# Analysis of Speech recognition

Ms. Mehak Mehraj[1], Ms. Sukhvinder Kaur[2], Muheet Ahmed Butt[3], Majid Zaman[4]
*[1]M. Tech Student, Department of Electronics and Communication, Swami Devi Dyal Inst. of Engg. & Technology, Kurukshetra University, Kurukshetra*
*[2]Assistant Professor and Head of Department of Electronics and Communication, Swami Devi Dyal Inst. of Engg. & Technology, Kurukshetra University, Kurukshetra*
*[3]Scientist, PG Department of Computer Sciences, University of Kashmir, Srinagar*
*[4]Scientist, Directorate of IT&SS. University of Kashmir, Srinagar*

***Abstract-*** The aim of this work is to investigate and analyze the algorithms of speech recognition. The proposed algorithm is programmed and simulated in MATLAB. In this work, two systems are designed. First system is depends on the information of shape of cross-correlation plotting. The second system is used to realize the speech recognition by using Weiner Filter. In the simulation, microphone is used in order to record the speaking words. This approach may provide 100% successful results if the speaker is the same person for three time recordings. Thus, the designed systems actually work well for the basic speech recognition.

***Keywords-*** *Speech recognition, Cross-correlation, Wiener Filter, Simulation.*

## I. INTRODUCTION

Speech recognition has become a house hold application. Modern electronic gadgets are equipped with speech recognition devices. Internet is flooded with audio data and software for speech detection and recognition. Instead of typing with the keyboard or operating with buttons, using speech makes it more convenient to operate electronic systems. Voice recognition programs are available which makes our life far better. Nowadays, this recognition system has numerous applications which requires interface such as automatic call processing, query-based information systems, weather reports etc. The speech recognition systems develop the efficacy of daily life and also get better the lives of human in diversified manner. Speech/voice recognition is one of the many available biometric recognition schemes. The past decade has seen dramatic progress in voice recognition technology, to the extent that systems and high-performance algorithms have become accessible. The efficiency of the daily life not only enhances, but also makes people's life more diversified.

According to Speech recognition is the technology that makes it possible for a computer to identify the components of human speech. The process can be said to begin with a spoken utterance being captured by a microphone and to end with the recognized words being output by the system. Speech is technically defined as a sequence of basic units called phonemes [5]. Automated Speech Recognition (ASR) systems convert analog speech signals received through microphones to digital signals that are segmented to retrieve phonemes. Using the phoneme sequence, the ASR system refers to the vocabulary and grammar rules to decipher words or phrases. It unlocks world of potential for developers; especially those building IVRs and other telephony applications, but speech recognition also has some challenges. User should articulate to the desktop rather than pushing buttons or interrelate with screen of desktop. This means uncertainty is linked with their input, as automatic speech recognition only returns probabilities, not certainties. Before discussing the many ways speech recognition is useful, it is important to consider its unique strengths and weaknesses. The most obvious weakness is the one mentioned above, namely the potential for misrecognition. There will always be times when the application misrecognizes user input. Because of this, it becomes important to provide for greater error handling than in other applications. It is significant to verify what user said, if the score on recognition system is low. Sometimes, user has to repeat them. In some cases, if speech engine gives less value for same users many times then it can be significant to send that user to a human operator so the user can conduct his or her transaction.

The speech signal conveys the information regarding the words or message being spoken along with the identity of the speaker. For the purpose of speaker identification speech signalis represented in terms of certain features. These features are grouped into feature vectors thatserve the purpose of reducing dimensionality and redundancy in the input to the speaker identification system, while retaining the speaker-specific information. As the presence of irrelevant information with regards to speaker discrimination is a common problem for all feature sets,it is the topic of ongoing research that strives to determine feature sets of reduced complexitythat can be applied to speaker identification. The exact nature of the feature set dependson what part of a speech signal the features are expected to represent and thus what type of information is to be extracted. This is why feature sets can be grouped as being source basedfeatures or system based features. The source is described as being the actualsound wave that is transmitted from the diaphragm through the glottis and so these feature concerned with determining the characteristics of the vocal cords, where this waveform isshaped. The most feasible parameter that can be determined is fundamental frequency. The system characteristics can be extracted for the vocal tract, the nasal cavity and the lip radiation. These features model the filter characteristics of the vocal tract that can be derived from information contained in voiced and unvoiced speech. This information includes the formant frequencies that are predominantly present in vowels. The system features reflect the physiology of the speaker. For each feature extraction method, it is therefore necessary to know exactly what is being extracted so as to avoid imprecision and ambiguity. As phase information in a speech signal is not significant for discrimination between speakers, it can be omitted In order to simplify calculations, i.e., the magnitude of the spectrum of the speech signal is used.

