Journal of Computing and Applied Informatics (JoCAI) Vol. 02, No. 2, 2018 | 74 - 86

DATA SCIENCE
Journal of Computing and Applied Informatics

A Two Microphone-Based Approach for Detecting and Identifying Speech Sounds in Hearing Support System

***Andre Sitompul*[1], *Masafumi Nishimura* [2]**

[1,2]*Graduate School of Integrated and Technology, Shizuoka University, Hamamtsu Campus - Japan*

Abstract. For people with hearing disabilities, not only would give them difficulties in going through their everyday lives but also sometimes could be life threatening. In this research we proposed a simple, yet robust approach for helping the hearing-impaired people in identifying the important sounds around them by using two microphones as input channel that could be worn around the person's head as a substitute for their ears. This device then could be used to record the situation of the surroundings, and then the system would estimate the Direction of Arrival (DOA) of the sound sources, then detect and classify them using Support Vector Machine (SVM) into target speech or noise category. As the results, system's classifier could produce FAR and FRR as low as 2%, in which 274 out of 280 samples were successfully classified as target speeches and 22 from the total of 27 noise samples were successfully classified as noise

Keyword: Hearing support system, Direction of Arrival, two-microphones, Support Vector Machine.

Abstrak. *Bagi orang dengan gangguan pendengaran, tidak hanya akan memberi mereka kesulitan dalam menjalani kehidupan sehari-hari mereka tetapi kadang-kadang juga dapat mengancam kehidupan. Dalam penelitian ini kami mengusulkan pendekatan yang sederhana namun sesuai untuk membantu orang dengan gangguan pendengaran dalam mengidentifikasi suara penting di sekitar mereka dengan menggunakan dua mikrofon sebagai saluran masukan yang dapat dipakai di sekitar kepala orang tersebut sebagai pengganti telinga mereka. Perangkat ini kemudian dapat digunakan untuk merekam situasi di sekitarnya, dan kemudian sistem akan memperkirakan Arah Kedatangan (DOA) dari sumber suara, kemudian mendeteksi dan mengklasifikasikannya menggunakan Support Vector Machine (SVM) ke dalam kategori noise atau target suara . Sebagai hasilnya, pengklasifikasi sistem dapat menghasilkan FAR dan FRR serendah 2%, di mana 274 dari 280 sampel berhasil diklasifikasikan sebagai pidato target dan 22 dari total 27 sampel suara berhasil diklasifikasikan sebagai noise.*

Kata Kunci: *Sistem pendukung pendengaran, arah kedatangan, 2-micropon,* Support Vector Machine

Received 09 April 2018 | Revised 07 May 2018 | Accepted 18 June 2018

*Corresponding author at: Shizuoka University, Hamamtsu Campus, 3 Chome-5-1 Johoku, Naka Ward, Hamamatsu, Shizuoka 432-8011, Jepang

E-mail address: andre.sitompul.15@shizuoka.ac.jp

ISSN: 2580-6769 | e-ISSN: 2580-829X | DOI: 10.32734/jocai.v2.i2-325
Journal Homepage: https://talenta.usu.ac.id/JoCAI

1. Introduction

According to the World Health Organization (WHO), there are approximately 360 million individuals suffer from deafness, which belong to the 5% of world population. From these numbers, it is obvious that the number of hearing-impaired people in this world holds a huge portion from world population, which leads to a situation that a support system is needed to assist the hearing-impaired people in detecting and identifying sounds around them.

The number of hearing-impaired people in this world holds a huge portion from world population, which leads to a situation that a support system is needed to assist the hearing-impaired people in detecting and identifying sounds around them.

Several existing technologies available in the market for assisting the hearing-impaired people to become aware of their surroundings. However, the high price and the lack of portability of some of the devices might trouble the hearing-impaired people in using the technologies for their everyday lives.

In this research, we proposed a simple, yet robust approach by using two microphones that could be worn on person's head to record the surrounding situation of the hearing-impaired people, then the detected sounds will be extracted. In order to make the system more robust, we combine the system with a classifier, so that the sounds that discovered in the earlier stage, which might consist of target speeches and noises (any sounds except target speeches), could be successfully classified into the right categories. Then the system would show the classification results of the existing sounds, whether it is a target speech or noise, to the hearing-impaired people. Using this approach, the hearing-impaired people could become more cautious with their surrounding environment and have a much better experience in living their lives.

2. Related Work

In the last couple of years, various solutions have been employed to help the people with hearing disabilities. Pioneered by hearing aid devices that amplify sound of the surroundings, so that the target speech or the target signal could be heard more clearly by the hearing-impaired people [1][2]. This type of assistive technology helps people with half-hearing loss to participate more fully in conversations. The technology works by amplifying any sound waves through the use of microphone, amplifier, and speaker. There are also various ways to use the hearing aid devices, including digital, in-the-ear, in-the-canal, behind-the-ear, and on-the-body aids.

Furthermore, as the hearing aids progress from analog to digital, it allows more precise corrections to the unique pattern of specific hearing losses. Later, it also allows people with hearing difficulties to focus on specific sounds by getting rid of extra background noises, reverberated sounds, and distractions.

