of classifiers on the data [1]. Studying over 21 frequency sub-bands regions of the spectrum shows that the frequency range less than 1.8 kHz and greater than 3.8 kHz are efficient in discriminating gender in older children. Frequencies greater than 1.4 kHz are useful for the youngest children [6].

In paper [4] MFCCs (39), Pitch (4), Formant (4), Shimmer (2), Jitter (2) and LPCCs (13) are used as features for children gender identification. Different combinations of these features are used for classification to evaluate the efficiency of the proposed approach. Classifiers that is efficient in discriminating non-linear data namely, ANN, DNN and RF are used and compare the classification accuracy results. The RF classifier exceed the other classifiers with an average accuracy of 84.79% for gender classification. DNN achieves an average accuracy of 78.25%. ANN did not give satisfying accuracy results as compared with other two classifiers.

## III. Methodology

The extraction of the relevant and important information from the speech signals of the human voice is an important task to produce a latter recognition performance. The result efficiency of feature extraction step is crucial for the next step like modeling, classification and feature matching since it affects its behavior. Linear Predictive Coding (LPC), MFCC, Zero Crossings with Peak Amplitudes (ZCPA), Dynamic Time Warping (DTW) and Relative Spectral Processing (RASTA) are commonly used as feature extraction techniques for speech recognition system. LPC is one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. ZCPA is mostly used for development of automatic speech recognition in noisy environments, speaker identification, throat signal analysis, development of noise robust speech recognition system etc. DWT has been applied to temporal sequences of video, audio, and graphic. DTW is commonly used for measuring similarity between two temporal sequences which may vary in time or speed. RASTA method is generally used for speech analysis in which speech signals are enhanced to develop noise robust speech recognition system and etc.

The most widely used features for speech recognition are the acoustic features namely MFCC. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound. The power spectrum describes the distribution of power into frequency components composing that signal. MFCCs are coefficients that collectively make up an MFC. The reason for MFCC being most commonly used for extracting features is that it is most nearest to the actual human auditory speech perception. This method is considered to be the best available approximation of human ear. MFCC feature extraction method is less complex in implementation and more effective and robust under various conditions. It is a standard method for feature extraction in speech recognition. Fig. 1 shows the detail process of MFCC feature extraction of audio files [5].

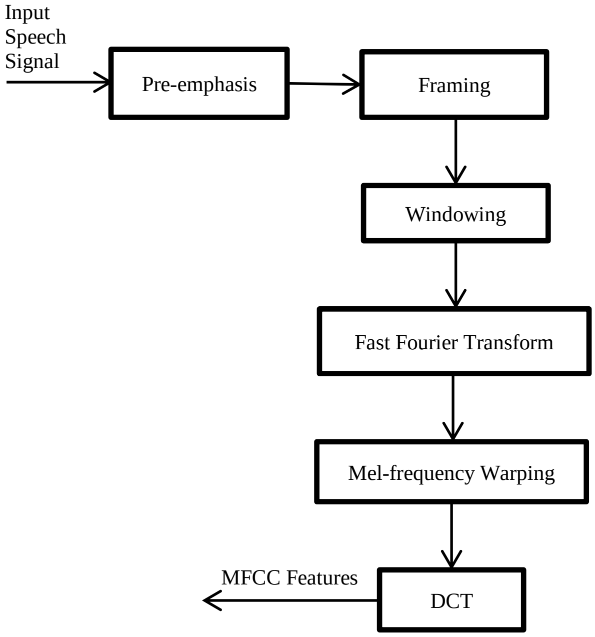


Figure 1. Steps for Computing MFCCs [5]

### A. Pre-emphasis

Due to the structure of human speech production system, damping occurs in high-frequency regions. Especially, higher frequencies will have less energy compared to the lower ones, thus, getting poor results with the prediction model. To cope with this, a high pass filter is applied on the signal in order to enhance these components. The filter increases the energy of higher frequency signal and decreases the energy of lower frequency signal and obtains a much evenly distributed spectrum. This is called the pre-emphasizing step [4]. Widely used pre-emphasis filter is given as,

$$Y(n) = x(n) - \alpha * x(n-1), \alpha \approx (0.95 – 0.97) \quad (1)$$

Where, Y (n) = pre-emphasis signal, x (n) = input signal and α= filter coefficient.

