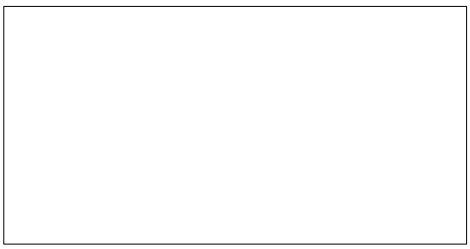
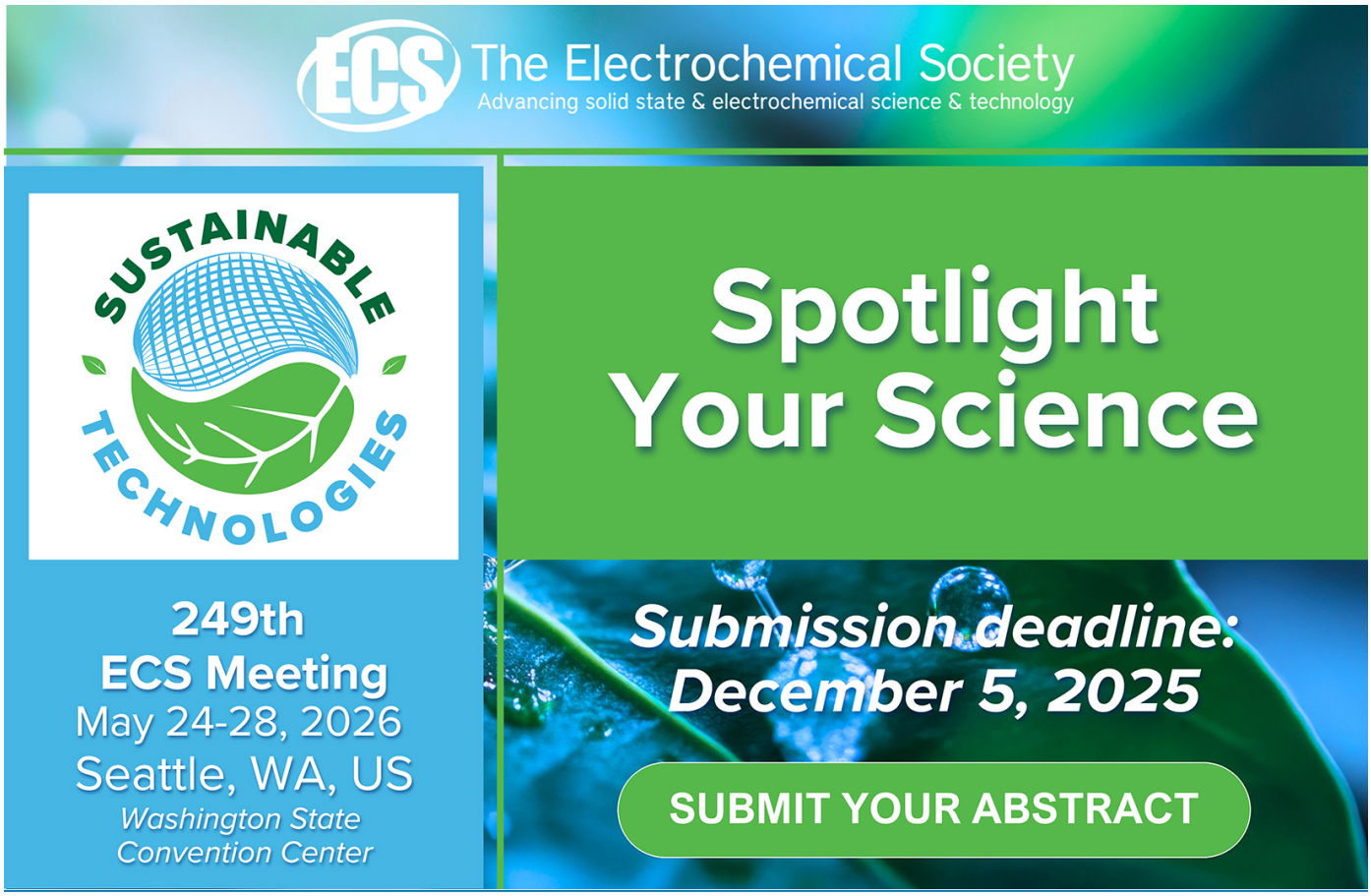


Training Based Noise Removal Technique for a Speech-to-Text Representation Model



Training Based Noise Removal Technique for a Speech-to-Text Representation Model

S Rashmi[1], M Hanumanthappa[2], B Gopala[3]