Afterwards, as the technology progresses, health professionals come up with another type of hearing assistive device called alerting devices [3]. This device helps the hearing-impaired people in identifying some specific sounds, such as doorbell, telephone sounds, fire alarm, etc. This device works by connecting an additional device, such as transmitter to a doorbell, telephone, or alarm. Then, the transmitter would transmit the information to the alarming devices, every time these sounds are available. Later, the device will produce a loud sound, vibrations, or blinking light to let the people with hearing difficulties aware with their surroundings.

In the last few years, a company called Tokyo Shinyu developed two different approaches for realizing assistive alarming devices [4]. The systems are called Silwatch and Cube Light. These two devices have the same objective which is to help the hearing-impaired people aware with some specific sounds existed in their surroundings. They also utilize a number of transmitters for realizing the fundamental functions of the systems. Both systems work almost in the same way as most of the alarming devices (see Figure 1).

Figure 1. Silwatch and Cube Light for assisting hearing-impaired people
NOTE: Reprinted from Tokyo Shinyu product catalogue, Copyright 2017 by Tokyo Shinyu Co., Ltd. | http://www.shinyu.co.jp/product/index.html

However, the interface used for displaying the sounds are quite different. As for the Silwatch, the information is transmitted and displayed on top of the watch as text notifications. And for the Cube Light, the information would be transmitted and displayed through a device that looks like a pager and would emits various different lights based on the sounds available at a specific point of time.

Another innovation related to the development of assistive technology which is now still in the developing phase, an innovation from Fujitsu called Ontenna. It is a wearable hearing assistive device and could be worn all over the body, for example as necklace, earrings, hair clips, etc [5]. Like most of other assistive technologies, it detects surrounding sounds such as boiling water, car horns, fire alarms, etc., and send the information to the hearing-impaired people through different types of vibrations. One of the important aspects of this device is, obviously, its portability. The possibility to wear it everywhere over one's body makes this body practical for everyday use. However, since it is still in the development phase, the price tag is yet to be decided. Some predictions also suggest that the cost would be quite expensive.

3. Methodology

In general, the developed system is divided into two different environments; internal and external environments. As the name represents, the internal environment mainly dealing with background processes in which the sounds must gone through before converted into usable and functional information, and later transmitted to the hearing-impaired people. On the other side, the external environment mostly deals with the technology to notify and to transmit the processed information from wearable technologies to the hearing-impaired people. The main focus of this research lies on the internal environment of the system which mostly concentrated on the background processes/methods used in retrieving the useful information. Furthermore, in order to measure whether the system is feasible to be implemented in wearable devices, Raspberry Pi was used as an operating system, with assumptions by using Raspberry Pi's moderate specifications, the system could also be run in most of existing wearable devices, such as smartphones, smart glasses, smart watches, etc.

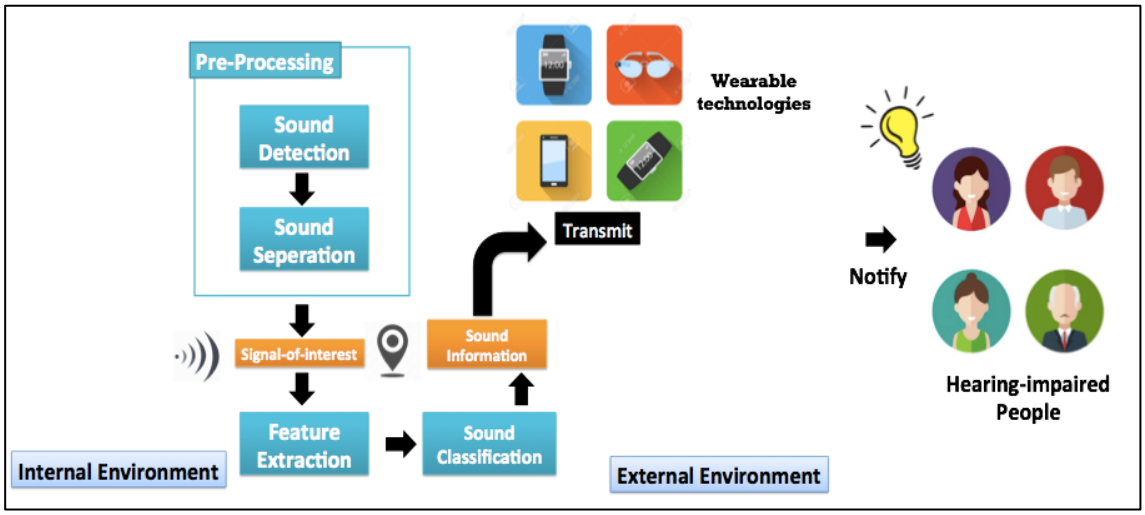

Figure 2. System framework

Figure 2 demonstrates, the internal environment of the system consists of several stages, which consists of pre-processing stage (Sound detection and sound separation), feature extraction, and lastly, sound classification. The classification results then would be processed and transmitted to wearable devices to notify the hearing-impaired people with the current condition of the surroundings. If target speech is present the system will notify the hearing-impaired people that there is something going on behind them, and if noises or other sounds are present at the moment, the system will not give any response to the wearable device. This is done so that the hearing-