## II. LITERATURE SURVEY

Review of literature on speech recognition systems authentically demands consideration towards the finding of Alexander Graham Bell regarding the method of converting sound waves into electrical impulses and the first speech recognition system developed by Davis et al. [6] for finding telephone superiority digits spoken at normal speech rate. This attempt for automatic speech recognition was mainly centered on the edifice of *an electronic circuit* for discovering ten digits of telephone superiority. Spoken words were examined to obtain a 2-D plot of formant 1 vs. formant 2. A circuit is developed for finding the greatest correlation coefficient among a set of novel incoming data for pattern matching. These features are grouped into feature vectors that serve the purpose of reducing dimensionality and redundancy in the input to the speaker recognition system. An indication circuit was created to exhibit the already discovered spoken digit. The proposed approach lays stress on recognizing speech sounds and providing appropriate labels to these sounds. Various approaches and types of speech recognition systems came into existence in last five decades gradually. This evolution has led to a remarkable impact on the development of speech recognition systems for various languages worldwide. The exact nature of the feature set depends on what part of a speech signal the features are expected to represent and thus what type of information is to be extracted. In the speech to text conversion system, the output of the system shows the text which is used to recognized the speech. Languages, on which automatic speech recognition system has been designed, is a portion of total around 7300 existing languages which are Hindi, English, Tamil, Bengali, Russian, Japanese, Portuguese, Sinhala, Chinese, Malayalam, Vietnamese, Spanish, Arabic, Filipino, Hindi are well-known among them. English is the language for which maximum work for recognition is done. Since 1930s, a simple speech machine that responds to a limited small set of words was invented. This proposed machine is capable to take actions on spoken words and create the speech. From that time, it becomes popular area of research to invent speech recognition system. The best example for this is done by Olson and Belar in 1950 in RCA Laboratories who build a system to identify 10 syllables of a single talker (Olson et al., 1956) and at MIT Lincoln Lab, Forgie and Forgie built a speaker-independent 10-vowel recognizer (Forgie et al., 1956). It is continued by the middle of 70's. The new system of speech recognition depends on LPC methods, were proposed by Itakura, Rabiner and Levinson (Itakura 1975; Rabiner et al., 1979) and others. This investigation brings main benefits where research shift the methodology from the more intuitive template-based approach towards a more precise statistical modeling outline (Juang et al., 2004) in 1980s.

## III. PROPOSED METHODOLOGY

In this work, there are two designed systems for speech recognition. Both of these two systems utilized the knowledge according to the theory part of this thesis which has been introduced previously. The author invited his friends to help to test two designed systems. For running the system codes at each time in MATLAB, MATLAB will ask the operator to record the speech signals for three times. The first two recordings are used as reference signals. The third time recording is used as the target signal.

### *A. Algorithm for Design System 1:*

1. Initialize the variables and set the sampling frequency.
2. Process recorded signals and get returned matrix signals.
3. Get the frequency spectrum by transposing the input signal.
4. Normalize the signal by Linear Normalization.
5. Do the cross-correlation for the third recorded signal with the first two recorded reference signals separately.
6. Check the frequency shift of the cross-correlations.
7. Continuously do the comparison by the symmetric property for the cross-correlations of the matched signals.

### *B. Algorithm for Design System 2:*

1. Initialize the variables and set the sampling frequency.
2. Record 3 voice signals. Make the first two recordings as the reference signals. Make the third voice recording as the target voice signal.
3. Process recorded signals and get returned matrix signals.
4. Get the frequency spectrum by transposing the input signal.
5. Normalize the frequency spectrums by the linear normalization.
6. Calculate the auto-correlations of 3 signals:
7. Calculate the filter coefficients by wiener filter mode.
8. Calculate the minimum mean square-error for each reference signal.
9. The better estimation should have the smaller minimum mean square-errors.

## IV. RESULTS

The information of the first statistical simulation results for system 1 is as following:

Reference signals: “on” and “off”:

Target signal: From time 1 to time 10, “on”.

From time 11 to time 20, “off”.