### B. Framing

Like in all voice analysis ways, also MFCC methodology is applied on the short parts where voice has stationary acoustic features. Framing is the process of blocking the speech signal into short portion of n samples known as frames in the time domain [2]. These frames are generally selected as 20-30 milliseconds, a shift of 10-15 milliseconds along the signal. If the signal is framed into 25ms, the frame length for a 16 kHz signal is 0.025*16000 = 400 samples. If the frame step is 10ms (160 samples), that permits some overlap to the frames, the earliest 400 sample frame starts at sample 0, the next 400 sample frame starts at sample 160 etc., until the ending of the speech file is reached.

### C. Windowing

After framing step, every individual frame is windowed using window function. Each of the above frames is multiplied with a window function to minimize signal discontinuities at the beginning and at the end of each frame [2]. This step makes the end of each frame connects smoothly with the beginning of the next. MFCC uses hamming window and the equation is as follows:

$$\mathrm{Y(n)} = \mathrm{x(n)} \,.\, \mathrm{W(n)} \tag{2}$$

Where, n = number of samples in each frame, Y (n) = output signal, x (n) = input signal and W (n) = Hamming window.

The graph of the Hamming window is as follows in Fig. 2 and Hamming window has the following form:

$$\mathrm{W(n)} = 0.54 - 0.46 \cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N\text{-}1}}\right) \tag{3}$$

Where, $0 \le n \le N-1$, $N$ = the window length [5].

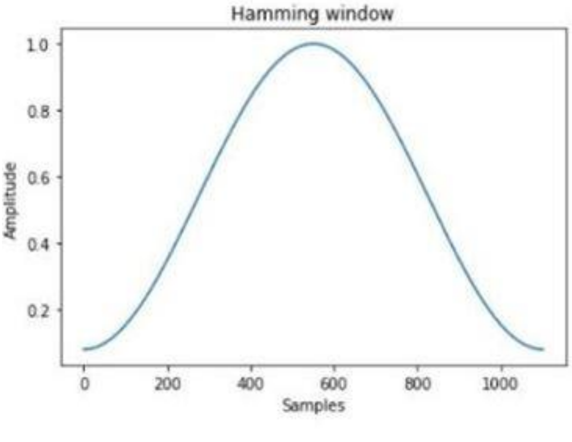


Figure 2. Hamming Window

### D. Fast Fourier Transform

Fast Fourier Transform (FFT) is a process of changing time domain into frequency domain. N-point FFT is applied on each frame to calculate the frequency spectrum, that is also called Short-Time Fourier-Transform (STFT), where N is typically 256 or 512. The FFT size must be longer than the frame length. If N is smaller than the length of the input, the input is removed. If it is larger, the input is filled with zeros. After that, compute the power spectrum (periodogram) using the following equation [2]:

$$\mathrm{P} = \frac{|\mathrm{FFT}(\mathrm{x_i})|^2}{\mathrm{N}} \tag{4}$$

Where, P= power spectrum (periodogram), $x_i$ = the $i^{th}$ frame of signal x.

### E. Mel-frequency Warping

The mel-scale correlates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at recognizing small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale make our features match more closely what humans hear. To convert the achieved amplitude spectrum into mel-scale, a filter set placed linearly with respect to mel-scale is used. This set consists of triangle band pass filters that are overlapping 50% and generally, filter coefficient is selected between 20 and 40.Each filter in the filter set is triangular starting at the first point, reach its peak at the second point, then return to zero at the 3rd point. The second filter will start at the 2nd point, reach its max at the 3rd, then be zero at the 4th etc. At this stage, mel filter bank energies is obtained by multiplying power spectrum of the signal with mel filter bank. After this stage, take the log of each of the energies from the above step and get log filter bank energies [2]. Fig. 3 describes a mel filter bank containing 26 filters. This filter bank starts at 0Hz and ends at 250Hz.