[1]Department of Computer Science, Kristu Jayanti College, Autonomous, Bangalore 560077, Karnataka, India, rashmi.s@kristujayanti.com
[2]Department of Computer Science & Application, Bangalore University, Bangalore 560056, Karnataka, India, hanu6572@bub.ernet.in
[3]Department of Computer Science & Applications, Bangalore University, Bangalore 560056, Karnataka, India, Gopala.nishanth@gmail.com

Abstract. The accomplishments of a Speech Recognition Process is significantly deteriorated by the presence of an unwanted speech signal called noise entity. This entity is present in the primary audio source. During the Speech – Recognition process, the presence of noise in the original audio signal adversely impact the output generated. Therefore, noise must be removed before performing any functions on the speech signal. With such observation of noise, it becomes essential to apply a unique procedure that perforce the noise without causing any distortion to the original audio. This research paper presents a novel approach to de-noise the given input audio signal based on the training method. Further, the paper explains the architecture adopted for Training Based Noise Removal Technique (TBNRT), steps of noise removal process, and the evaluation of the results obtained by the proposed procedure. The SNR values of the input are compared with the SNR values of the audio signal after applying the proposed TBNRT. Improvements in the SNR values were observed after the application of the proposed method. The obtained results were compared with the existing techniques and the proposed TBNRT gave promising results.

Keywords: *End Point Detection, Noise Removal, Training Based Noise Removal Technique (TBNRT), Praat System, Speech Recognition Model, Signal – to – Noise Ratio (SNR)*

1. Introduction

Audio / Speech signal is constantly prone to disturbances because of the surrounding environment. This disturbing entity is called Noise. Noise is an unwanted signal present in the required signal that causes disturbance to the main audio [3]. The presence of noise greatly affects the result in a Speech Recognition process [10]. Therefore noise has to be removed from the speech signal mainly to i) improve the SNR ii) conserve the meaning of the original noise with the identical speech characteristics without any distortion. Henceforth, the primary convergence of this research paper is on the noise removal process. Here, a Training Based Noise Removal Technique (TBNRT) is proposed for a speech signal and the results are evaluated. To measure the presence of noise, Signal to Noise Ratio (SNR) is used. SNR imparts on what is wanted over what is not wanted by comparing the signal power to the background noise. Thus, by the definition of SNR, it can be noted that the ratio of average signal power over the average noise power gives the SNR measure as shown in equation 1.

$$SNR = \frac{\text{Average Signal Power}}{Average\ Noise\ Power} \quad \text{--- [equation .1]}$$

The value of SNR defines the intensity of the noise present in the original audio. The SNR value must be relatively small and proportionate to 1:1. A high SNR value indicates that there is less noise disturbance and conversely, a less SNR value shows that the noise in the signal is high [14]. This is further subjected to the non-uniform pattern of the speech signal. For instance, consider a sound signal utterance of 1 second duration, uttered with some energy say E. A noisy version of the sound signal can be constructed by finding the noise sample (white noise) of the same signal and then scaling the SNR values to match the required result. Consider another example, where the continuous speech signal has 1 second of sound utterance and 1 second of silent interval making the total length of the signal to 2 seconds nevertheless the energy of the signal remains unchanged i.e., 'E'. However, to include 2 seconds of noise sample with the same energy as it was for 1 second, then the amplitude has to be reduced by almost 30% [1]. Thereby, the actual amount of noise in a sound signal depends on the length of that signal. This indeed includes the sound utterances and also silent intervals. Therefore, in this research work, only the sound segments are considered for the noise removal stage and it is assumed that the silent patterns have already been removed from the original signal.

Organization of the paper: The current research paper gives an overview of Noise Removal Stage in a Speech Recognition Model. Noise-free speech signal is achieved by the proposed Training Based Noise Removal Technique. The research paper is organized as follows: section 2 provides a brief idea about the existing noise removal techniques and the algorithms, section 3 describes the architecture of the proposed approach, section 4 explains the process of Noise Removal technique adopted in the current research work, section 5 shows the implementation of TBNRT using Praat tool and the results are evaluated using SNR value. In addition to this, the yielded results are compared with the existing methodologies. Finally in section 5, the paper concludes by briefly talking about the future epilogue for the current work.

