Bilge International Journal of Science and Technology Research

Web : http://dergipark.gov.tr/bilgesci - **E-mail:** kutbilgescience@gmail.com

Received: 08.11.2017
Accepted: 06.03.2018
DOI: 10.30516/bilgesci.350016

ISSN: 2587-0742
e-ISSN: 2587-1749
2(1), 40-46, 2018

Real-time Parental Voice Recognition System For Persons Having Impaired Hearing

Mete Yağanoğlu[1*], Cemal Köse[2]

Abstract: Persons having impaired hearing do not live a comfortable life because they can't hear sounds when they are asleep or alone at home. In this study, a parental voice recognition system was proposed for those people. Persons having impaired hearing are informed by vibration about which one of their parents is speaking. By this means, the person having impaired hearing real timely perceives who is calling or who is speaking to him. The wearable device that we developed can real timely perceive parental voice very easily, and transmits it to person having impaired hearing, while he/she is asleep or at home. A wearable device has been developed for persons having impaired hearing to use easily at home environment. Our device is placed on user's back, and just a ring-sized vibration motor is attached to the finger of person. Our device consists of Raspberry Pi, usb sound card, microphone, power supply and vibration motor. First of all, the sound is received by a microphone, and sampling is made. According to the Nyquist Theorem, 44100 samples are made per second. Normalization during preprocessing phase, Mel Frequency Cepstral Coefficients (MFCC) during feature extraction stage, k nearest neighbor (knn) during the classification phase were used. Statistical or Z-score normalization was used in the pre-processing phase. By means of normalization of the data, it is ensured that each parameter in the training input set contributes equally to the prediction of the model. MFCC is one of the feature extraction methods that are frequently used in voice recognition applications. MFCC represents the short-time power spectrum of the audio signal, and models the manner of perception of human ear. Knn is an educational learning algorithm, and its aim is to classify the existing learning data when a new sampling arrives. The sound data received via microphone were estimated through preprocessing, feature extraction and classification stages, and the person having impaired hearing was informed through real time vibrations about to whom this voice belongs. This study was tested on 2 deaf, 3 normal hearing persons. The ears of normal hearing persons were covered with a earphone that gives out loud noise. Persons having impaired hearing estimated their mothers' voice by 76%, and fathers' voice by 81% accuracy in real-time tests. The success rate decreases due to the noise of environment especially while watching tv. In the tests performed while these persons are asleep, a person having impaired hearing perceives his/her mother's voice by 78%, and father's voice by 83% accuracy. In this study it was aimed for persons having impaired hearing to perceive their parents' voice and accordingly have a more prosperous standard of living.

Keywords: Wearable Processing, MFCC, Raspberry Pi, Vibration for Deaf.

1. Introduction

Engineers have been working on technological developments providing the relation between machines imitating human behaviors and humans for many years. Human-being wants discoveries that would understand him and make meeting his expenses easier with developed technology. One of the most remarkable discoveries is speech recognition applications. Speech is the most appropriate and effective way of communication for people. The technology of speech recognition aims

[1]Ataturk University, Faculty of Engineering, Department of Computer Engineering, 25240, Erzurum, Turkey.
[2]Karadeniz Technical University, Faculty of Engineering, Department of Computer Engineering, 61080, Trabzon, Turkey.
*Corresponding author: yaganoglu@atauni.edu.tr

Citation: Yaganoglu, M., Kose, C. (2018). Real-time Parental Voice Recognition System for Persons Having Impaired Hearing. Bilge International Journal of Science and Technology Research, 2 (1): 40-46.

at developing and producing systems that can get data by talk and move on data. At the same time, it aims at providing data not only from humans to machines but also from machines to humans through speech synthesis. Works on speech recognition can be considered part of a function of artificial intelligence machines that can get, understand talked messages, turn them into text, move according to these messages and complete data transmission by talking. Speech recognition systems are in the first place in research subjects providing this kind of developments and are the systems including the recognition of voices recorded by a microphone through computer programs. The problem of speech recognition has had a very large place and a lot of various systems have been developed from the past to today in the literature. Speech processing is a place of study many various signal processing techniques can be applied.

