Measuring the Effect of Music Therapy on Voiced Speech Signal

Pradeep Tiwari[1(✉)], Utkarsh V. Rane[1], and A. D. Darji[2]

[1] MPSTME, NMIMS University, Mumbai, India
pradeep.tiwari@nmims.edu, raneutkarsh@gmail.com
[2] SVNIT, Surat, India
addarji@gmail.com

Abstract. With the rapid development in the field of speech processing, the human speech is being analyzed from different perspectives. Now-a-days impact of external factors like music on speech are also being studied by the researchers. It is widely accepted fact that the music plays important role in refreshing the mood when we see most of the people listening to the music in train or bus to get rid of boredom. This paper deals with the relation between music & its effect on human speech based on the fact that brain (cerebrum) has control over vocal tract (speech). It is also observed that the people work efficiently while listening music to increase their alertness & concentration. By studying voice samples of fatigued persons (physically or mentally fatigued) of different age-groups, it has been observed that listening to music reduces considerably the average mean & the average standard deviation feature of the speech waveform. It has also been observed that average energy of the speech waveform gets reduced & its zero crossing rate (ZCR) gets increased.

Keywords: Music therapy · Speech Feature Extraction · Stress

1 Introduction

Stress is called the lifestyle disease in today's world that not only limits individual's capabilities, interest levels and mood but also causes physical and mental health problems as in [1]. Music therapy is getting more and more reputation nationally and internationally since it is painless, no side effects, and low cost treatment for depressed patients in [2]. Background music can also help enhancing the efficiency of individuals who work with their hands as it increases their alertness and concentration in [3]. Mental condition plays an important role in the course of recovery and affect the efficiency of administered medicines in the process of disease and cure and should be taken into consideration during diagnosis and treatment in [4].

Speech signal is composed of a sequence of sounds which are produced as a result of acoustic excitation of the vocal tract when air is expelled from the lungs as shown in Fig. 1. The first step of speech production is when the speaker formulating the message in mind which he intends to transmit to the listener via medium.

The message is then converted into language code with the help of set of phoneme sequences corresponding to the sounds that make up the words. The prosody which

Z. Patel and S. Gupta (Eds.): ICFITT 2017, LNICST 220, pp. 147–156, 2018.
https://doi.org/10.1007/978-3-319-73712-6_15

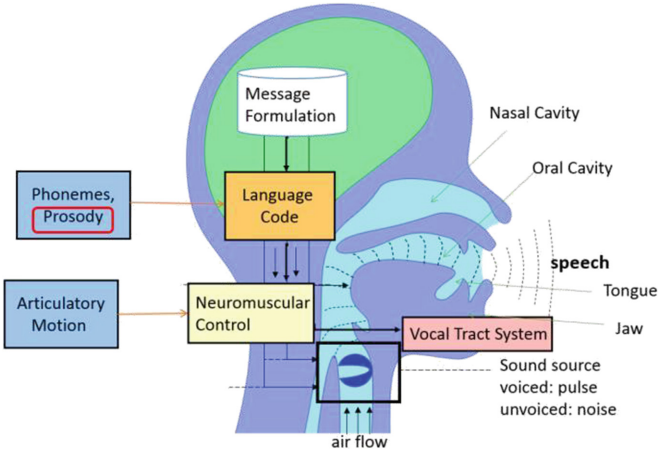

Fig. 1. Process of speech production

defines the stress is also added at this step in accordance with duration, loudness of the sounds & also the pitch associated with it as in [5].

Stressed Speech is defined as the speech produced under any condition that causes the speaker to vary the speech production from the neutral condition as in [6]. If a speaker is in a 'quiet room' with no task responsibilities, then the speech produced is considered neutral. Stress can be classified as (a) Emotionally induced stress: Speech produced under change in the emotional or psychological state of the speaker such as angry speech, sad speech, happy speech etc. (b) External Environmentally induced stress such as Lombard Speech (c) Pathological Stressed Speech such as Cold affected speech, Old age Speech. In this paper, emotionally and External environmentally induced stressed speech are considered.

Different subjects with stress were selected. Speech signal are acquired from each subjects before and after they were subjected to listening music. The recorded speech signal is sampled at sampling frequency of 44100 Hz and 16 bits per sample. The speech feature selection is an important problem in stress identification. Speech features such as average Energy, average mean, average Standard Deviation, average Zero Crossing Rate (ZCR) are obtained for the acquired speech signal from the subjects before and after hearing music. It was clearly seen that average energy, average mean & average standard deviation is found decreasing after hearing the music. Also it is observed that ZCR is increasing after listening the music.