2. Literature Survey

Noise is a varying entity because of its changing characteristics. Therefore it becomes extremely cumbersome to remove it from the original signal. Perhaps, due to this varying characteristic of noise, to devise one standard algorithm to remove it is difficult [7]. Some of the famous noise reduction algorithms are i) Modulation Detection: These algorithms differ for different users but ensure to provide better signal to the listener [Starkey laboratories, 2003] ii) Synchrony Detection: These algorithms look for energy level of the speech signal and compare them with the noise energy [Tyranski & Pogash (AAA 2004)]. Alcantara et al. (2003) observed that there were no improvements in the speech quality when the noise reduction feature was set on. Some of the algorithms that work on the above mentioned approach are Starkey Strata 312, Oticon Spirit II [Boymans & Dreschler (2000)] iii) Fourier and Hadamard Transforms [2] works on generalized transform domain and the matrix of feature points are used as transform locations. There are various Voice Activity Detection (VAD) algorithms as well whose primary target is to identify and separate the spectral features containing utterances and silence segments. This is also termed as End Point Detection. Different levels of decibels are considered for a VAD process such as -5db, -10db, 0db and so on. At these levels the SNR is calculated and the performance will be gauged as a ratio of Utterance Hit Ratio versus Silent Hit Ratio [9]. There are few Noise Estimation algorithms that are based on a-priori and covariance components. Some of the observation of these algorithms are as follows; i) if the SNR of signal power is low, then SNR value of noise residue will be high thereby reducing the overall SNR value of speech signal ii) If the signal power is too high, then the overall quality of the speech signal will be distorting leading to low quality output. iii) Consider End Point Detection before noise estimate algorithms are applied. This invariably reduces the amount of work needed to identify the noise in silent segments. iv) The power of the adjacent spectral components must be considered and a correlation must be established using the probabilities of power of the present and the neighbouring frames [16]. There are few pattern based noise detection algorithm techniques such as Dynamic Time Wrapping (DTW) and Hidden Markov Model (HMM). DTW ideally works on the transient noise by identifying them and removing the segment as a function of time. In HMM, the current speech spectrum is evaluated by looking at the previous spectral feature. The noise of previous type is propagated though making the model to learn the pattern and then remove it from the speech signal [5]. Authors Huy Toan Nguyen et al. [8] have worked on Noise reduction techniques using Kalman Filter and have been shown the noise removal approaches with respect to Gaussian filters. This has

been done with respect to feature extraction of the noise-prone input and the results of the values are evaluated using homograph matrix construction. The problem of linear filtering is addressed by Jesper Rindom Jensen et al. [11]. Instead of using the linear filtering technique for noise removal, the authors have shown the usage of variable span filters which greatly avoids the distortion caused due to the presence of noise. The minimum and the maximum SNR filter designs are considered. The noise signals are removed from the original speech signal using the filter designs and eigenvector approximation.