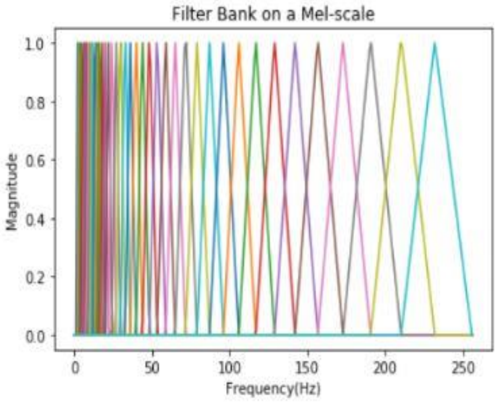


Figure 3. Mel Filter Bank

### F. Discrete Cosine Transform (DCT)

DCT is applied on the log energy $E_k$ obtained from the triangular band pass filters to have L mel-scale cepstral coefficients. DCT formula is shown below:

$$C_m = \sum_{k=1}^{N} \cos\,[m*(k\text{-}0.5)*\pi/N] * E_k\,,\ m = 1,2 \ldots L \tag{5}$$

Where, $C_m$ = coefficient values ,N = number of triangular band pass filters, L = number of mel-scale cepstral coefficient, $E_k$ = log energy.

DCT transforms the frequency domain into a time domain. Typically, for ASR, the resulting cepstral coefficients 2-13 are retained and the rest are discarded. The resulting features (12 numbers for each frame) are called MFCCs [2].

## IV. System Overview

The proposed framework for the classification of children's gender is revealed in Fig. 4. The system has three main processing stages. The first stage is preprocessing the speech signal. In preprocessing, noise reduction and removal of silences and unvoiced regions are included. The second stage involves feature extraction. In the last stage, extracted features are evaluated using RF classifier to predict children's gender.

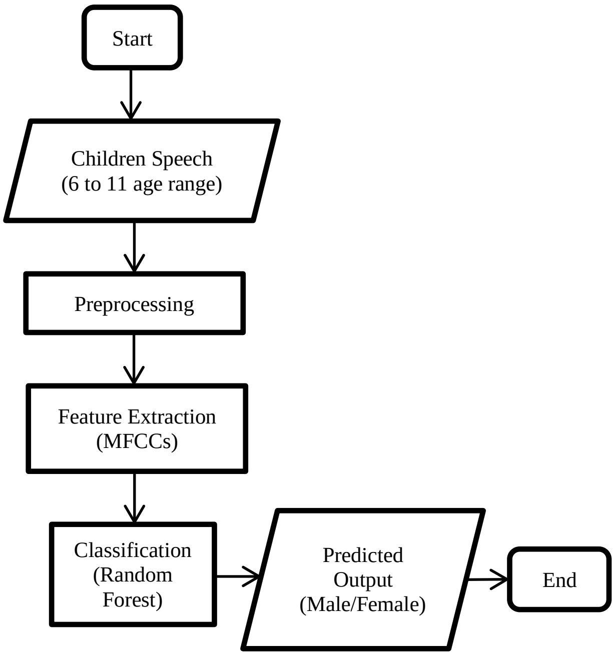


Figure 4. Flow Diagram of the Proposed System

### *A. Data Collection*

In this system, the main objective is distinguishing between the male and female voice, the 750 audio records were provided from the children of a primary school. All the speakers were asked to utter the same 10 sentences and recorded utterances with SONY Digital Stereo High Definition. These voice clips are preprocessed and evaluated.

### *B. Preprocessing*

Only voiced region of speech contains most of the gender related information. Many silence and unvoiced regions may be consisted in the recording files of children. These regions are removed from the speech using librosa.effects.trim function which is a build in function of Python. Noise reduction from these voice clips is done with using ACID pro voice editing software.

### *C. Feature Extraction*

MFCCs which are short term spectral based features are extracted from children speech. MFCC features are frequently used by many researchers for speech recognition and in music/ speech classification problem.