3.1. Pre-Processing

In the pre-processing stage, the system conducts sound detection process to the streamed sounds retrieved from the surrounding environment and performs sound separation to the tracked sounds. First, the system would estimate the Direction of Arrival (DOA) of the available sound sources based on Generalized Cross-Correlation method weighted by the

phase transform (GCC-PHAT), also known as Cross Spectrum Phase (CSP) [6][7]. Also, referring to [8], the localization performance could be improved by performing signal-to-noise ratio (SNR) weighting. The binaural sound source localization formula is described based on the formula below [9]:

Let $\theta_{arr} = \{\theta_1, \theta_2, \theta_3, \ldots\}$ represents an array of directions of the existing sound sources, therefore:

$$\theta_{arr} = argmax_{\theta} \frac{1}{F} \sum_{f=1}^{F} \frac{SNR_{weighting}[f,n]}{1+SNR_{weighting}[f,n]} \cdot \frac{X_{left}[f,n]X^*_{right}[f,n]}{\left|X_{left}[f,n]X^*_{right}[f,n]\right|} \exp(j2\pi \frac{f}{F} fs\tau_{multi}(\theta),$$

$$SNR_{weighting}[f,n] = \frac{\left|X_{left}[f,n]X^*_{right}[f,n]\right| - E[\left|N_{left}[f,n]N^*_{right}[f,n]\right|]}{E[\left|N_{left}[f,n]N^*_{right}[f,n]\right|]}, \text{ and} \quad (1)$$

$$\tau_{multi}(\theta) = \frac{d_{left\,right}}{2v}\left(\frac{\theta}{180} + sin(\frac{\theta}{180}\pi)\right) - \frac{d_{left\,right}}{2v}\left(sgn(\theta)\pi - \frac{2\theta}{180}\pi\right).\left|\beta_{multi} sin\left(\frac{\theta}{180}\pi\right)\right| \quad (2)$$

where X[*f*,*n*] represents an input audio signal at frequency bin *f* and time frame *n.* Furthermore, GCC algorithm, which is suggested could only measure and track a single sound source, by adding dynamic k-means clustering to the GCC-PHAT algorithm, multiple sound sources could be maintained and retrieved in real-time [6].

For sounds that fulfill the GCC-PHAT threshold that has been initialized beforehand, would be separated into a number of sound files that might composed of target speeches or noises. As mentioned in the previously, the system covers a 180 degrees coverage for sound detection and separation process. The direction orientation of the system could be seen in Figure 3. The sound detection process represents a single sound source detection process which is arriving approximately 30 degrees from the back left of the person.

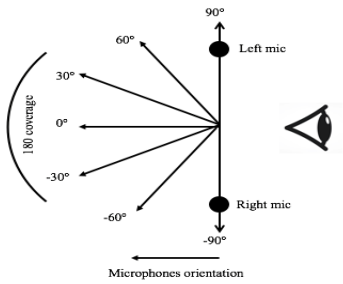

Figure 3. Direction orientation of the system

Basically, the threshold is used to avoid estimating DOA of noises and reverberated sounds. The value of this threshold also should be carefully considered, because if the threshold is set too high, the DOA of the target speech could not be detected and such information would be lost and rejected by the system, and if the threshold is set too low, the DOA of noises and reverberated

sounds would be included and handled in the next stage of the system. Finally, after estimating the DOA of existing sound sources, the system would use that information to track and separate the sound sources from the recordings. Since there is no guarantee that the separated sounds retrieved in the pre-processing stage always consist of target speech, the system requires further inspections or processes so that the possibility of having inaccurate n3.2otifications from the system (false alarm) could be minimized.

3.2. Feature Extraction and Sound Classification

Following the pre-processing stage, it is assumed that the separated sound signals retrieved by measuring the DOA of the sound sources, do not always guarantee that the sound signals would only compose of target speech, but could also compose of noises and reverberated sounds. Therefore, in order to get more robust results in notifying the hearing-impaired people regarding the existing sounds, a classifier is introduced and added into the system. The classifier will perform as filter and classify the retrieved sounds into two different categories; target speech or noise. However, before going into the classification process, features extraction must be conducted to obtain the unique features of each separated files. While there are many possible feature representations, in this thesis, we utilize 12-dimension of Mel-Frequency Cepstral Coefficients (MFCC) to represent the sound signal. This is because according to [10] MFCC could imitate the logarithmic perception of loudness and pitch of human auditory system and it could to eliminate speaker-dependent characteristics by excluding the fundamental frequency and their harmonics. In this research, we have confined the target of our research, specifically, into speech sounds and noises and we would also like to be able to classify as much as possible target speech uttered by several different persons. Accordingly, we assume MFCC is suitable for representing our purpose. Moreover, it is also powerful and robust for sound classification.

