Person Identification Based on Humming Using MFCC and Correlation Concept

Nikunj Patel[1], Prof. Vinesh Kapadia [2], Prof. Kinnal Dhameliya [3], Dr. Ninad Bhatt [4]

[1] M.E-Research student, [2] Associate Professor, E & C Department, GTU, SNPIT & RC, Bardoli, Gujarat, India
[3] Assistant Professor, E & C Department, UTU, CGPIT, Bardoli, Gujarat, India
[4] Associate Professor, E & C Department, GTU, CKPCET, Surat, Gujarat, India

***Abstract*—In this paper, an attempt is made to identify persons with the help of their hum. Normally, speech is used as an input to thebiometric system for recognition of a person, but here,instead of using speech, hum of a particular person is used for the sametask. Hum is a sound that is emerged from the nose, in which themouth is completely closed and vocal tract is directly coupled to nasal cavity. Humming is also applicable for a deaf personas well as an infant person who is not able to speak.Humming is not produced as the same way as production of normal speech, so in this paper speaker identification is referred to as person identification.Here, Mel Frequency CepstralCoefficients (MFCC) is used as a feature extraction technique and Pearson Correlation concept is used to identify the person. Identification rate is measured as a performance parameter of the system.**

***Keywords*— Person Identification, Humming, MFCC, Pearson Correlation, Identification rate**

I. Introduction

Speaker Recognition is the process through which one can identify or verify a person's identity from his or her speech.Since humming is not the same process as the process of production of normal speech, so speaker recognition is referred to as person recognition. Speaker recognition is further divided in to Speaker Identification and Speaker Verification. Speaker Identification is the process which gives an output that, which registered speaker provides a given utterance from a set of known speaker and in that, number of decision alternative is equal to number of the classes. Speaker Verification deals with the process of accepting or rejecting the identity claim of a speaker and it provides two choices, accept or reject, regardless of the number of classes.Hum is a sound produced by nasal cavity, in which mouth is completely closed and the hum sound is emerged from the nose. Some of the facts about humming are as follows:

- The results obtained from the humming pattern of humans vary on mathematical scale. For each person, the patterns of hum signal, formant contour, pitch contour and spectrogram are distinct [1].
- Humming is naturally available and that too for every living beings let it be disordered or an infant. It is universally available on everyone than speech. Pitch is a fundamental frequency arising due to vibration of vocal tract. Pitch is also there during hum of a signal which is not very much useful for identification and verification of a person. It is very much effective for humming based music information retrieval [2].
- During production of humming, nasal cavity remains steady and it is known to be person-specific. Particular feature of nasal sound could also be very useful to recognize a person under forensic conditions [3].
- While producing thehum sound, oral cavity remains closeand vocal tract is coupled to nasal cavity. In most of the time, many people keep their mouth open while singing or humming, and it is difficult to produce hum without significant connection of nasal cavity and oral cavity. Sound which is coming out from nose has more inter-person variability and less intra-person variability [4].
- Velar movements generally occurs out of control of the brain for some subject so feature of nasal could also be useful to identify a person under global security conditions [5].
- Recent studies based on morphological and acoustic data proved that there are considerable acoustic contributions (in terms of anti-resonances) of different paranasal sinus cavities, viz., maxillary, sphenoidal, frontal, and ethmoidal to the transmission characteristic of the nasal tract. Thus, it is difficult to manipulate deliberately nasal spectra and hence features extracted from humming sounds could serve as a strong biometric signature [5].
- When there is a case of person-dependent telephone dialling, person authentication can be done using humming rather than normal speech. In this application, detection of a speaker is carried out by exploiting person-specific information through humming pattern of a person. So the approach can be based on extracting features which distinguish pattern of humming which is specific to a person [6].
- Humming must require significant connection of nasal cavity and oral cavity. So during humming, it is always difficult to ask the person to hum for a long duration because after some time breathing may be recorded so it can interrupt during recording [7]

The rest of this paper is organized as follow: Section II provides overview of different features and classifiers those have been used for humming based recognition system. Section III describes about Mel Frequency Cepstral Coefficient (MFCC) and Pearson Correlation concept. Section

IV demonstrates parameter specification required for experiment and result of the experiment with different sampling frequency and different number of MFCC features and section V ends with the conclusion based on experiment result.