### *D.Classification*

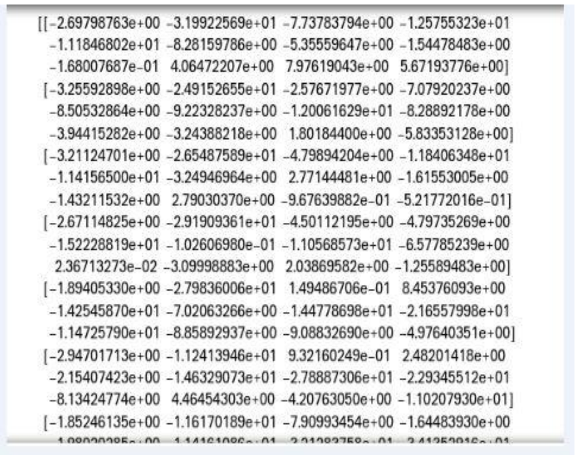
RF is a supervised learning algorithm. It is an ensemble of decision trees, most of the time trained with the "bagging" method. The main idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: RF builds multiple decision trees and merges them together to get a more accurate and stable prediction.

## V. Results and Discussion

The first stage of speech recognition is compressing a speech signal into streams of acoustic feature vectors, referred to as speech feature vectors. The extracted vectors are assumed to have sufficient information and to be compact enough for efficient recognition. The concept of feature extraction is actually divided into two parts: first is transforming the speech signal into feature vectors; secondly is to choose the useful features which are insensitive to changes of environmental conditions and speech variation.

The present system is based on converting the source wave file into a speech signal and then extracting MFCCs for each frames of the converted signal. But only the first 12 of the DCT coefficients are kept. This is as a result of the upper DCT coefficients mean quick changes within the filterbank energies and it seems that these quick changes truly degrade ASR performance, therefore higher classification results will get by dropping them.

Fig. 5 describes a MFCC matrix having features vectors extracted from all the frames of a speech signal. It is a numpy array of size (numbers of frames) and each frame contains 12 MFCC features. Each row represents one feature vector. After feature extraction, a feature dataset consisting of coefficients values of each audio file is established and is used for modeling classifier.



```
[[-2.69798763e+00 -3.19922569e+01 -7.73783794e+00 -1.25755323e+01
  -1.11846802e+01 -8.28159786e+00 -5.35559647e+00 -1.54478483e+00
  -1.68007687e-01  4.06472207e+00  7.97619043e+00  5.67193776e+00]
 [-3.25592898e+00 -2.49152655e+01 -2.57671977e+00 -7.07920237e+00
  -8.50532864e+00 -9.22328237e+00 -1.20061629e+01 -8.28892178e+00
  -3.94415282e+00 -3.24388218e+00  1.80184400e+00 -5.83353128e+00]
 [-3.21124701e+00 -2.65487589e+01 -4.79894204e+00 -1.18406348e+01
  -1.14156500e+01 -3.24946964e+00  2.77144481e+00 -1.61553005e+00
  -1.43211532e+00  2.79030370e+00 -9.67639882e-01 -5.21772016e-01]
 [-2.67114825e+00 -2.91909361e+01 -4.50112195e+00 -4.79735269e+00
  -1.52228819e+01 -1.02606980e-01 -1.10568573e+01 -6.57785239e+00
   2.36713273e-02 -3.09998883e+00  2.03869582e+00 -1.25589483e+00]
 [-1.89405330e+00 -2.79836006e+01  1.49486706e-01  8.45376093e+00
  -1.42545870e+01 -7.02063266e+00 -1.44778698e+01 -2.16557998e+01
  -1.14725790e+01 -8.85892937e+00 -9.08832690e+00 -4.97640351e+00]
 [-2.94701713e+00 -1.12413946e+01  9.32160249e-01  2.48201418e+00
  -2.15407423e+00 -1.46329073e+01 -2.78887306e+01 -2.29345512e+01
  -8.13424774e+00  4.46454303e+00 -4.20763050e+00 -1.10207930e+01]
 [-1.85246135e+00 -1.16170189e+01 -7.90993454e+00 -1.64483930e+00
```

Figure 5. The Numeric Representation of MFCC Matrix

Separating data into training and testing sets is an important part of evaluating classification models. Classification model is built on the training set and check the accuracy of the model by using it on the testing set. Currently the system uses 630 audio records are used as training dataset and 120 records are used for testing. From table I, the performance of RF classification model can be found. RF classification model achieves average accuracy of 93%.