Speaker: Speaker 1 for both reference signals and the target signal.
Environment around: almost no noise
The figure 4.1 is about frequency spectrums for three recorded signals, but the axis is not the real frequency axis since the figure is got by STFT. The information of STFT can be found in the part in chapter 2.

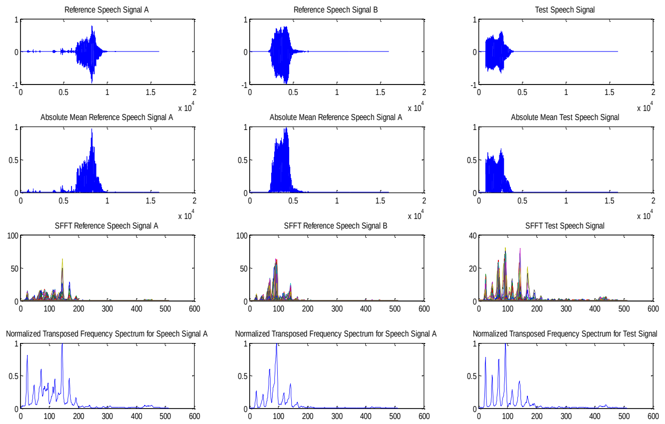


Fig.1: Frequency spectrums for three speech signals: “on”, “off”, “on”

The figure of cross-correlations between the target signal “on” and the reference signals is as below (the reference signal of the left plotting is “on”; the reference signal of the right plotting is “off”):

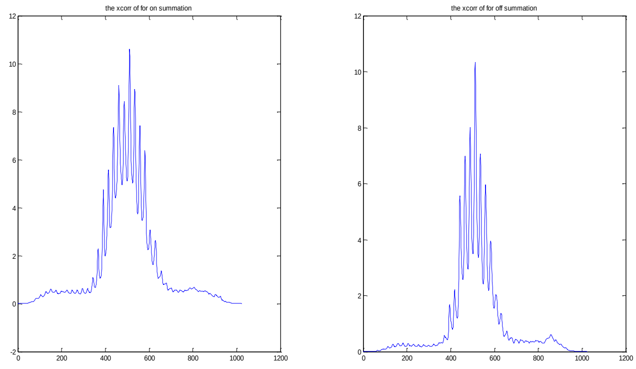


Fig.2: Cross-correlations between the target signal “on” and reference signals

In Fig. 2 shown above, there is no big difference between two graphs, since the pronunciations of “on” and “off” are close.

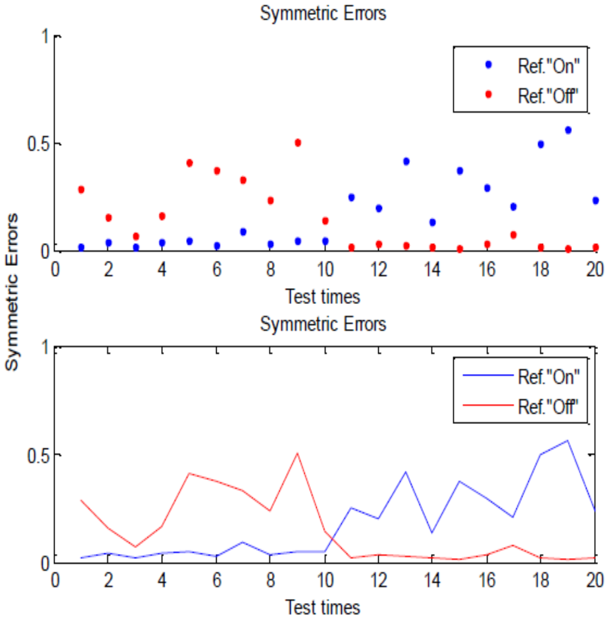


Fig.3: Symmetric errors in 20 times simulations for reference “on” and “off”

As shown in Fig. 3, the blue curve is simulated result when the reference speech word is “on”. The red curve is the simulated result when the reference speech word is “off”. As information given at the beginning, the target speech word is “on” in the first 10 times simulations and the target speech word is “off” in the second 10 times simulations. From Fig. 4.4, it is shown that in the first 10 times simulations the reference “on” curve has lower value and in the second 10 times the reference “off” curve has lower value. The results have shown that when the reference speech signal and the target speech signal are matched, the symmetric errors are smaller. The judgments are totally correct.

The information of the second statistical simulation results for system 1 is as following:

Reference signals: “Door” and “Key”:

Target signal: From time 1 to time 10, “Door”.