After performing feature extraction to the sound signals, the extracted features then would be handled as input vectors of the classifier. We utilized Support Vector Machine (SVM) for its robustness in performing classification tasks. SVM uses kernel trick to transform the signal data and then based on these transformations it finds an optimal boundary between the possible outputs, for this case, it predicts if the specific input vector belongs to target speech category or noise category. In a simpler term, it draws an optimal boundary line that separates input vectors into target speech class or noise class. Moreover, since SVM is a kernel-based machine learning algorithm, we could utilize different kernels and choose the one that best fits our model with low computational complexity. With this we can capture more complex relationships between our input vectors without having to perform difficult transformations by our own. Therefore, for its robustness in performing classification tasks and its low computational complexity, we assume that SVM is well suited to be used for classifying the sound signals retrieved in the pre-processing stage of the proposed system.

5. Experimental Condition

We conducted the experiment in a noise-free, 3 x 5 meters room, and performed the experiment with different number of microphones and setup. The sound data is divided into two groups: 1) Repeating sounds (Fire alarm, Door bell, Phone ring), 2) Irregular sounds (Speech sounds, three types Each sound data was evaluated by putting them in five different locations; 30, 60, 90, 120, and 150 degrees and separated around one meter from the microphone. In total, there are 30 sound files were evaluated in this experiment. The layout of the room is described in Figure 4.

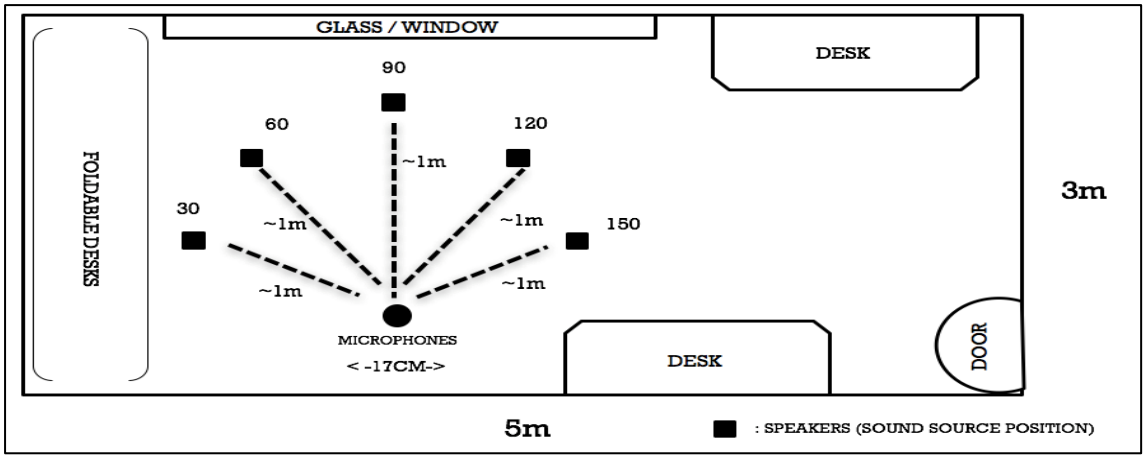

Figure 4. Room layout and experiment settings

ways: Firstly, by calculating the root mean squared error of each sound files based on its predefined location (see Figure 5), and secondly, by calculating the total root mean squared error of each methods (Figure 6).

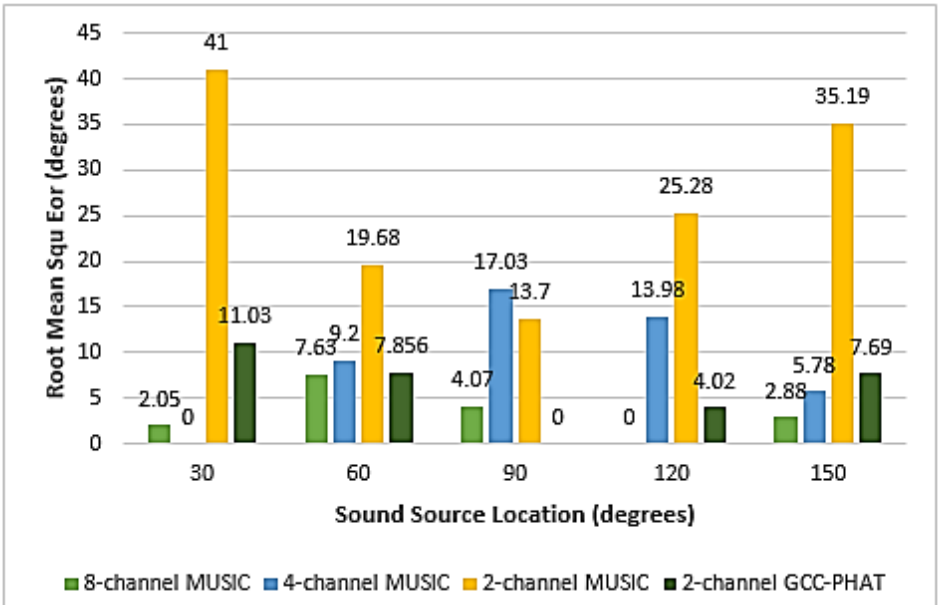

Figure 5. DOA performance comparison between MUSIC and GCC-PHAT method

By referring to Figure 6, MUSIC method produced the best results with 8-channel microphone settings. However, when the number of microphones is reduced, there is a quite significant error leaps in using the MUSIC method. The highest error was obtained when using MUSIC method with two-channel microphone and the sound file were located 30 degrees away from the microphone, while the lowest error (0 degrees) was obtained with 8-channel MUSIC method, and when the sound file was located 120 degrees away from the microphone. In comparison, the GCC-PHAT method shows some quite consistent results in various locations with the highest error around 11 degrees when the sound file was located 30 degrees away from the microphone, and the lowest error (0 degrees) was obtained when the sound file was located 90 degrees away from the microphones.