II. BACKGROUND STUDY

Speaker recognition system includes three basic steps: Pre-processing, Feature extraction and Classifier. Pre-processing part includes sampling, frames, windows, start point-end point detection, and so on. Ultimately, it is a data compression processing, in that useful information are selected from signal data available [8].

A. *Feature Extraction*

Feature extraction, is a process of extracting best parametric representation of signals in order to produce a better recognition performance. Different features which have been used for humming based recognition system are as follow:

*1) LP based features:*This feature set includes LPC and Linear prediction cepstral coefficient (LPCC). The optimal value of the LPC are found out by using the least square error formulation of the linear prediction. Cepstral coefficients can be calculated from the LPC via a set of recursive procedure and which isknown as LPCC. LPC and LPCC both gives speaker specific information but LPCC gives better result than LPC [1].

*2) Perceptual Linear Predictive Coding (PLP):*In PLP, speech auditory spectrum is obtained, approximation of auditory spectrumis done by all-pole zero model and optionally transform the model spectrum back to the original amplitude-frequency domain. The order of PLP model is half of LP model which allows computational and storage saving [9]-[10].

3) Pitch: It is the fundamental frequency arising due to the vibration of vocal tract. The pitch information from humming is not conductive to human verification and identification. Pitch information is to be effective for humming-based music retrieval. However, pitch contained in humming is highly dependent on the melody and not on the person who is humming so the pitch information in hum is not that much useful for person identification and verification task [2].

*4) MFCC:*It is a static feature which is developed to mimic human perception process of hearing [1]. MFCC is the most axiomatic and popular feature extraction technique for speech recognition. In MFCC, frequency bands are placed logarithmically so it approximates the human system response more closely than any other system response. As a result of MFCC, there is a collection of coefficients which are called acoustic vectors. So each input utterance is transformed into a sequence of acoustic vector [11].

*5) Variable length teager energy based MFCC:*Teager energy operator (TEO) gives the result that how much energy is required to generate the signal. This features uses a TEO, a non-linear energy-tracking operator, for analysis of signals. Variable length teager operator (VTEO) is used to capture airflow properties in vocal tract which is responsible for another kind of excitation for both oral as well as nasal cavity [12].VTMFCC is obtained by passing the Speech signal through pre -processing stage which includes pre-emphasis, frame blocking, windowing. Next the VTEO of the pre-processed signal is calculated and the magnitude spectrum of the VTEO output is computed and warped to Mel frequency scale which is followed by usual log and DCT computation [3].

*6) Temporal and Spectral Features:*Zero crossing rate (ZCR) and Short time energy (STE) are used as temporal features which are extracted from the hum signal. MFCC, Spectral centroid (SC) and Spectral Flux (SF) are extracted as spectral features from a hum signal [4].

*7) Magnitude and Phase Information features via TEO:*Here magnitude and phase spectrum information of humming signal is captured during mel-filtering in frequency-domain and then mel filtered signal is converted back into time-domain to preserve phase information inherently. From the result, again VTEO is calculated from the sub band signals in time domain. This newly derived feature set is called as Mel frequency cepstral coefficients to capture magnitude and phase spectrum information via VTEO (MFCC-VTMP) [5].

*8) Static and Dynamic:*Static features are extracted from pre-processed frame and over which, the humming signal is stationary.Dynamic features are extracted from multiple frames, typically adjacent to the current frame. MFCC and VTMFCC are extracted as a static features and Δ cepstral, Shifted Delta Cepstral (SDC) are extracted as a dynamic features. Δ Cepstral and SDC conveys more meaningful temporal information hidden in the sequence of samples of humming signal. One SDC feature means it contains many Δ cep in its calculation and these are accumulated over longer time during humming [6].