TABLE I. CONFUSION MATRIX AFTER EVALUATING RF CLASSIFIER

| **Actual Group** | **Predicted Group** | |
|---|---|---|
| | ***Male*** | ***Female*** |
| ***Male*** | 61 (TP) | 4 (FN) |
| ***Female*** | 5 (FP) | 50 (TN) |

$$\text{Accuracy}= \frac{\text{TP+TN}}{\text{N}} \times 100\% = \frac{61+50}{120} \times 100\ \% = 93\% \quad (6)$$

$$\text{Specificity}= \frac{\text{TN}}{\text{TN+FP}} \times 100\% = \frac{50}{50+5} \times 100\ \% = 91\% \quad (7)$$

$$\text{Sensitivity}= \frac{\text{TP}}{\text{TP+FN}} \times 100\ \% = \frac{61}{61+4} \times 100\ \% = 94\% \quad (8)$$

In (6), (7) and (8), TP is true positives (the positive tuples that were correctly labeled by the classifier).TN is true negatives (the negative tuples that were correctly labeled by the classifier). FP is false positives (the negative tuples that were incorrectly labeled as positive). FN is false negatives (the positive tuples that were mislabeled as negative) and N is total number of tuples of testing dataset.

Table II shows the comparison of the proposed MFCC feature extraction technique with other feature extraction technique namely Linear Predictive Analysis (LPC). MFCC is derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence had the better response compared to LPC parameters. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other techniques.

TABLE II. COMPARISON OF FEATURE EXTRACTION TECHNIQUES

| No. | Technique | Classification Accuracy |
|---|---|---|
| 1 | Mel Frequency Cepstral Coefficient (MFCC) | 93% |
| 2 | Linear Predictive Analysis (LPC) | 87% |

## VI. Conclusion

In this paper the major discussion is the feature extraction method with the detailed process for the MFCC coefficients. There are a large number of techniques for the feature extraction but the advantage of using MFCC is producing robust and concise features that come out with high accuracy in the recognition process. The paper also presents complete overview of the mechanism of feature extraction derived from the phase of speech recognition.

## References

[1] E. S. Parris and M. J. Carey, “Language independent gender identification,” in Acoustics, Speech, and Signal Processing, 1996. ICASSP96. Conference Proceedings., 1996 IEEE International Conference on, vol. 2. IEEE, 1996, pp. 685–688.

[2] Fawaz S. Al-Anzi, Dia AbuZeina, “The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition ,” World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering,Vol:11, No:10, 2017

[3] K. Wu and D. G. Childers, “Gender recognition from speech. part I: Coarse analysis,” The journal of the Acoustical society of America, vol. 90, no. 4, pp. 1828–1840, 1991.

[4] Parwinder Pal Singh, Pushpa Rani, “An Approach to Extract Feature using MFCC,” OSR Journal of Engineering (IOSRJEN), vol. 04, Issue 08,pp. 21-25,Auguest, 2014.

[5] Pravin Bhaskar Ramteke , Amulya A. Dixit , Sujata Supanekar , Nagraj V. Dharwadkar , and Shashidhar G. Koolagudi , “ Gender Identification From Children’s Speech ”, Published in: Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India, pp. 2–4.

[6] S. Safavi, M. Russell, and P. Jancovic, “Identification of age-group from children’s speech by computers and humans,” in Fifteenth Annual Conference of the International Speech Communication Association, 2014, pp. 243–247.

[7] Sherry Vijh, Parminder Singh and Manjot Kaur Gill, “Feature extraction using MFCC for speech recognition”, Published in: Fourth International Conference on Recent Trends in Communication and Computer Networks - ComNet 2016.

[8] S. I. Levitan, T. Mishra, and S. Bangalore, “Automatic identification of gender from speech,” in Proceeding of Speech Prosody, 2016, pp. 84–88.

[9] Speech Recognition. (n.d.) In Wikipedia. Retrieved April 17, 2019 from https://en.wikipedia.org/wiki/Speech_recognition.

[10] Why is children’s Automatic Speech Recognition special? retrieved on April 17,2019 from http://www.italk2learn.eu/automatic-speech-recognition-childrens-speech-special.