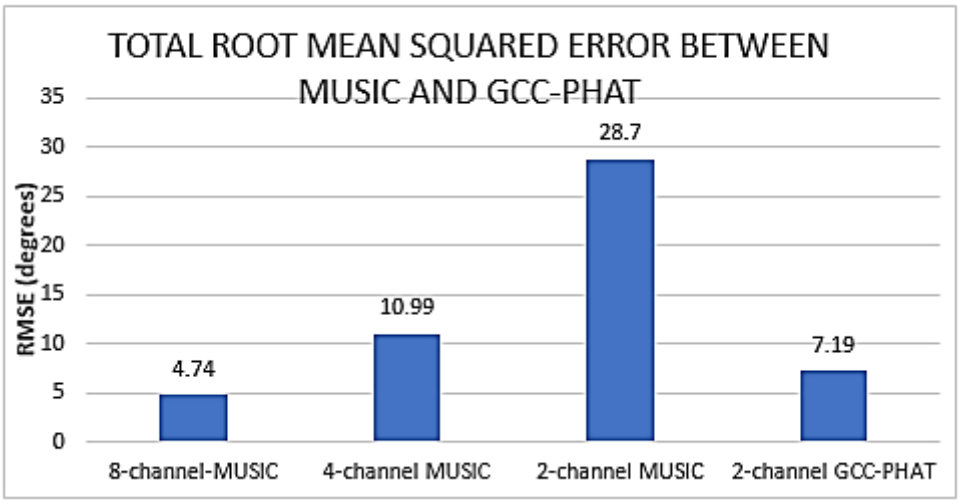

Figure 6. The Overall Root Mean Squared Error between MUSIC and GCC-PHAT method

From the results of the first experiment, we could see that that the GCC-PHAT method seems to be more suitable than the MUSIC method for developing the proposed hearing support system, which demands low-cost and portability. We also suggest that the distance between each microphone in the TAMAGO microphone might has some connection to the performance of DOA estimation. Based on this outcome, we decided to continue using the GCC-PHAT method to measure the DOA of the sound source and we perform another experiment by increasing the number of data and changing the type of sound data by confining it into speech sounds only. Besides that, we also decided to utilize two mono-aural microphones that separated around 16 cm from each other. 16 cm was decided as the distance between the two microphones because this number represents the proximate distance of person's left ear and right ear. We suggest with this setting the system could be implemented and worn around a person's head. The description of the new simulated data is as follows:

1. 10 types of speech sounds, which consists of greeting sounds and attraction sound:
   a. Good Morning b. Good Afternoon c. Good Night d. Konnichiwa
   e. Hi f. Sumimasen g. Ano
   h. Excuse me i. Sorry j. Hi, excuse me
2. Each sound was uttered by four different persons
3. Each sound was put in the same location settings with the previous experiment
   a. 30 degrees and 1 meter
   b. 60 degrees and 1 meter
   c. 90 degrees and 1 meter
   d. 120 degrees and 1 meter
   e. 150 degrees and 1 meter

Total number of simulated sound data: 200 files

The experiment is evaluated in the same ways as the previous experiment, which are through RMSE calculation for each location of the sound files, and RMSE calculation for the whole experiment. Th first calculation of RMSE is represented in Figure 7.