From time 11 to time 20, “Key”.

Speaker: Speaker 1 for both reference signals and the target signal. Environment around: almost no noise. The figure 4 about frequency spectrums for three recorded signals is got by the same way as the figure 4.1.

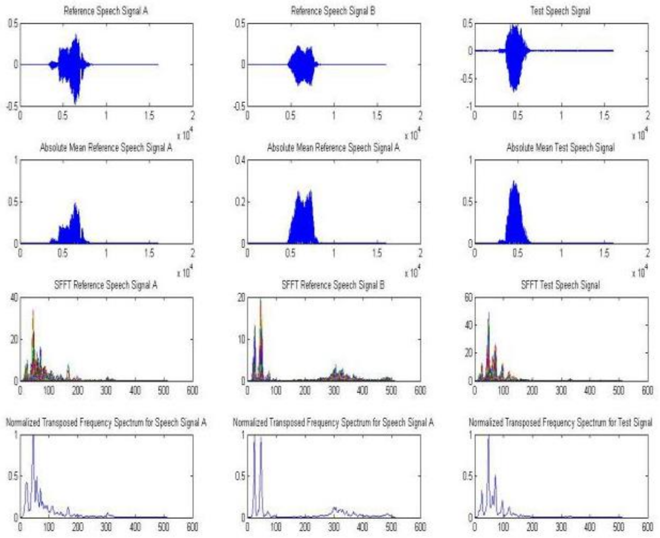


Fig.4: Frequency spectrums for three signals: “Door”, “Key”, and “Door”

The figure of cross-correlations for the target signal “Door” with reference signals is as below (the reference signal of the left plotting is “Door”; the reference signal of the right plotting is “Key”)

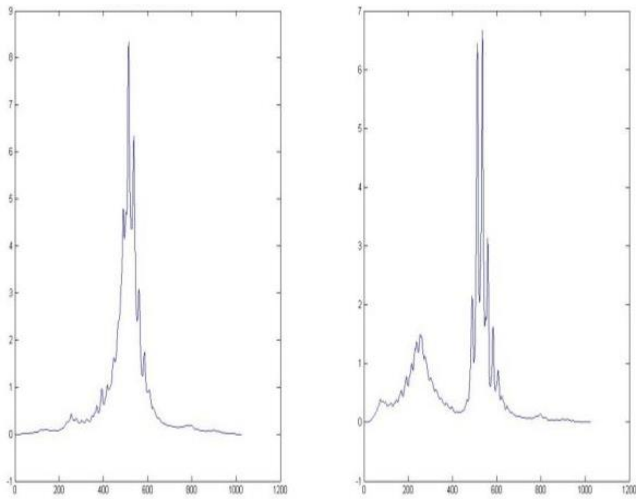


Fig.5: Cross-correlations for the target signal “Door” with reference signals

As Fig.5 shown above, there is large difference between two graphs. Since the pronunciations of “Door” and “Key” are totally different.

As introduced in theory part, the better matched signals have better symmetric property of the cross-correlation. The Fig.5 approved this point.

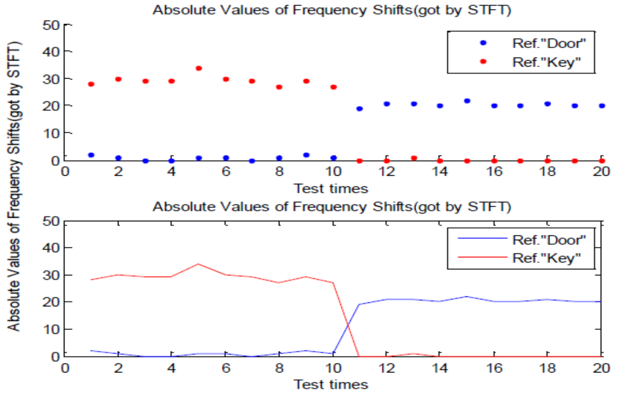


Fig.6: Frequency shits in 20 times simulations for reference “Door” and “Key”

From Fig.6, it can be seen that the frequency shifts have large differences. So the designed system will directly give the judgments according to the frequency shifts.

The information of the third statistical simulation results for system 1 is as following:

Reference signals: “on” and “off”:

Target signal: From time 1 to time 10, “on”.

From time 11 to time 20, “off”.

Speaker: Speaker 2 for both reference signals and the target signal.