It is too complicated problem to rapidly realize speech recognition process that people can easily do in daily life through an algorithm on computer. By getting rid of this problem, an important step would be taken in realizing a capability belonging to people by means of machines. This way, an application, which can rapidly work in a simple embedded system, be used by disabled people for example and place of use under the control of a device, can be developed.

The most important benefit in speech recognition applications is to be able to design a system deaf and blind people can understand. In this kind of applications, it is purposed to make a disabled person's life better than normal. For example, it can be provided that a blind reader can read records gotten before at any speed (Schafer and Rabiner, 1975). In this study, a system that deaf people can sense father and mother voices is designed. A deaf person doesn't hear his parents talking to him and his standard of life decreases. For example, he can understand during sleeping when his parents talk to him especially in emergency.

In the study of Phoophuangpairoj, multiple Hidden Markov Model (HMM) voice recognition system is used. The used system is composed of the results of acoustic model, spelling dictionary, robot command grammar, recognition of robot command and voice recognition algorithms. An average success at 98% is obtained with the system developed by using three different user groups as gender-dependent, gender-independent and those whose genders are known (Phoophuangpairoj, 2011).

In the studies of Muda et al., various commands are separately tried for women and men by using Mel Frequency Cepstrum Coefficients (MFCC) and Dynamic Time Bending algorithms (Muda et al., 2010).

In the studies of Fezari and Salah, trials for voice recognition application are done under 4 different conditions. Average success at 85%, 73%, 78% and 65% is respectively obtained in the end of the trials as regular and irregular noisy inside and outside lab (Fezari and Salah, 2006).

In the studies of Babu et al., Linear Predictive Coding, Vector Quantization algorithms and HMM algorithm are separately used. In the developed system, environment noisy is determined and used as threshold value, numbers between 0 and 9 are firstly sensed through voice determination system independently from time and voice processing techniques are used. In various gain values, an average success at 80% is obtained (Babu et al., 2012).

In the studies of Leechor et al., while an average success at 98% is obtained noiseless environment a remote-control car is controlled by voice commands, it decreases up to 44% in a noisy room (Leechor et al., 2010).

In the studies of Jiang et al., MFCC and Chirp Z Transform algorithms are used. An average success at 86% is obtained in the study on 120 voice examples with the numbers between 0 and 9 (Jiang et al., 2009).

In their studies, Shearme and Leach develops an application providing the recognition of disjoint words voiced by any speaker. In their experiments, they express each one of words with a value sets obtained with specific intervals of the normalized spectrum and thus realize the recognition process by comparing the values for every one of words. They create a dictionary composed of 32 words and test their applications with 10 speakers. They determine recognition percentages as 90% (Shearme and Leach, 1968).

Chang et al. obtain 90% as the lowest success and 95% as the highest success for a speech recognition system which MFCC is used as feature and whose

education is realized with DTW (Chang et al., 2012).

Nandyala and Kumar realizes DTW algorithm whose education phase is very short compared to HMM, YSA and SVM with dynamic programming and a speaker-dependent Hindi disjoint speech recognition system which MFCCs are used as feature and suggest a success of recognition at 82% (Nandyala and Kumar, 2010).

2. Material and Method

While game consoles, computers, smart phones and tablets take place in the revolution of technology, smart and wearable devices are taking the first place today. Wearable devices particularly have an important place for blind and deaf people. Because of them, their life quality is tried to be increased. In this study, a wearable device is developed to make disabled people's usage. Our wearable device (see in Fig.1.) is placed on back of the user. The wearable device we develop is composed of Raspberry Pi, Usb sound card, Microphone, power supply and vibration motor.