3. Architecture of the Proposed Method

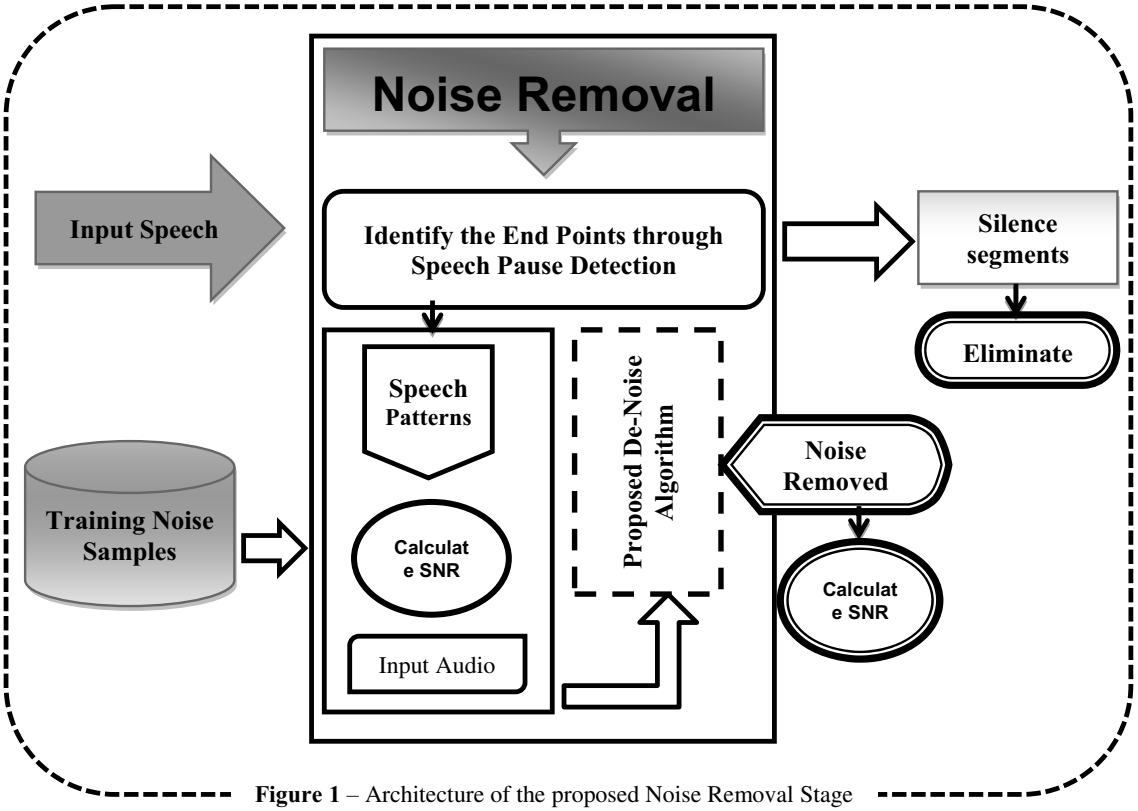


Figure 1 – Architecture of the proposed Noise Removal Stage

Figure 1 shows the proposed architecture of training based noise removal process. The audio/sound signal serves as the input to the system. The speech is continuous in nature and it is liable to many types of noise. The main objective of this paper is to remove noise from the original audio. In order to do so, the silent intervals present in the input speech have to be eliminated. To do so, End Point Detection algorithm is proposed ***. Once the silent segments of the speech are separated, the noise from the remaining signal is removed using the proposed Noise – Removal algorithm. Here, a set of all the noise patterns is collected and stored in the corpus and this data set serves as the training data. The speech segments of the original data are given as the input for the noise removal stage. The speech segments are compared with the noise samples present in the corpus. If there is a match, then the noise will be removed through de-noise function. This de-noise function, identifies the spectral features where the noise patterns are observed without affecting the primary features of the original speech. After the de-noise function, the output is evaluated for the efficiency in terms of SNR. This is compared with the SNR values of the original speech signal.

*** *The current paper is a part of the Speech-to-Text conversion in our research work. Earlier to this, an End-point Detection (EPD) algorithm was proposed and adopted in the research work. This EPD was based on identifying the frequency and the intensity levels observed in the input signal. Before an utterance, the sound levels and the frequency will be minimal (at least for few fractions of seconds); the intensity of that segment is identified and removed.*

4. Methodologies

In order to remove noise from a given input signal in a Speech-to-Text conversion, PRAAT, a phonetic tool has been used [6]. PRAAT, famously known as 'speech analysis by computer' is an efficient tool for analysis speech spectrum. It includes many features such as recoding of various types of sound, displaying it in different forms (table, grid, matrix and so on), analyze it in terms of periodicity and spectrum, sound signal annotate, convert it into formats such as intensity, stereo, mono and filtering techniques. The Praat system also includes many routines to implement signal processing, segmentation, and re-synthesis of the speech formants. Praat also provides aesthetic features for displaying the sound signals indicating pitch and intensity levels and also provides facility for drawing the same signal on a picture layout. Furthermore, Praat has 'remove noise' option under filter techniques. The tool efficiently removes the noise present in the input audio however; it can only remove the standard types of noises such as background hiss, and certain type of white noise. Further, it cannot remove environmental noise. This noise class is huge which includes so many different types of noises such as door tap, dog bark, strong wind or rain, baby cry, thunderstorm, crackle, other people on phone, someone's turning the switch on/off in the room while you are recoding the audio and many more. When such disturbing entities are found in the original audio, it becomes extremely difficult for the conversion of Speech-to-Text to take place efficiently [4]. Therefore, to remove these types of noises from the primary speech signal is the main objective of this research paper. This primary objective is achieved by the proposed Training Based Noise Removal Technique (TBNRT). According to the proposed technique, a noise class is created by collecting different types of environmental noises and manually keeping in the database as noise tag set. Each tag set in the database is saved with the specific name manually. There are around 500 different types of noise sundries that act as tag set. This tag set serves as the training data set and when the input audio is given for de-noising; the corresponding type of noise is matched from the tag set and is removed from the input data without tampering the original speech signal. As a rule, the user may input the entire speech signal for de-noising process [10]. Despite, sometimes the input file may be too long and it becomes time consuming for the entire file to be de-noised or the noise trials may not be found throughout the input data and it may be present only in some segments of the input speech. In such scenarios, the user can avail the option of click and drag in Praat system and select only those segments where the user finds presence of noise [12].