Environment around: there is some noise sometimes

Since “on” and “off” have small frequency shifts’ difference, so the designed system will only give the judgments with symmetric errors. The plotted simulation result is as below:

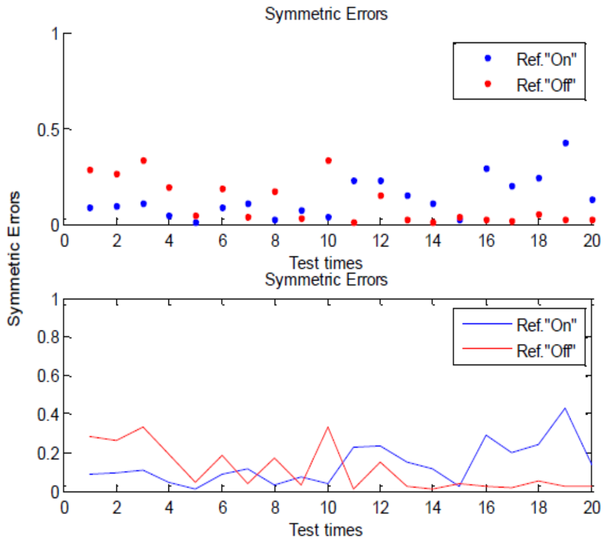


Fig.7: Symmetric errors in 20 times simulations for reference “on” and “off” (noisy)

(4) The information of the fourth statistical simulation results for system 1 is as following:

Reference signals: “Door” and “Key”:

Target signal: From time 1 to time 10, “Door”.

From time 11 to time 20, “Key”.

Speaker: Speaker 2 for both reference signals and the target signal.

Environment around: there is still some noise sometimes

The plotted simulation result is as below:

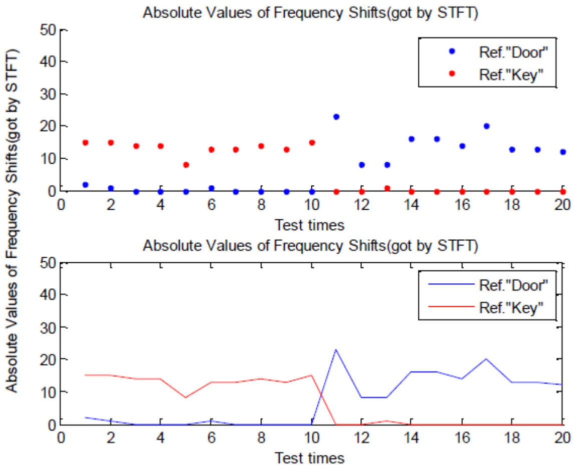
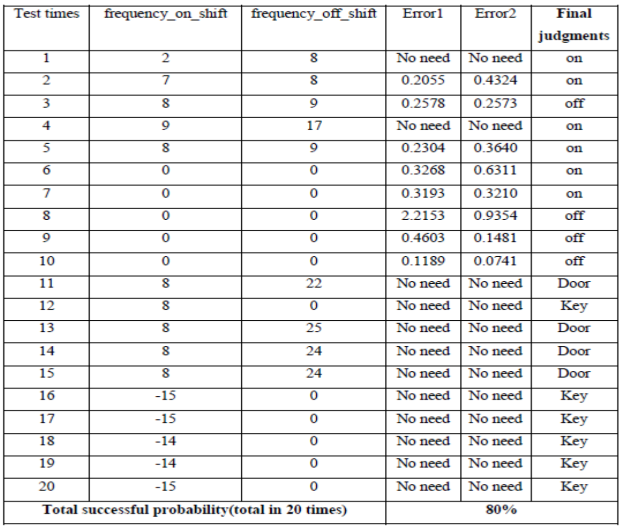


Fig.8: Frequency shits in 20 times simulations for reference “Door” and “Key” (noisy)

Table 1 indicates the simulation results for reference signals “Door” and “Key” as the information given at the beginning of this section.