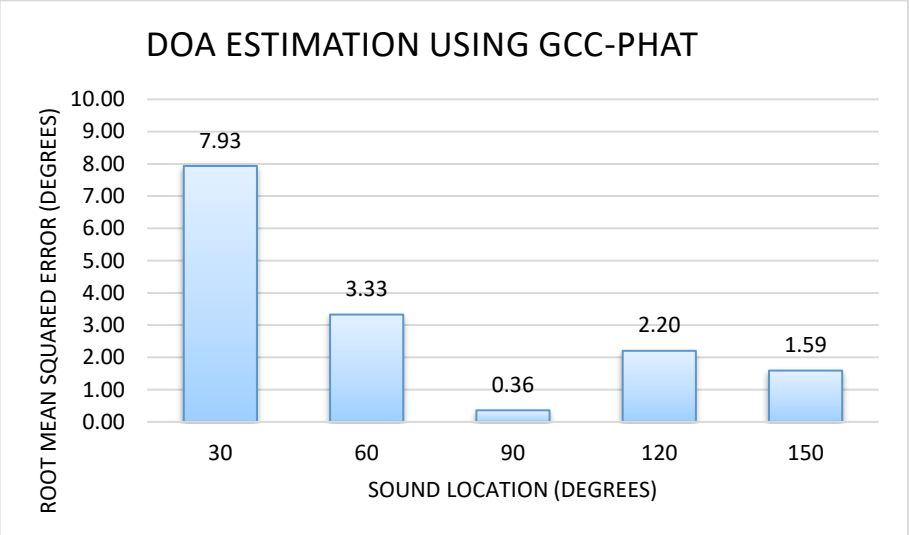

Figure 7. RMSE of DOA estimation using GCC-method

From Figure 8, the most accurate direction of arrival estimation was produced when the sound source is located 90 degrees from the center of the microphones (right behind the microphones). It also shows, almost in every case (30, 60, 90, 120, 150) the error is reduced compared to the previous settings of the experiment. This means several adjustments carried out to the microphone settings and the amount of sound file leads to the improvement in performing the direction of arrival estimation toward the sound signals.

In Figure 8, the overall RMSE results between 8-channel MUSIC and 2-channel GCC-PHAT is only around 3 degrees. From these results, we can conclude in terms of accuracy, MUSIC with 8-channel settings produced much better results than the 2-channel GCC-PHAT.

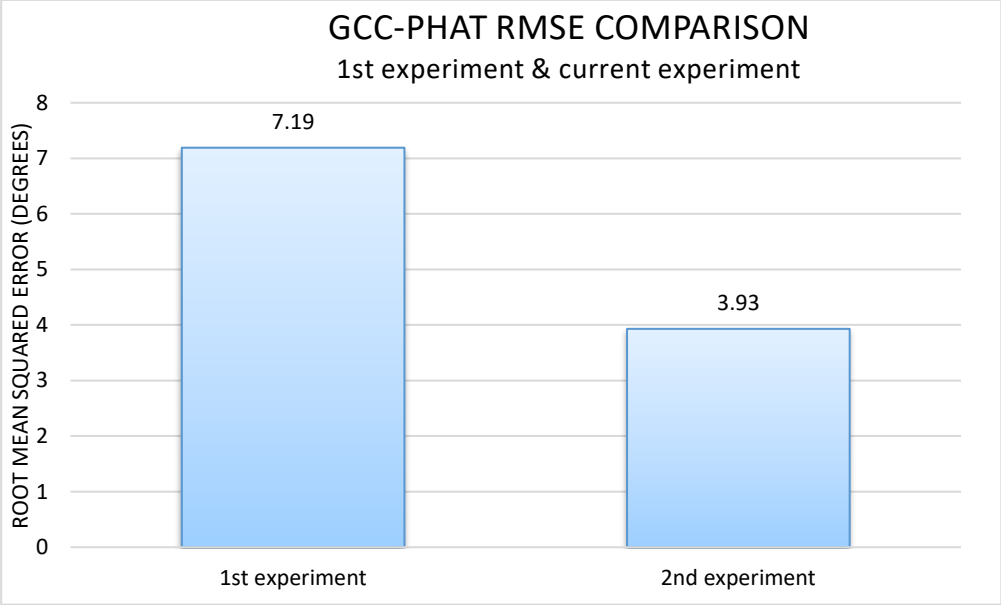

Figure 8. Overall RMSE calculation between the 1st experiment and the 2nd

However, if cost and portability factors are also considered in developing a system, 2-channel GCC-PHAT offers a better performance than the 8-channel microphone, despite the slightly lower results in accuracy compare to the 8-channel MUSIC. If the overall RMSE calculation of the 1st experiment and the 2nd experiment are compared, the 2nd experiment displays a much better degree of error compare to the 1st experiment. Based on the reassuring results from the previous two experiments, GCC-PHAT with two microphones seems to be the most suitable approach for implementing the proposed hearing support system, especially in estimating the direction of arrival of the sound signals

Signal to noise ratio is one way to measure the power of the intended signal relative to background noise. SNR of the sound data is calculated to describe the characteristic of the simulated data. In doing so, the SNR ratio could be used as a reference to assess the performance of sound data. Furthermore, two classifiers are used to evaluate the performance of the system. They are Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), see Figure 9 and Figure 10 respectively. The classifiers are going to be evaluated by calculating the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) of each classifier. False Acceptance Rate represents the ratio of noises incorrectly classified as target speeches, while False Rejection rate represents the ratio of target speeches incorrectly rejected by the system. Lastly, Equal Error Rate (EER), in which the point where FAR and FRR is equal will be considered as the best settings for the classifiers. Through FAR and FRR evaluation, the robustness of the system could also be analyzed