4.1. Experimental Steps for the proposed de-noising or noise removal step –TBNRT Approach;

- Feed the input file to the Praat system
- The user can manually select portions (Drag and drop) of the sound signal if the input file is too long and if the noise is present only in few segments of the input
- Select remove noise under filter option on the right panel of the Praat object system
- If the noise is a standard type, the Praat removes it automatically. However if the noise is not in boundary of Praat system, then the control goes to the trained data set.
- Search the test data and the trained data for a match. When the appropriate match or a closely appropriate match is found, the corresponding noise is removed using low and high band pass filter technique available in Praat.
    - *According to Low and High Band Pass filter technique [13], the segments which has a pattern match is identified and only those high band filter are marked and converts it into low band filter which has frequency less than the actual expected frequency of a sound file. Since the frequency of these segments whose pattern have been matched with the existing training tag set is reduced to a very less frequency number, these signals will be surpassed during the conversion as it would not be identified.*
- Once a noise is removed, a copy of noise-free signal file will be automatically stored with the original file name + de-noised.

The process of performing the noise removal process is explained in the algorithm 4.1 below;

Algorithm 4.1- TBNRT

Input: ∑Noise signal + Clean Speech utterance = Noisy Speech Signal
Hypothesis: End point Detection i.e., the speech utterances and pause intervals are separated
Output: Noise-free / Noise-suppressed speech signal

1) Start
2) For each noise utterance
    a. If noise utterance is identified by the Praat system
        i. Then use remove noise option of Praat System
        ii. Exit
    b. Else, initiate TBNRT – Go to Step:3
3) Compare the speech utterance with the training noise tag-set
    a. If noise pattern found,
        i. Remove noise using high pass filter and low pass filter technique of Praat system
        ii. Exit
    b. Else,
        i. Return the original speech signal to the user
        ii. Save the current noise type in the training database of noise tag-set
        iii. exit
4) Stop

5. Experimental Results

NoteTo implement TBNRT, initially a noise tag set was created. This tag set is the data repository containing different variants of environmental noises. In order to carefully consider the space complexity, only 3 sec of the noise types are collected from various sources available online. A small code is written to iterate the noise collected to match the length of the noisy input. Using the options 'intensity', 'pitch', 'formants' and 'frequency' available in the Praat system, each noise type can be transformed to have different intensity levels. For instance; the tag set of dog bark has any of the above mentioned factors with 'x' value, and in the input speech signal if the dog barks with the factors having 'x+1' or 'x-1' value [15], then the factors can be increased or decreased accordingly with the help of code. By doing so, the comparing parameters of trained and the test data will be identical or nearly identical. Figure 2 shows a sample noise tag set used in the proposed TBNRT. The noise tag set contains around 500 noise types collected from online sources. It includes noise types such as animal sounds, birds chirping, object destruction, musical instruments, and other environmental noise types.

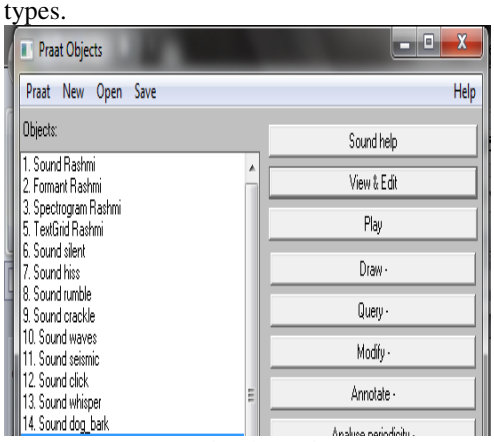


Figure 2 – Trained Noise Tag Set to implement the proposed TBNRT

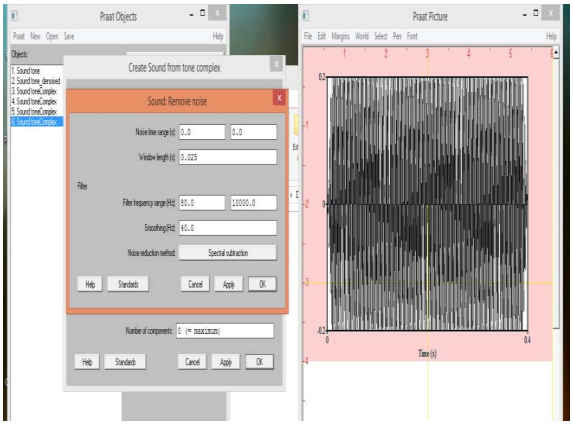


Figure 3 – Intrinsic Noise Removal option for a pre-recorded speech signal

As mentioned earlier, Praat system has an inbuilt noise remove option. This is shown in figure 3. After loading the input speech signal, there is an option 'filter' on the right side of panel. Under 'filter', there is an option to 'remove noise'. After clicking on 'remove noise', a new sound file will be created.