2 Related Works

There have been considerable research works in the field of Neuropsychology and speech processing since past two decades. The assessment carried in [7] indicates that the relaxation and concentration improves using Alpha music, which influence the alpha and beta rhythms significantly. When subjects faced Alpha music, they felt

considerable reduction in fatigue/stress along with the increase in the physical relaxation. Thus it is evident that music affects human brain & relaxes it during fatigue. Another kind of work showing effect of music on human body & mind where subjects were made to listen to particular music for particular span of time & their heart rate variability was measured using ECG machine. It has been observed that the single cluster formed by the volume of the Point Care Plots of Spherical Coordinate is reduced remarkably in the data acquired during music state as compared to the data of pre music state & also the amount of reduction is not the same for all the subjects. This proves that music has some definite effect on human physiology [8]. Some investigations of researchers are even showing improvement in the typewriting work performance due to impressions of music by measuring some bio-signals to monitor participant's condition. Furthermore, the music impression causes activation in saliva amylase which decreases fatigue/stress as in [9]. A study is presented in [10] depicts that when a subject is exposed to live violin music performance, its brain induces theta, alpha and beta brainwaves to get balanced. Another study shows that to relieve users from depression, an electroencephalogram (EEG) based music therapy system was used to identify the user and to measure the degree of depression giving results which gives conclusion that the EEG approach is user-based approach for preventing depression [11]. A number of studies have been done on understanding the relation between music & brain & one such suggests that injury to brain can drastically impair musical activities except leaving intellectual and linguistic abilities. This research indicates that music cognition is not affected entirely, but in particular abilities [12]. Another work carried by a researcher presents that music enhances spatial-temporal reasoning when the data collected from College students & Preschool children. It also showed that the effect is more if the subjects are exposed to longer duration of time [13]. In [13] a structured neural model has been analyzed, describing a certain kind of relationship between music and spatial-temporal reasoning. The study presented in [14] has developed relationship between music, and subject's emotion and concentration with the help of EEG device. In [15] EEG device is used to check the emotional responses while listening music with the help of low cost cloud based architecture. The functioning of the human brain while listening to music is studied in [16] with the help of Natural stimulus functional magnetic resonance imaging (N-fMRI). The observation derived in this experimentation shows that music like classical music, pop etc. affects the attention and emotions of human brain [16].

3 Process Description

This section describes the block diagram of the project (Fig. 2).

3.1 Data Acquisition

This is the first step of the project. In this speech signal of sampling frequency 44.1 kHz & 16 bits/s using microphone has been acquired. The microphone of cellphone HTC Desire 620G is used for recording voices. The vowels a, e, i, o, u are considered to be spoken by each speaker as they are voiced sounds, hence these vowels

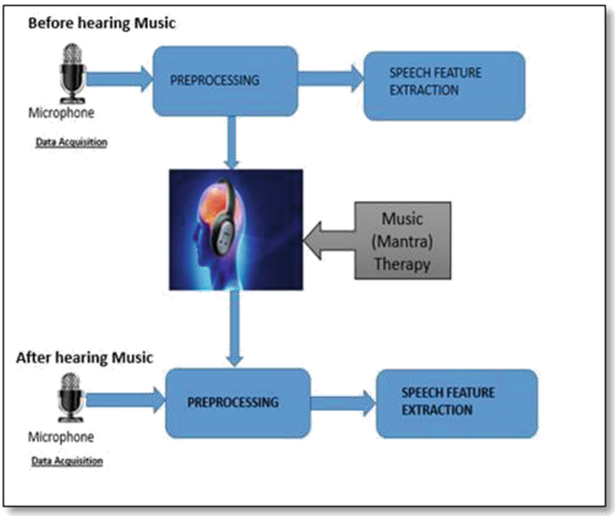

Fig. 2. Block diagram

contain high energy & maintains uniformity in the sentence spoken by each speaker. Ten voice samples of ten subjects, both male & female of various age groups are taken before hearing the music & then again after hearing the music. In this the subjects are made to hear sweet & soft romantic music. The phone's speakers used have following specifications: Stereo speakers of HiFi edition with one at the top of the phone acting as tweeter & other at the bottom acting as woofer & working together to offer experience of high quality surround effect with Dolby Audio 4. All the ten subjects are from Mumbai, India region speaking typical Indian accent English. The recording time for each sample was around 3–6 s. Mainly the age-groups are youngsters like between ages 18 to 26 of which the three voice samples of male & female each are acquired. Two voice samples of one male of age 64 & one female with the age of 62 are taken. Finally, the two voice samples of middle-aged group around 40 to 50 years of age are taken which are both females. So different aged voices are used to make project more extensive & authentic.