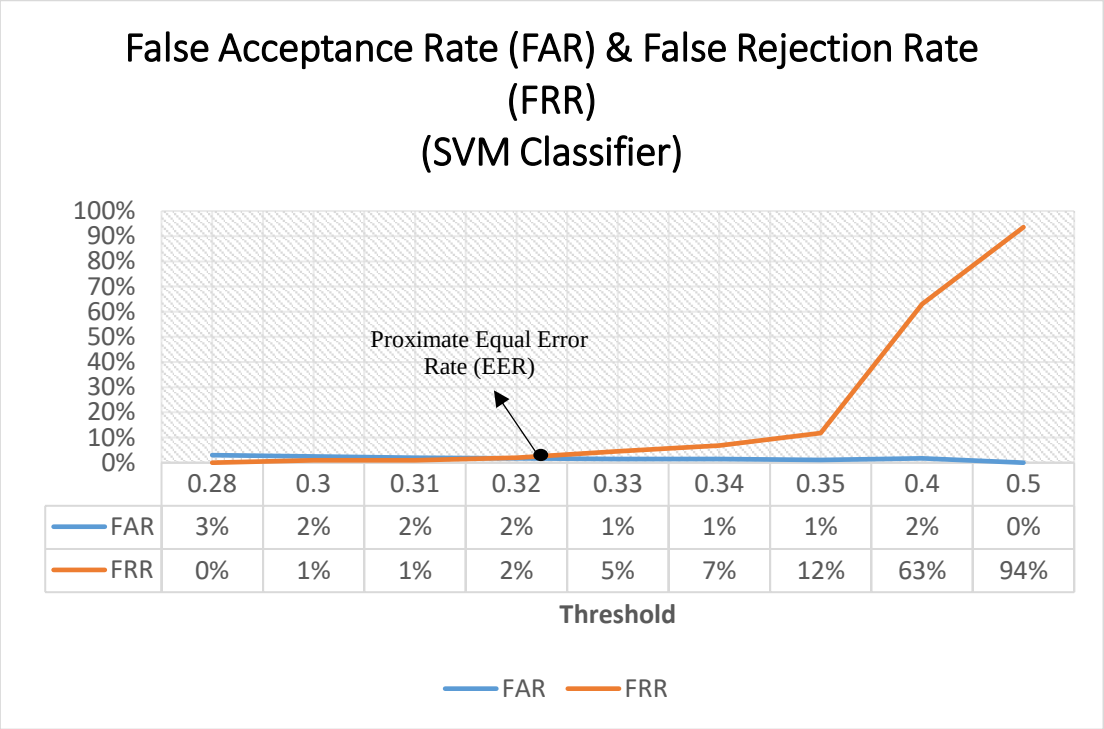

Figure 9. FAR and FRR results using SVM classifier

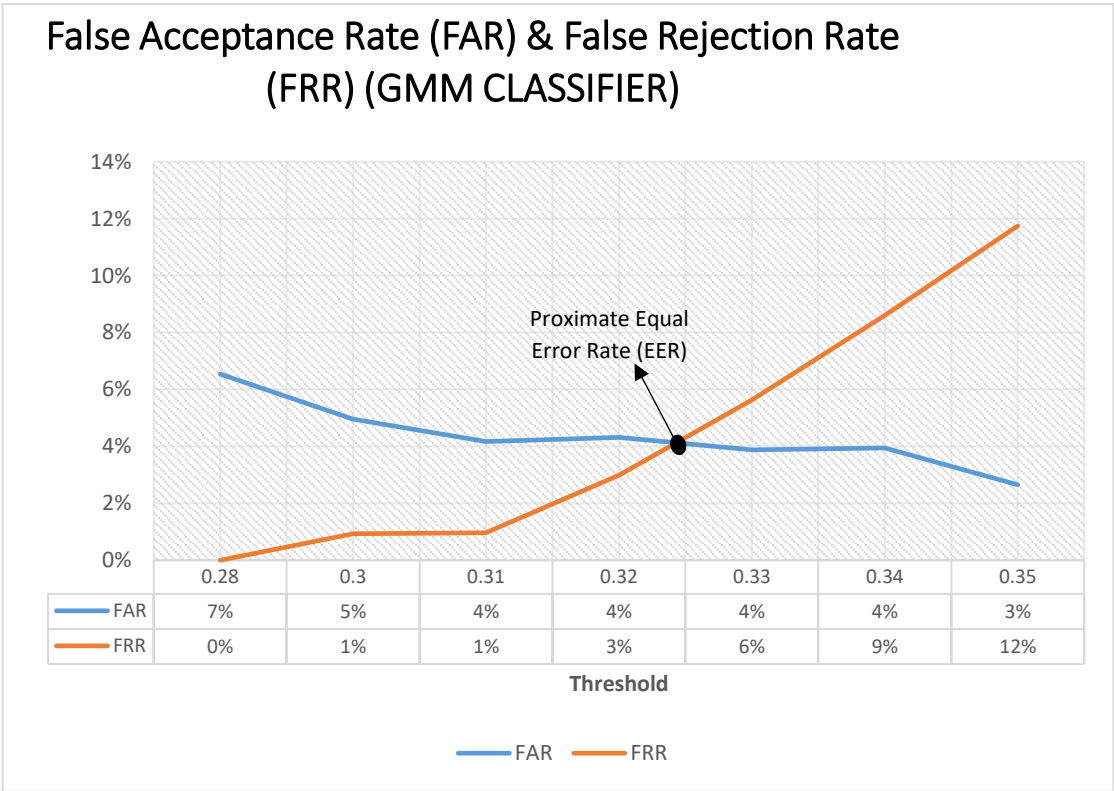

Figure 10. FAR and FRR results using GMM classifier

6. Result and Discussion

Based on preliminary experiments performed, GCC-PHAT method seems to be the most suitable method to be implemented to the proposed hearing support system. The method offers portability, low cost, and accuracy in estimating the direction of arrival of the sound signals. This shows the utilization of two channel microphones in the implementation of hearing support system in a