The proposed TBNRT is implemented in Praat system and the output of the implementation is shown in the figure 4. The trained noise tag set is searched for a match looking at the test data i.e., input file. When there is a match, those segments are eliminated from the original audio without tampering the quality. This is shown in the figure 5.

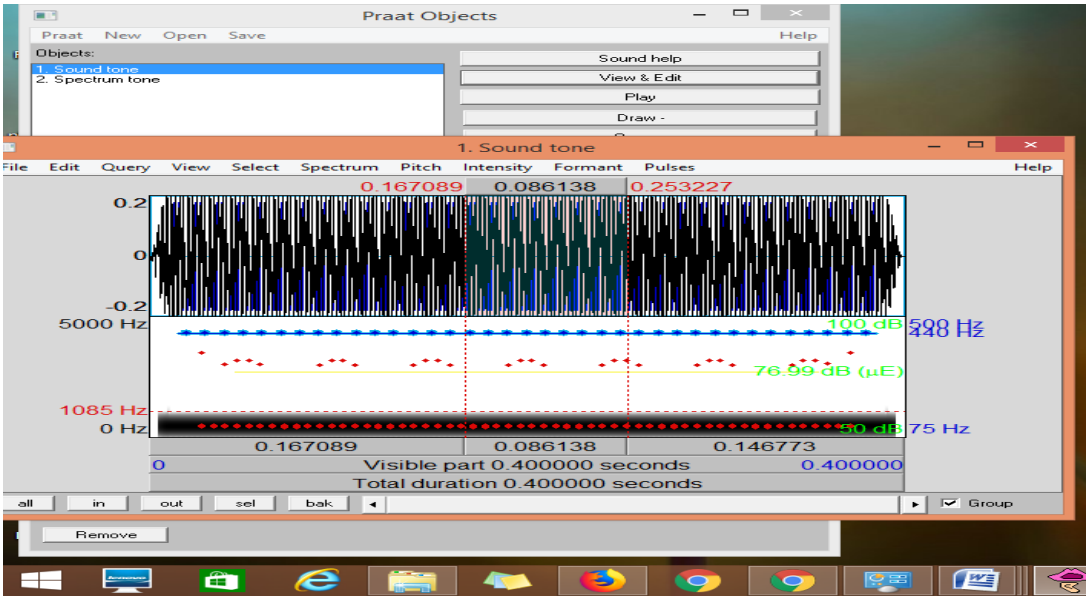


Figure 4 – Proposed Training Based Noise Removal Stage

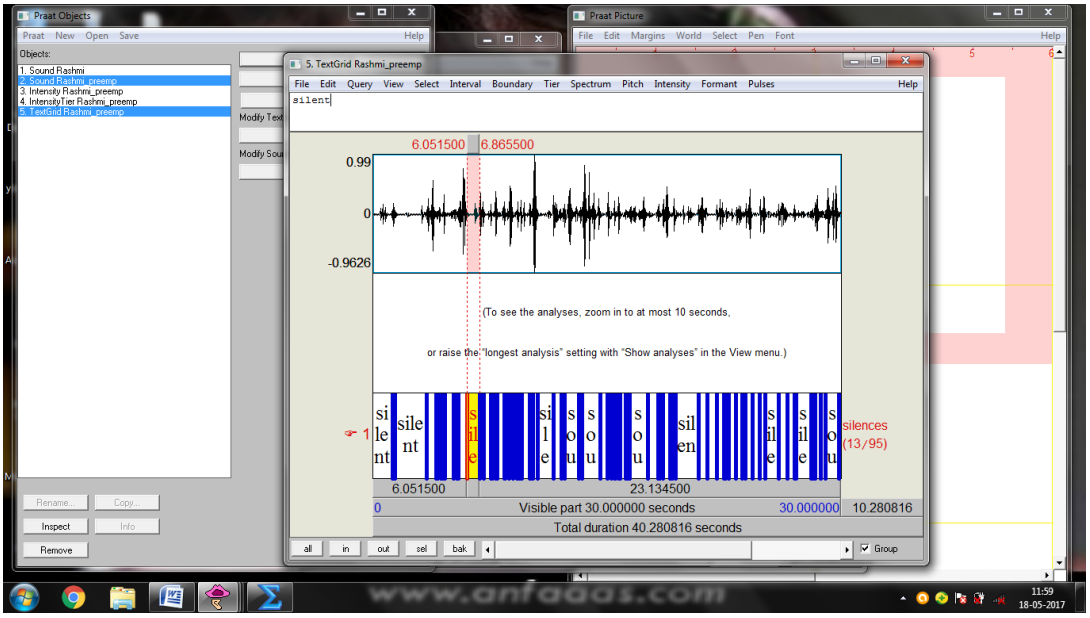


Figure 5 – Sample of a speech signal after TBNRT

Since the intrinsic method of Praat system removes only the standard noise types, the proposed Trained Based Noise Removal Technique (TBNRT) considers a bigger noise class. Different types of noises are trained and it becomes a training noise tag set. The tag set consists of various types of noises that can tamper the original speech signal. Each of the tag set is collected and stored in the repository of the Praat system.

The results are evaluated using the SNR values. This entity, as described in section 1, is an efficient method for calculating the efficiency of signal power. Initially, the SNR for the original speech signal without application of TBNRT is calculated and after feeding the input to the proposed TBNRT, the SNR is calculated and tabulated again. The comparative analysis of these two values shows a significant improvement in the original speech signal. The result of this comparison is tabulated in the table 1. The values of SNR were observed from the Praat system. Praat System evaluates the SNR using [equation.1] as shown earlier. The values of SNR were looked after and before the application of the proposed TBNRT.

Table 1. Experimental Results of the Proposed TBNRT evaluated for 10 samples chosen which contains different variants of noises