B. *Classifier*

The classifier takes the features which are the results of the feature extraction technique and then performs either template matching or probabilistic likelihood computation on the features, depending on the type of algorithm employed [8]. In most of the cases Polynomial classifier is used as a basic classifier because these are the universal approximators to the optimal Bayes classifier. It is processed by the Polynomial discriminant function. As the order of classifier is increased, computation time and memory requirement increases severely and hence 2nd order polynomial classifier is normally used as the basis in most of the cases [13]-[14]. Other classifier, Gaussian Mixture Model Universal background Model is also used for humming based speaker recognition. For Human Verification, GMM is based on the hypothesis and for the hypothesis, GMMUBM is used. The UBM is trained on humming from various people not included in the target and impostor list. The model of target person is adapted from the

UBM using enrolment data of in the maximum a posterior sense and accordingly accept/reject decision is made [2].

As from the study of all literatures and results included in that, it is found that MFCC and combination of MFCC with different feature gives good success rate as well as better equal error rate. Here, MFCC and Pearson Correlation concept theory is used for the humming based person recognition task which is explained in detail in the next section.

III. Mel Frequency Cepstral Coefficient and Pearson Correlation Concept

A. *Mel Frequency Cepstral Coefficient (MFCC)*

MFCC is the most axiomatic and popular feature extraction technique for speech recognition. It approximates the human system response more closely than any other system because frequency bands are placed logarithmically here [11]. The overall stepwise process of the MFCC is described and shown in below figure.

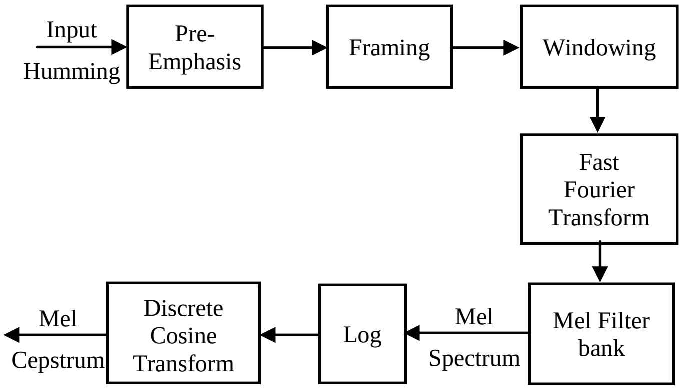

Fig. 1Steps involved in MFCC

1) Pre-Emphasis: Pre-Emphasis is a technique used in speech processing to enhance high frequencies of the signal. It is done by using FIR high-pass filter.

$$y[n] = x[n] \times w[n] \qquad 0 \le n \le L-1 \qquad (1)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample [11].The humming signal generally contains more person specific information in high frequencies than the lower frequencies as well as pre-emphasis removes some of the glottal effects from the vocal tract parameter[15].

*2) Framing:*During this step, the continuous speech signal is divided into frames in which each frame consist of N samples. Two adjacent frames are separated by M samples (M<N). For long time period, speech waveform is not considered stationary but that over a short time interval say about 10-30 ms, it can be considered stationary. Usually the frame size is equal to power of two in order to facilitate the use of FFT. Successive frames are overlapping with each other by M samples[16].Overlapped frame is used for maintaining continuity between frames. Overlapping ensure high correlation between coefficient of successive frames [17].

3) Windowing: To window each individual frame means to minimize the signal discontinuities at the beginning and end of each frame. To minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame.

$$w[n] = 0.540 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) \quad 0 \le n \le L-1 \qquad (2)$$
$$= 0 \qquad\qquad Otherwise$$

Where, $w[n] =$ Window operation, $n =$ Sample andL =Number of samples per frame. Windowing plays an important role before converting the signal in to thefrequency domain. Fourier transform is done by assuming that the signal repeats,and the end of oneframe does not connect easily with the starting of the next one.This indicates some glitches at regular interval. So we have to make the ends of each frame smooth enough to connect with each other. This is called windowing [18].