Raspberry Pi (Rpi) is developed to improve computer-based education and teach computers to children at schools. It is at the size of a credit card and Linux, Android and Windows can be set up on it. Rpi is defined a mini computer with ARM architecture. Because of its low cost, power consumption and little size, Rpi's popularity has increased recently. You can get footage by connecting it onto TV and connect a keyboard. For example, you can do work you do on desktop computers and play various games working with Word processors and calculation programs by Rpi we call capable little computer (Balasubramaniyan and Manivannan, 2016).

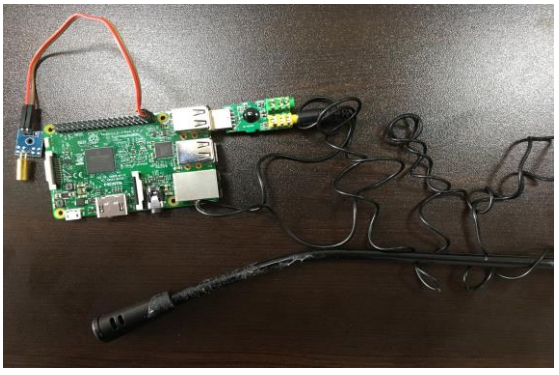

Figure 1. Our Wearable Device

Figure 2. Raspberry Pi

The Raspberry Pi 3 technical features are:

- 4× ARM Cortex-A53, 1.2GHz CPU
- 1GB RAM
- 10/100 Ethernet, 2.4GHz 802.11n wireless
- Bluetooth 4.1 Classic
- GPIO: 40-pin header
- HDMI, 3.5mm analogue audio-video jack, 4× USB 2.0, Ethernet, Camera Serial Interface (CSI), Display Serial Interface (DSI)

There should be Usb sound cart to connect microphone. Because there is no sound card on Rpi, Usb sound card is used to get sound. Microphone is plugged onto Usb sound card and used to sense sound. Sound is transmitted into Rpi by getting it by microphone. Thus, necessary processes are done for determined sound. As power supply, chargeable battery or usb power supply can be used. In this study, usb power supply is used. Vibration motor is placed on finger of the user. By giving different vibrations for mother and father voice with vibration motor, it is provided that users sense who talks.

The first step of speaker recognition is to obtain voiced expression. For this purpose, microphone or telephone is generally used. Speech signal obtained in this step is analogue. First, analogue signal in continuous time should turn into discrete time. The process of transformation into discrete time is done with sampling. Value of signal is measured at specific time intervals and this obtained value is called sample. If analogue signal has high frequency components, you should do sampling at higher rates not to lose data. Generally, we need sampling as much as double highest frequency not to lose data from analogue signal. This is known as Nyquist ratio. Otherwise, degradation occurs because of aliasing and original signal cannot be obtained

again. In this study, sound is gotten by microphone firstly and sampling is done. According to Nyquist Theorem, 44100 samplings are done per second. A general structure of sound recognition systems is given in Figure 3.

Figure 3. Speech Recognition System

2.1. Preprocessing

In the step of preprocessing, Statistical or Z-Score normalization is used. The fact that some values on the same data set are below 0 and some are bigger than it suggests that these distances between data and data at the initial or final point would be more effective on results. Because of data normalization, it is provided that every parameter on the input set equally contributes to prediction process of the model. By normalizing data, the distances between them are removed and extreme points on data are decreased (Yağanoğlu et al., 2014).

2.2. Feature Extraction

In the step of feature extraction, Mel Frequency Cepstrum Coefficients (MFCC) is used. MFCC features model human ear. Because it is obtained by using filter group, it shows a better success compared to farseeing coefficients in noisy sounds. MFCC is used in both speech recognition and speaker recognition applications and gives successful results. That's why, today's most widely used characteristic is vectors. The method of revealing feature taking human ear's hearing as example is one of the most widely used algorithms giving the highest success. In Figure 4, steps of deriving MFCCs are shown. MFCCs compose features by using mel scale taking human's hearing as model.