3.2 Data Preprocessing

The acquired data of ten subjects which is in '.wav' format is plotted as shown in Fig. 3. The waveforms of the speech sample before hearing and after hearing music is analyzed which reflect the fact that there is variation in the speech produced on hearing music. The five vowels were clipped separately for both the cases (i.e. before hearing the music & after hearing the music). Further preprocessing is done for '.wav' files. Preprocessing includes normalization & pre-emphasis. Direct current offset (de-offset) carries no useful information rather it can carry disturbing information. Removal of de-offset is called normalization. The statistical normalization which is widely given by,

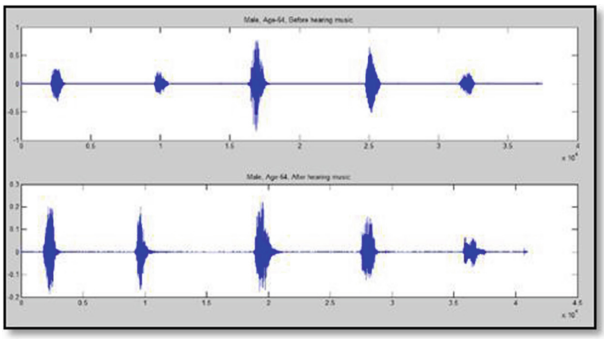

Fig. 3. Speech sample (a, e, i, o, u) waveform before and after hearing music

$$S_n = \frac{S - mean(S)}{Variance(S)} \tag{1}$$

The higher frequency components of speech signals are suppressed while speech production. Pre-emphasis increases the magnitude of the higher frequencies with respect to the magnitude of the lower frequencies. A simple first-order high pass, FIR filter is generally used for pre-emphasis as given below:

$$H(z) = 1 - kz^{-1} \quad \text{where k} \in [0.9, 1] \tag{2}$$

Next step is Speech Feature Extraction to identify the changes in each vowel spoken in both the cases.

3.3 Speech Feature Extraction

This is the next step in this research paper. In this step the features considered for analysis are Short time Energy, Zero Crossing Rate (ZCR) and Statistical features (Mean, Standard Deviation).

3.3.1 Short-Time Energy

It has been watched that the amplitude of the speech signal fluctuates considerably with time. Specifically, the sufficiency of the unvoiced portions is by and large lower than the amplitude of voiced portions. The short-time energy of speech signal gives a helpful portrayal that mirrors these amplitude fluctuations. As a rule, it can be characterized the short-time energy as in [17] as, this expression can be written as (Fig. 4),

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{3}$$

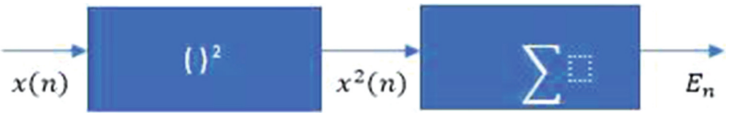

Fig. 4. Block diagram representation of the short-time energy [17]

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \quad where\, h(n) = w^2(n) \tag{4}$$

The real noteworthiness is that it gives a premise for distinguishing voiced speech segments from unvoiced speech segments. Also, for very high quality speech (high signal-to-noise ratio), the energy can be used to distinguish speech from silence.

3.3.2 Short-Time Average ZCR

With regards to discrete time signals, a zero-crossing is said to happen if successive specimens have distinctive algebraic signs as in [17]. The rate at which zero-crossings happen is a basic measure of the frequency content of a signal. This is especially valid for narrowband signals. For example, a sinusoidal signal of frequency F_0 sampled at a rate F_s, has F_s/F_0 samples per cycle of the sine wave. Each cycle has two zero-crossings so that the long-time average rate of zero-crossings is

$$Z = \frac{2F_0}{F_s} crossing/sample \tag{5}$$

Thus, the average zero-crossing rate gives a reasonable way to estimate the frequency of a sine wave. Speech gives rough estimates of spectral properties can be acquired utilizing a representation of zero-crossing rate for the speech. A suitable definition is (Fig. 5),

$$Z_n = \sum_{-\infty}^{w} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|w(n-m) \tag{6}$$