*4) Fast Fourier Transform (FFT):*During FFT, each frame of N samples is converted from time domain to frequency domain. It reduces computation time required to compute a DFT and improve performance by a factor of 100 or more over direct evaluation of the DFT. The Fourier Transform convertsthe convolution of the glottal pulse U[n] and the vocal tract impulse responseH[n] in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) * x(t)] = H(w) \times X(w)(3)$$

Where,$H(w)$, $X(w)$and $Y(w)$are the Fourier transform of $h(t)$, $x(t)$and $y(t)$respectively [11].

*5) Mel Filter Bank:*Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency fin Hz:

$$mel(f) = 2595 * \log_{10}(1 + f(Hz)/700)(4)$$

One approach to Simulink the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale. The filter bank has a triangular band pass frequency response. The bandwidth of each filter is determined by the centre frequencies of the two adjacent filter and is dependent on the frequency range of the filter bank and number of filter chosen for design.Each filter's magnitude frequency is triangular in shape and equal to unity at the centre frequency and decline linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components [18].

*6) Log and Discrete Cosine Transform (DCT):*Cepstrum is defined as the Fourier transform of the logarithm of the spectrum. Any repeated patterns or periodicity in a spectrum is considered as one or two specific components in a cepstrum [18]. So output of the mel filter bank is spectral components which is given to the Log block. This log Mel spectrum is converted into time domain using Discrete Cosine Transform (DCT). The result of the reconstruction is called Mel-Frequency Cepstrum Coefficient. The collection of coefficient is called acoustic vectors. So that, each input utterance is transformed into a sequence of acoustic vector [11].

B. *Pearson Correlation Coefficient*

In statistics, the Pearson correlation coefficient is a measure of the linear relationship between the two sets of data. It gives values between +1 and −1 inclusive, where 1 indicates positive correlation, 0 indicates no correlation, and −1 indicates negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s. Formula for Pearson Correlation Coefficient is given by

$$r = \frac{n(\sum xy)-(\sum x)\times(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (5)$$

Where, x and y are two feature vectors.The correlation coefficient ranges from −1 to 1. A value 1 of correlation coefficient shows that there is an exact linear relation between two variables X and Y perfectly, which means that all the data which are there in Y increases as X increases. A value of −1 shows that the data which are there in Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables [19].

C. *Performance Measurement*

% Identification Rate is familiar measurement of the performance of a speaker recognition system. For good Speaker recognition system, % ID should be high. Identification rate is calculated as

$$SR = (N_c/N_t) \times 100 \quad (6)$$

Where,
N_c = Number of correctly identified speakers and
N_t = Total number of speaker used for machine learning

IV. Parameter Specification and Experiment results

The database is prepared from 10 subjects in the seminar room of SNPIT & RC, Umrakh. The humming samples are taken from speakers on voluntary basis. Different persons are asked to hum for the part of the songs as per person's comfort level. The hum for 20 popular songs of Arijit Singh, Atif Aslam, Shreya Ghoshal and Alka Yagnik are recorded with Audacity software of version 2.0.6 which is open source available on internet [21]. Out of which 16 songs are used for training and 4 songs are used for testing. Below table I shows finalized parameters specification for experiment.

In pre-processing step, exact start point-end point and silence period removal is done with the help of Audacity software. In framing the continuous signal is divided into frames of N samples and then windowing is done to minimize the signal discontinuities at the beginning and end of each frame. Here non overlapping window is used for windowing process. By considering the different segments of the songs, MFCC features are found out by implementing the steps of MFCC which is described in above section IV. Similarly MFCC features are found for testing file and the features of testing files are correlated to MFCC features that has been found in earlier stage. This correlation is done with the help of Pearson Correlation Coefficient formula and among all maximum value of correlation gives the output. Below table 2 shows the success rate for 44100 Hz sampling frequency and for different number of MFCC features. Similarly table 3 shows the success rate for 22050 Hz sampling frequency and for different number of MFCC features.