Name and length of the noisy input signal (in Millisecond)	**$SNR_{Original}$**	**SNR_{TBNRT}**
Sample 1 : 130 ms	20	170
Sample 2 : 145 ms	68	270
Sample 3 : 180 ms	190	245
Sample 4 : 200 ms	198	395
Sample 5 : 230 ms	207	432
Sample 6 : 270 ms	167	210
Sample 7 : 330 ms	193	352
Sample 8 : 410 ms	291	419
Sample 9 : 500 ms	301	506
Sample 10 : 630 ms	328	329

The graphical evaluation of the results obtained in the proposed TBNRT is put forward in the figure 6. It is evident from the table and the graph that, with the proposed TBNRT, the SNR values are higher indicating a better quality speech signal. The amount of noise present in the original audio is also reduced. The speech signal with better quality is stored for further processing of the Speech-to-Text conversion.

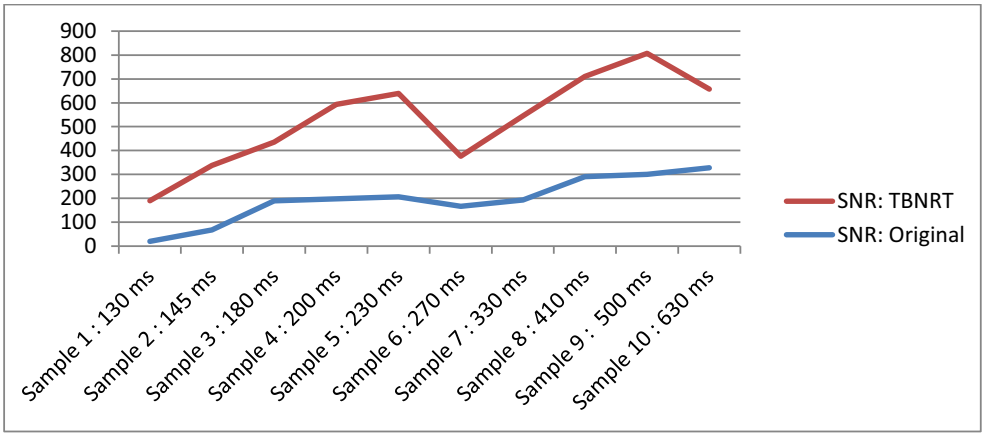


Figure 6 – Graphical representation of the result evaluation of the proposed TBNRT for values as tabulated in table 1

5.1. Comparative Study of the TBNRT with the existing Algorithms

The overall average of SNR values of the proposed TBNRT and the original speech signal can be calculated by taking the average values of table 1 and substituting in [equation.2].

$$\boldsymbol{SNR(average)} = \frac{\sum_{Sample\ i=1}^{N=10} \boldsymbol{SNR(i)}}{\boldsymbol{N}} \text{ --- [equation.2]}$$

From [equation.2], it was calculated that the average SNR for original signal was approximately 196 db and the average SNR of speech signal after applying TBNRT was found to be approximately 332 db. This average value of SNR for TBNRT is compared against the existing algorithms and the comparison study is indicated in table 2. It can be observed from the table 2 that the proposed TBNRT performs well when compared to the existing noise removal and noise estimation techniques.