Where,

$$sgn[x(n)] = 1 \quad x(n) \geq 0$$
$$= -1 \quad x(n) < 0$$

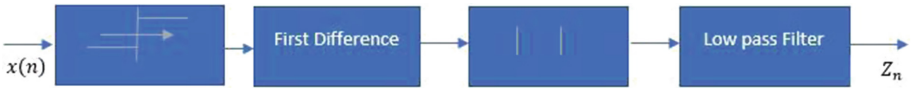

Fig. 5. Block diagram representation of short-time average zero-crossings [17]

And

$$w(n) = \frac{1}{2N} \quad 0 \le n \le N-1 \\ = 0 \qquad \text{otherwise} \tag{7}$$

Zero Crossing Rate (ZCR) represents the frequency content of a signal, since it checks number of times the amplitude of the speech signals passes through a value of zero in a given time interval/frame. The interpretation of average ZCR is less exact since speech or voice signals are broadband signals, however it can be generally estimated.

3.3.3 Statistical Features

The mean and standard deviation are the two prominent statistical features explored in this paper. The following equation gives Mean value for every vowel:

$$M = \sum_{i=1}^{n} \frac{x_i}{n} \tag{8}$$

The Standard Deviation value for every vowel is given by following equation:

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - M)}{n-1}}. \tag{9}$$

4 Implementation and Results

The core idea of undertaking this project is to identify & analyze the impact of music on the speech of person under fatigue. Both the types of fatigue are considered: mental & physical. The acquired speech samples generated for this project are vowels a, e, i, o & u. The ten samples of each five vowels are considered for before hearing and after hearing the music. The energy of each vowel is calculated as shown in the Fig. 6.

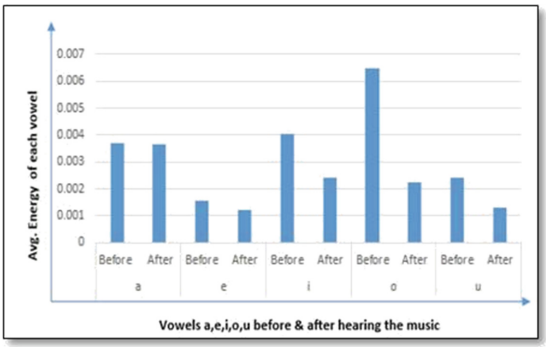

Fig. 6. Average Energy of vowels a, e, i, o, u before & after hearing music

As it can be seen from Fig. 6 that average energy values of each vowel are higher before hearing the music as compared to the value of average energy values after hearing the music. Similarly, it is depicted in the Fig. 7 that average mean values of each vowel are more before hearing the music than after hearing the music.

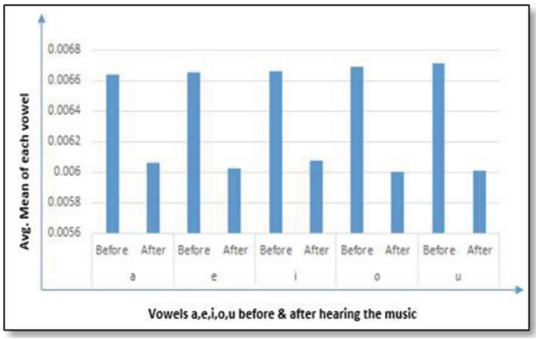

Fig. 7. Average Mean of vowels a, e, i, o, u before & after hearing the music

Also from the Fig. 8, it can be observed that average standard deviation values before hearing the music for each vowel is more than the values for vowels after hearing the music.

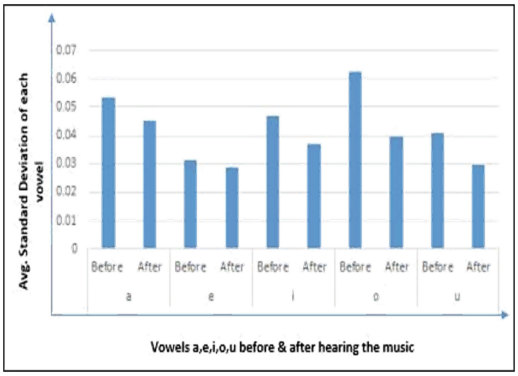

Fig. 8. Average Standard Deviation of vowels a, e, i, o, u before & after hearing the music

There is a difference which can be seen in the Fig. 9 for ZCR feature. The ZCR values for each vowel after hearing the music are more compared to the values before hearing the music. Though the values for vowel O are showing a kind of deflection from the general pattern but it can be generalized that the ZCR values are generally more after hearing the music than the ZCR values before hearing the music.