Characteristics of speech signals remain stable at a sufficiently little time interval. That's why, sound signals are processed at little time intervals (Schafer and Rabiner, 1975). Researches suggest that characteristics of sound signal remain stable at a sufficiently little time interval. Therefore, sound signals are processed at little time intervals. Signals are divided by frames with length varying from 20 to 100 milliseconds generally. Mostly, the most effective time interval is between 20 and 30 ms (Atal, 1976). Every frame includes N speech examples and M (M<N) examples of the previous neighboring frame. So, every frame covers a specific part of the previous frame. The goal of the method of covering is to provide that transition from a frame to another one would be soft

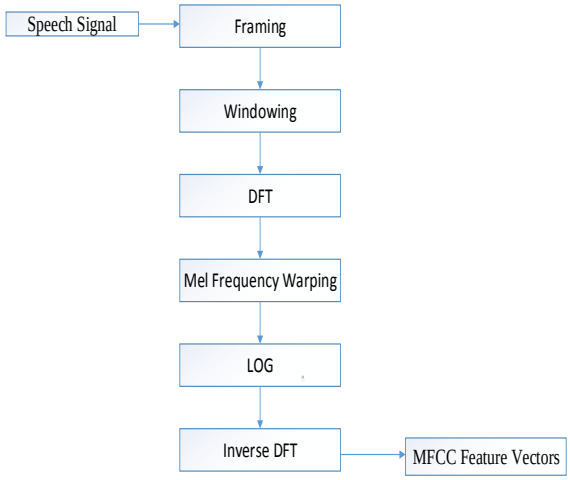

Figure 4. MFCC Steps

Windowing is the process done to prevent discontinuity in the beginning and end of every one o frames. Types of windowing are Hamming, Hanning, Rectangular, Barlett, Kaiser Mel Cepstrum Framing Windowing FFT Mel-frequency transform Kepstrum 9 and Blackman windows. The most widely used structure of window is Hamming. It is used to derive frequency components of windowed signal.

In the step of DFT, Fast Fourier Transform (FFT) is applied to transform every frame with N samples from the domain of time into of frequency. Methods allowing for rapid calculation of discrete fourier transform are fast fourier transform. Fast fourier transform provides that discrete fourier transform is widely used in signal processing application. Discrete fourier transform of a sign can be calculated the following equation. In that case, there should be N complicated multiplication and N-1 complicated addition operations for every k value of the transform.

According to Mel frequency skewness, width of triangle filters changes and therefore daily total energy in a critical band around central frequency is included. Numbers of coefficients are yielded after warp. In the end, inverse DFT is used to calculate cepstral coefficient (Bhattacharyya, 2015).

As the last step of revealing feature, every frame is exposed to inverse fourier transform and the domain of frequency is turned into the domain of time again. As a result of it, Mel-Frequency Cepstral Coefficients are yielded.

2.3. Classification

K Nearest Neighbor (Knn) is one of the mostly preferred machine learning and classification methods in different applications. In this preference, the main reason is simplicity of the method. The method of Knn was firstly recommended by Fix and Hodges as the method of nonparametric pattern classification in 1951. In the classification of Knn, the first step is the process of learning. In this process, data in the form of feature vectors is taught to classification system and it is determined what class it belongs to. The next step is testing. In the step of testing, distance metrics of test data to learning data are simply calculated. After calculating distance metrics, results are increasingly listed (Yağanoğlu and Köse, 2017).

Bozkurt et al. used Knn in their study and obtained the classification results with 10-fold cross validation (Bozkurt et al., 2018). In our study, Leave-one-out cross validation was used in order to select the most suitable k value. Knn is a trained learning algorithm and its goal is to classify on the existing learning data when a new sample comes. The algorithm decides class of the example by considering a new sample's nearest k neighbor when it comes. In the algorithm of Knn, k value should be determined firstly. The best k value becomes determined. k = 5 value, having the best rate, is selected. After determining k value, its distance between all learning samples should be calculated. Then, the process of listing with respect to minimum distance is done. After the process of listing, what class value it belongs to is found.