Table 1: Simulation results for speech words “On”, “Off”, “Door” and “Key”



| Test times | frequency_on_shift | frequency_off_shift | Error1 | Error2 | Final judgments |
|---|---|---|---|---|---|
| 1 | 2 | 8 | No need | No need | on |
| 2 | 7 | 8 | 0.2055 | 0.4324 | on |
| 3 | 8 | 9 | 0.2578 | 0.2573 | off |
| 4 | 9 | 17 | No need | No need | on |
| 5 | 8 | 9 | 0.2304 | 0.3640 | on |
| 6 | 0 | 0 | 0.3268 | 0.6311 | on |
| 7 | 0 | 0 | 0.3193 | 0.3210 | on |
| 8 | 0 | 0 | 2.2153 | 0.9354 | off |
| 9 | 0 | 0 | 0.4603 | 0.1481 | off |
| 10 | 0 | 0 | 0.1189 | 0.0741 | off |
| 11 | 8 | 22 | No need | No need | Door |
| 12 | 8 | 0 | No need | No need | Key |
| 13 | 8 | 25 | No need | No need | Door |
| 14 | 8 | 24 | No need | No need | Door |
| 15 | 8 | 24 | No need | No need | Door |
| 16 | -15 | 0 | No need | No need | Key |
| 17 | -15 | 0 | No need | No need | Key |
| 18 | -14 | 0 | No need | No need | Key |
| 19 | -14 | 0 | No need | No need | Key |
| 20 | -15 | 0 | No need | No need | Key |
| **Total successful probability(total in 20 times)** | | | **80%** | | |

Since the simulation results are not good as the expected values. So only the table results are shown here.

## V. CONCLUSION

For general conclusions, the designed systems for speech recognition are easily disturbed by the noise, which can be observed from Table 1. For the designed system 1, the better matched signals have the better symmetric property of their cross-correlation. For the designed system 2, if the reference signal is the same word as the target signal, so using this reference signal to model the target will have less error. This conclusion can be proved by the all the simulation results for the designed system 2. When both reference signals and the target signal are recorded by the same person, two systems work well for distinguishing different words, no matter where this person is from. But if the reference signals and the target signal are recorded by the different people, both systems don’t work that well. So in order to improve the designed systems to make it work better, the further tasks are to enhance the systems’ noise immunity and to find the common characteristics of the speech for the different people. Otherwise, to design some analog and digital filters for processing the input signals can reduce the effects of the input noise and to establish the large data base of the speech signals for different words. And studying more advanced algorithms for signal modeling can give a lot of help to realize the better speech recognition.

## VI. REFERENCES

[1]. B.H. Juang& Lawrence R. Rabiner, Automatic Speech Recognition – A Brief History of the Technology Development, 2004

[2]. Deepak, M. Vikas, “Speech Recognition using FIR Wiener Filter”, International Journal of Application or Innovation in Engineering & management (IJAIEM),pp.204-20,2013.

[3]. Christine Englund, Speech recognition in the JAS 39 Gripen aircraft - adaptation to speech at different G-loads.

[4]. SushilPhadke,” The Importance of a Biometric Authentication System”, The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 4, September-October 2013.

[5]. Speech Recognition Technology & Patent Landscape, iRunway 2015.

[6]. Davis, K., Biddulph, R., and Balashek, S., “Automatic Recognition of Spoken Digit,” J. Acoust. Soc. Am. 24: Nov 1952, p. 637.

[7]. S. Furui, .Digital speech processing,. Synthesis and Recognition, New York, Marcel Dekker, 2001.

[8]. Malcolm Slaney, .Auditory Toolbox,.Version 2, Technical report, Interval Research Corporation, 1998.

[9]. John G.Proakis, DimitrisG.Manolakis, Digital Signal Processing, Principles, Algorithms, and Applications, 4th edition,Pearson Education inc., Upper Saddle River.

[10].John Wiley, Sons,Inc. Statistical Signal Processing And Modeling, Monson H Hayes,Georigia I nstitute of Technology.

[11]. Luis Buera, Antonio Miguel, Eduardo Lleida, Oscar Saz, Alfonso Oretega, "Robust Speech Recognition with On-line Unsupervised Acoustic Feature Compensation", Communication Technologies Group (GTC),13A, University of Zaragoza, Spain.

[12]. HartmutTraunmüller , Anders Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults", Institutionenförlingvistik, Stockholmsuniversitet, S-106 91 Stockholm, Sweden.

[13]. Jian Chen, JiwanGupta,"Estimation of shift parameter of headway distributions using crosscorrelation function method", Department of Civil Engineering, The University of Toledo.

[14]. Buera, Luis, et al. "Cepstral vector normalization based on stereo data for robust speech recognition."*Audio, Speech, and Language Processing, IEEE Transactions on* 15.3 (2007),pp, 1098-1113,2007.