TABLE I
Parameter Specification for Experiment

Parameters	**Details**
Number of Speakers	10 (5 Male, 5 Female)
Data Type	Hum for a Hindi Song
File Type	.wav
Total Songs	20 popular songs from Arijit Singh, Atif Aslam, Shreya Ghoshal, and Alka Yagnik
Training Hum	16 Songs
Testing Hum	4 Songs
Training Time Duration	1s, 3s, 5s, 10s, 20s, 40s, 60s
Testing Time Duration	1s, 3s, 5s, 10s, 20s, 40s, 60s
Sampling Rate	44100 Hz, 22050 Hz
Frame Size	23.22 ms
Number of Samples per Frame	1024, 512
Recording Software	Audacity 2.0.6 (Open Source)
Microphone	Sennheiser PC 8 USB
Environment	Seminar room of SNPIT & RC

TABLE 2
SUCCESS RATE FOR FS=44100 HZ AND DIFFERENT NUMBER OF MFCC FEATURES

Sr. No.	Timing Segments	Success Rate (%)		
		Fs = 44100 Hz		
		MFCC 12	MFCC 20	MFCC 30
1	1s	65.0	82.5	87.5
2	3s	67.5	77.5	85.0
3	5s	67.5	82.5	85.0
4	10s	67.5	80.0	85.0
5	20s	70.0	80.0	85.0
6	40s	70.0	90.0	95.0
7	60s	72.5	90.0	90.0

TABLE 3
SUCCESS RATE FOR FS=22050 HZ AND DIFFERENT NUMBER OF MFCC FEATURES

Sr. No.	Timing Segments	Success Rate (%)		
		Fs = 22050 Hz		
		MFCC 12	MFCC 20	MFCC 30
1	1s	72.5	80.0	80.0
2	3s	80.0	80.0	85.0
3	5s	82.5	82.5	92.5
4	10s	75.0	87.5	90.0
5	20s	80.0	90.0	90.0
6	40s	85.0	92.5	95.0
7	60s	85.0	90.0	95.0

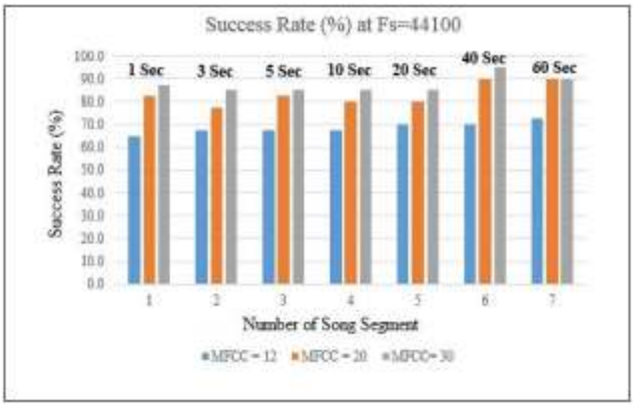

Fig. 2 Plot showing success rate for Fs=44100 & different number of MFCC features

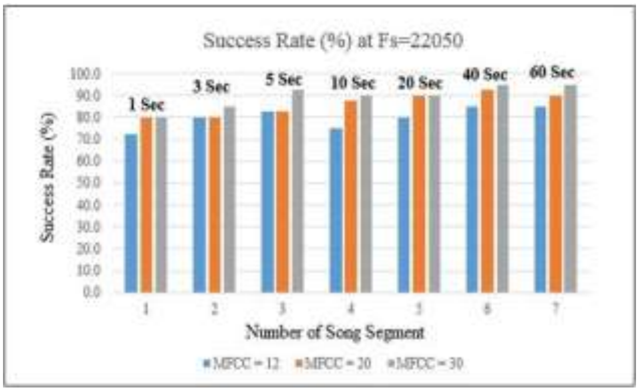

Fig. 3 Plot showing success rate for Fs=22050 & different number of MFCC features

V. CONCLUSIONS

From the above all results it is observed that as the number of MFCC features are increased, success rate is also increased at a cost of complexity. Moving towards sampling frequency, it is better to record the files at lower frequency rather than considering high frequency. The success rate can also be increased by considering the different features like VTMFCC,Δ Cepstral, MFCC-VTMP, MFCC and VTMFCC with Δ-Cep and shifted delta cepstrum (SDC).The work by considering overlapping of frames and different features is under progress.