3. Results

This study has been tested on 2 deaf and 3 normally hearing people for 30 days. Ears of normally hearing people are covered by loud sound earphones. Our device mounted on backs of the people gives data to deaf person. Thus, deaf person could very comfortably predict parent's voice. While sleeping, our device put with his side warns of the person when he hears parent's voice.

In Figure 5, success rates of deaf and normally hearing people in home environment are seen. While deaf people's rate of prediction of mother's voice is 74% on average, normally hearing people's rate of prediction of mother's voice is 79% on average. While deaf people's rate of prediction of father's voice is 78% on average, normally hearing people's rate of prediction of father's voice is 83% on average. Considering general average, mother's and father's voices are respectively predicted at 76% and 81% in home environment.

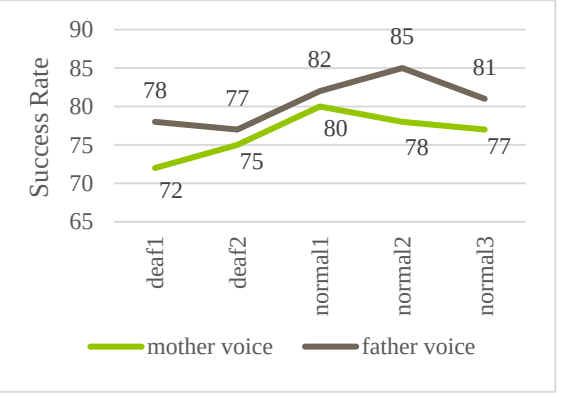

Figure 5. Success rates of deaf and normally hearing people in home environment

In Figure 6, success rates of deaf and normally hearing people while sleeping are seen. While deaf people's rate of prediction of mother's voice is 76% on average, normally hearing people's rate of prediction of mother's voice is 80% on average. While deaf people's rate of prediction of father's voice is 79% on average, normally hearing people's rate of prediction of father's voice is 83% on average. Considering general average, mother's and father's voices while sleeping are respectively predicted at 78% and 83%.

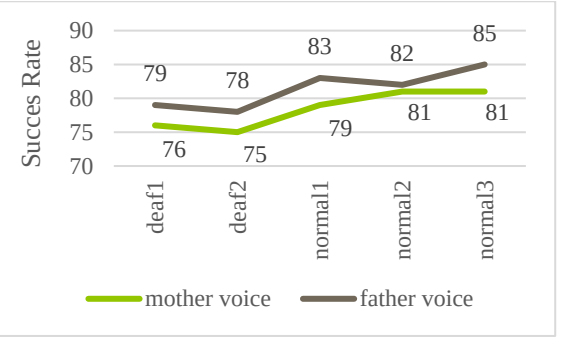

Figure 6. Success rates of deaf and normally hearing people while sleeping

Only success rates of deaf people are like in Figure 7. Deaf people respectively predicted mother's voice at 74% and 76% in home environment. While sleeping, they respectively predicted father's voice at 78% and 79%.

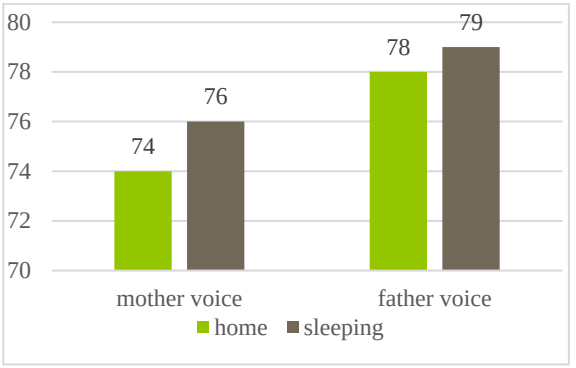

Figure 7. Success rates of deaf people

4. Discussion and Conclusions

Telling the device to deaf people and educational process take time. Understandability in telling the device to them and first usages decreased. However, the user could better sense after using the device.