realistic room environment is feasible. However, in a more complex room environment, where a number of environmental noises are available, the current sound detection stage is not enough for distinguishing between target speeches and noise. Referring to the sound detection results, in a condition where the average SNR of the sound data less than 10 dB, most of the target speeches were not be able to be retrieved. As for sound data that have SNR more than 10dB, all the target speeches were in each case could be identified correctly, but in some cases, some noises, especially reverberated noises were also incorrectly identified by the system. Therefore, a further processing or method are needed to be included to the system, so that target speeches and noises could be distinguish correctly, and the system could produce an accurate notification system that consists of less false alarms. In addition, according to [11], in a realistic living room environment, even though we could retrieved SNR ranging from 20dB – 29dB for a normal talk, the system could produce a good performance even in a not realistic situation or condition (less than 20dB). From the results and SNR characteristic of the simulated test data, the system shows a could detect all the target speeches in a lower SNR (more than 10 dB).

7. Conclusion

In this research , we proposed a simple, portable, and robust a two-microphone based approach for detecting and identifying speech sounds in hearing support system, in assumption that it would offer a better experience for the hearing-impaired people in living their lives. From several experiments that have been performed to analyze and evaluate the performance of the system, the proposed approach and several additions added to the system had shown that the system shows a pretty accurate and robust to be implemented in a realistic room environment. In a constrained, quite room environment, Generalized Cross Correlation with Phase Transform (GCC-PHAT) shows promising results in estimating the Direction of Arrival (DOA) of the sound signals and detecting the available sounds in the environment at the time of recording. However, in a more complex environment, where some common room noises are included to the system, the GCC-PHAT seems to be deteriorated. It failed to detect the target speeches and incorrectly identified noises and consider them as target speeches. Consequently, if no further processing conducted after the sound detection stage, the system might produce a fair amount of false alarm and could give a bad experience and influence the hearing-impaired people. Therefore, here we proposed an approach to combine the system with classifiers. The addition proves that some improvements were occurred in the system and the system could produce a pretty accurate detection rate in a not ideal SNR environment with SNR below the normal realistic SNR (20dB – 29db) and more than 10dB. By adding Support Vector Machine (SVM) classifier to the system and with the limited number of training data, SVM shown a more better performance compare to the Gaussian Mixture Model (GMM). From the total of 274 target speeches retrieved in the sound detection stage, all the target speeches could be correctly identified as target speech, and for the total of 27 noise samples detected in the sound detection stage, 22 noise sample could be identified successfully

as noise. Thus, false alarm occurrence might be minimized to certain extent and the hearing-impaired people might be spared from having a poor sound notification system.

REFERENCES

[1] U.S. Department of Health & Human Services, National Institutes of Health, and National Institute on Deafness and Other Communication Disorders, "NIDCD Fact Sheet "Assistive Devices for People with Hearing, Voice, Speech, or Language Disorders," *NIH Publ. No11 -7672*, Dec. 2011.

[2] Florida Medical Hearing Centers, "Introduction to Hearing Aids," *Florical Medical hearing Centers*, 31-Jul-2017. [Online]. Available: https://floridamedicalhearing.com/main/about-us-2/. [Accessed: 25-Jul-2017].

[3] S. Jones, "Alerting devices," May 2017.

[4]. T. Shinyu, "聴覚障がい者のより自由で安心できる生活を求めて," 27-Jun-2017. [Online]. Available: http://www.shinyu.co.jp.

[5] 本多達也, "髪の毛で音を感じる装置「Ontenna」を世界中のろう者へ (JISA Digital Masters Forum 2016)," *JISA Q. Bull.*, no. 124, pp. 37–39, 2017.

[6] U. Kim and H. G. Okuno, "Improved Binaural Sound Localization and Tracking for Unknown Time-Varying Number of Speakers," vol. 27, no. 15, pp. 1161–1173, Jul. 2013.

[7] K. C. H. and C. G. C., "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans Acoust. Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[8] K. U, H. G. Okuno, and K. Nakadai, "Improved Sound Source Localization in Horizontal Plane for Binaural Robot Audition," *Appl. Intell. Springer*, Mar. 2014.

[9] HARK, "HARK Binaural + 2.3.0 documentation," 2014. [Online]. Available: http://www.hark.jp/document/2.3.0/packages/hark-binaural+/index.html#. [Accessed: 25-Jul-2017].

[10] D. O'Shaughnessy, "Invited Paper: Automatic Speech Recognition: History, Methods and Challenges," *Pattern Recogn*, vol. 41, no. 10, pp. 2965–2979, Oct. 2008.

[11] The Institute for Enhanced Classroom Hearing, "Poor Acoustics," 25-Jun-2017. [Online]. Available: http://www.classroomhearing.org/acoustics.html.