REFERENCES

[1] HemantA. Patil, Robin Jain and Prakhar Jain, "Identification of speakers from their hum", springer-verlag berlin heidelberg, pp. 461 - 468, 2008.

[2] Minho Jin, Jaewook Kim and Chang D. Yoo, "Humming-Based Human Verification and Identification", International Conference on Acoustic, Speech and Signal Processing, ICASSA, pp. 1453-1456, April 2009.

[3] Hemant A. Patil and Keshab K. Parhi, "Novel Variable length Teager Energy Based features for person recognition from their hum", International Conference on Acoustics, Speech and Signal Processing, pp. 4526-4529, March 2010.

[4] Hemant A. Patil, Maulik C. Madhavi, Rahul Jain and Alok K. Jain, "Combining Evidence from Temporal and Spectral Features for Person Recognition Using Humming", Springer-Verlag Berlin Heidelberg, pp. 321 -328, 2012.

[5] Hemant A. Patil and Maulik C. Madhavi, "Significance of Magnitude and Phase Information via VTEO for Humming Based Biometrics", International Conference on Biometrics, pp. 372-377, 2012.

[6] Hemant A. Patil, Maulik C Madhavi and Keshab K. Parhi , "Static and dynamic information derived from source and system features for person recognition from humming", Springer Science+BusinessMedia, pp. 393-406, LLC 2012.

[7] Hemant A. Patil, Maulik C Madhavi and Nirav H. Chhayani, "Person Recognition using Humming, Singing and Speech", International Conference on Asian Language Processing, pp. 149- 152, Nov. 2012.

[8] Hector Perez Meana, *Advances in Audio and speech signal processing technologies and applications*; Idea group publishing, pp. 377, 2007.

[9] H. Hermansky, B. A. Hanson, H. Wakita, "Perceptually based Linear Predictive Analysis of Speech", Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 509-512, August 1985.

[10] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based Processing in Automatic Speech Recognition", Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 1971 -1974, April 1986.

[11] Nidhi Desai, Prof.Kinnal Dhameliya and Prof. Vijayendra Desai, "Recognizing voice commands for robot using MFCC and DTW", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6456-6459, May 2014.

[12] Vikrant Tomar and Hemant A. Patil, "On the development of variable length Teager energy Operator (VTEO)", Proc. Interspeech, Brisbane, Australia, pp. 1056-1059, September 2008.

[13] Hemant A. Patil, Prakhar Kant Jain and Robin Jain, "A Novel Approach to Identifjication of Speakers from their Hum", Seventh International Conference on Advance in pattern Recognition, pp. 167-170, Feb. 2009

[14] W. M. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifier", IEEE Trans. on Speech and Audio Processing, vol. 10, no. 4, pp. 205-212, May 2002

[15] DipakHarjani, MohitaJethwani and Ms. Mani Roja, "Speaker Recognition System using MFCC and Vector Quantization Approach",

International Journal for Scientific Research & Development, Vol. 1, Issue 9, pp. 1934-1937, 2013

[16] Prof. Ch.Srinivasa Kumar and Dr. P. Mallikarjuna Rao, "Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm", International Journal on Computer Science and Engineering, 2011, pp.2942-2954.

[17] Samiksha Sharma, Anupam Shukla and Pankaj Mishra, "Speech and Language Recognition using MFCC and DELTA-MFCC", International Journal of Engineering Trends and Technology (IJETT), Volume 12, Number 9, pp. 449-452,June 2014

[18] Arjun Rajsekhar, "Real time Speaker Recognition using MFCC and VQ", M.TechThesis,National Institute of Technology, Rourkela, India, 2008

[19] (2015), Pearson product-moment correlation coefficient, [Online]. Available: http://en.wikipedia.org/wiki/Pearson_productmoment_correlation_coefficient

[20] (2015), Audacity Software, version 2.0.6, [Online]. Available: http://www.fosshub.com/Audacity.html/audacity-win-2.0.6.exe