Because of noise in home environment, lower rates of success are obtained. Especially due to such sounds as TV and telephone sounds, success rates become lower. However, success rates increase while sleeping because there is no a lot of noise.

In this study, normalization in the step of pre-processing, MFCC in the step of feature extraction, Knn for classification are used. In the next studies, success rate would be tried to be increased by other methods of revealing feature and classification.

In this study, wearable device is designed to increase life quality of deaf people. Deaf people could sense who talks from parent's voice. Particularly in emergency, deaf people can sense that they are called when somebody calls them or while sleeping.

References

Atal, B.S. (1976). Automatic recognition of speakers from their voices, Proceedings of the IEEE, vol. 64, pp. 460-475, 1976.

Bhattacharyya, S. (2015). Handbook of Research on Advanced Hybrid Intelligent Techniques and Applications, IGI Global.

Babu, C.G., Kumar, R.H., Vanathi, P. (2012). Performance analysis of hybrid robust automatic speech recognition system, in Signal Processing, Computing and Control (ISPCC), 2012 IEEE International Conference on, 2012, pp. 1-4.

Balasubramaniyan, C., Manivannan, D. (2016). IoT Enabled Air Quality Monitoring System (AQMS) using Raspberry Pi, Indian Journal of Science and Technology, vol. 9.

Bozkurt, F., Köse, C., Sarı, A. (2018). An inverse approach for automatic segmentation of carotid and vertebral arteries in CTA. Expert Systems with Applications, 93, 358-375.

Chang C.-H., Zhou, Z.-H, Lin, S.-H, Wang, J.-C. And Wang J.-F. (2012). Intelligent appliance control using a low-cost embedded speech recognizer, in Computing and Networking Technology (ICCNT), 2012 8th International Conference on, 2012, pp. 311-314.

Fezari, M., Bousbia-Salah, M. (2006). A voice command system for autonomous robots guidance," in Advanced Motion Control, 9th IEEE International Workshop on, 2006, pp. 261-265.

Jiang, Z., Huang, H., Yang, S., Lu, S., Hao, Z. (2009). Acoustic feature comparison of MFCC and CZT-based cepstrum for speech recognition, in Natural Computation, 2009. ICNC'09. Fifth International Conference on, 2009, pp. 55-59.

Leechor, P., Pompanomchai, C., Sukklay, P. (2010). Operation of a radio-controlled car by voice commands, Mechanical and Electronics Engineering (ICMEE), pp. V1-14-V1-17.

Muda, L., Begram, M., Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, Journal of Computing, Volume 2, Issue 3.

Nandyala, S.P., Kumar D.T.K. (2010). Real Time Isolated Word Speech Recognition System for Human Computer Interaction, International Journal of Computer Applications, vol. 12, pp. 1-7.

Phoophuangpairoj, R. (2011). Using multiple HMM recognizers and the maximum accuracy method to improve voice-controlled robots, Intelligent Signal Processing and Communications Systems (ISPACS), pp. 1-6.

Schafer, R.W., Rabiner, L.R. (1975). Digital representations of speech signals, Proceedings of the IEEE, vol. 63, pp. 662-677.

Shearme, J, Leach, P. (1968). Some experiments with a simple word recognition system, IEEE Transactions on Audio and), 1427-1432. Electroacoustics, vol. 16, pp. 256-261.

Yağanoğlu, M, Bozkurt, F., Günay, F.B. (2014). Feature Extraction Methods for EEG Based Brain-Computer Interface Systems, Journal of Engineering Sciences and Design, vol. 2, pp. 313-318.

Yağanoğlu, M., Köse, C. (2017). Wearable Vibration Based Computer Interaction and Communication System for Deaf. Appl. Sci. 2017, 7, 1296.