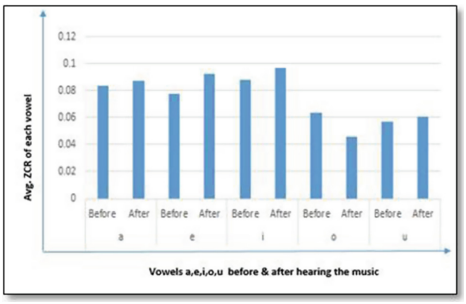

Fig. 9. Average Zero Crossing Rate of vowels a, e, i, o, u before & after hearing the music

5 Conclusion

In today's world, due to the competition and fast paced life, students, corporates and family often experience fatigue or stress which reduces their expected performance. This paper clearly gives an evidence that the speech is connected to & is in the control of brain by means of prosody. Furthermore, the paper concludes that the music affects the speech features, hence the stress level of the subject. The future scope of this project is that the statistical features of the speech & the facial features can be considered for better analysis of the impact of music on the speech.

References

1. Vandyke, D.: Depression detection & emotion classification via data-driven glottal waveforms. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 642–647. IEEE (2013)
2. Zhou, P., Lin, D., He, W., Li, G., Shang, K.: Influence of musicotherapy on mental status and cognitional function of patient with depression disease. In: 2009 2nd International Conference on Biomedical Engineering and Informatics, pp. 1–4. IEEE (2009)
3. Restak, R.: Mozart's Brain and Fighter Pilot. Crown Publications, New York (2003)
4. Cornelia, B., Richardson-Boedler, C.: Applying Bach Flower Therapy to the Healing Profession of Homoeopathy. B. Jain Publishers, New Delhi (2003)
5. Rabinar, L., Juang, B.H., Yeganarayana, B.H.: Fundamental of Speech Recognition. Pearson (2010). Second Impression
6. Ramamohan, S., Dandapat, S.: Sinusoidal model-based analysis and classification of stressed speech. IEEE Trans. Audio Speech Lang. Process. **14**(3), 737–746 (2006)
7. Vijayalakshmi, K., Sridhar, S., Khanwani, P.: Estimation of effects of alpha music on EEG components by time and frequency domain analysis. In: 2010 International Conference on Computer and Communication Engineering (ICCCE), pp. 1–5. IEEE (2010)
8. Das, M., Jana, T., Dutta, P., Banerjee, R., Dey, A., Bhattacharya, D.K., Kanjilal, M.R.: Study the effect of music on HRV signal using 3D Poincare plot in spherical co-ordinates-a signal processing approach. In: 2015 International Conference on Communications and Signal Processing (ICCSP), pp. 1011–1015. IEEE (2015)

9. Iwaki, M., Nakano, K.: Typewriting performance affected by music impression in working environment. In: 2013 Proceedings of SICE Annual Conference (SICE), pp. 1539–1543. IEEE (2013)
10. Hassan, H., Murat, Z.H., Ross, V., Buniyamin, N.: A preliminary study on the effects of music on human brainwaves. In: 2012 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 176–180. IEEE (2012)
11. Peng, H., Hu, B., Liu, Q., Dong, Q., Zhao, Q., Moore, P.: User-centered depression prevention: an EEG approach to pervasive healthcare. In: 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, pp. 325–330. IEEE (2011)
12. Peretz, I., Hébert, S.: Music processing after brain damage: the case of rhythm without melody. In: Steinberg, R. (ed.) Music and the Mind Machine, pp. 127–137. Springer, Heidelberg (1995). https://doi.org/10.1007/978-3-642-79327-1_13
13. Shaw, G.L.: Computation by symmetry operations in a structured neural model of the brain: music and abstract reasoning. In: Cabrera, B., Gutfreund, H., Kresin, V. (eds.) From High-Temperature Superconductivity to Microminiature Refrigeration, pp. 287–311. Springer, New York (1996). https://doi.org/10.1007/978-1-4613-0411-1_25
14. Sourina, O., Kulish, V.V., Sourin, A.: Novel tools for quantification of brain responses to music stimuli. In: Lim, C.T., Goh, J.C.H. (eds.) 13th International Conference on Biomedical Engineering, pp. 411–414. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92841-6_101
15. Guo, Y., Wu, C., Peteiro-Barral, D.: An EEG-based brain informatics application for enhancing music experience. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) International Conference on Brain Informatics, pp. 265–276. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35139-6_25
16. Fang, J., Xintao, H., Han, J., Jiang, X., Zhu, D., Guo, L., Liu, T.: Data-driven analysis of functional brain interactions during free listening to music and speech. Brain Imaging Behav. **9**(2), 162–177 (2015)
17. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals. Prentice Hall, Englewood Cliffs (1978)