Table 2. Comparative Study of existing approaches and the proposed TBNRT for Noise Removal

Method	**Average SNR values***
TBNRT	**332**
Non-TBNRT	196
Fourier Transform	201
A-priori Noise estimation algorithms	212
HMM	262
DTW	257
VAD algorithms	308
Kalman filter/Gaussian filter	292
Linear Filter	312

**-values are approximated and rounded off to get an integer value*

5.2. Challenges of the Proposed TBNRT

The proposed TBNRT process appears simple since the approach involves feeding the input to the system; a search is made in the trained tag set to locate an ideal match of noise type and finally arranging them into a noise cluster. Such conventional method is often observed in sound acoustics. Whilst the process of TBNRT appeals to be straight forward and it sure bespoke some challenges. The two main challenges are mentioned below;

- ***Vocabulary Size:*** The efficiency of TBNRT strongly mandates on the data tag set which is trained as noise cluster. The primary apprehension here is the size of this noise class. In real world, any sound that causes disturbance for the original data is bad. Identifying such noise types and storing it in the database could be tedious. The question of how big or how large the noise cluster should be is difficult to answer. For an approach like TBNRT whose accuracy depends on the volume of the trained data set, the large data is good. This in turn will be a hit on the vocabulary size since it demands lot of variations to be stored.
- ***Hybrid Noise:*** If the noise cluster contains data of a particular style, then it is sufficient to identify the noise type of the same kind. However, when the noises of more types are mixed, even though the individual noise type is present in the cluster, the system fails to identify it accurately.

6. Conclusion

The proposed TBNRT has given a speech signal whose quality is better and noise is removed. This approach is less complicated however efficient enough to yield good results. The suggested approach of noise removal shows that a noise of any variant can be removed robustly. Various test case samples include different types of noise and the noise types were removed with at most certainty. However, sample 10 contains a song clip. The motto was to remove the instrumental noise at the background but as seen from the table [1], the SNR values are almost same and the noise was not reduced. Further, when the sound sample with an individual sound type was included, then the noise was efficiently removed. The bottleneck of the proposed TBNRT lies here. More generally, the system cannot identify the hybrid noise which is a mix noise type from the existing cluster. The approach cannot be applied for the audio sample which includes songs, jazz and instrumental beats.

Future scope: In the future works, the proposed TBNRT can be applied for noise removal stage in End Point Detection of a continuous speech signal. The noise free speech signal can be used in the later stages of Speech – to – Text Model. Further enhancement to the proposed TBNRT algorithm would be to extend the techniques to efficiently identify the noise in the presence of background music. These include songs and music files. The limitation of the current technique is that the system is not capable of identifying the noise containing background music. Another direction of the investigation is on developing a speech synthesis which includes emotion identifier. For an emotion to be identified, it obviously demands for a noise-free data. So the proposed TBNRT can be implemented in such a system.

References

[1] Akbacak, M., "*Spoken Proper Name Retrieval for Limited Resource Languages Using Multilingual Hybrid Representations",* Audio, Speech, and Language Processing, IEEE Transactions on Volume: 18 , Issue: 6 Digital Object Identifier: 10.1109 /TASL .2009 .2035785 Publication Year: 2010 , Page(s): 1486 – 1495

[2] A. Zehtabian and H. Hassanpour, A Non-destructive Approach for Noise Reduction in Time Domain, World Applied Sciences Journal 6 (1): 53-63, 2009

[3] C. Barras et al. "Transcriber: development and use of a tool for assisting speech corpora production," Speech Communication, Vol. 33, pp. 5-22, 2001.

[4] Frances Alias et al, "Towards High-Quality Next Generation Text-to-Speech Synthesis:A multi domain Approach by Automatic Domain Classification",IEEE transactions on audio, speech and language processing, vol16,no,7 september 2008.

[5] G. Tamulevičius, A. Serackis, T. Sledevič, D. Navakauskas, "Bidirectional Dynamic Time Warping Algorithm for the Recognition of Isolated Words Impacted by Transient Noise

Pulses", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, No:4, 2014, scholar.waset.org/1999.4/9998049

[6] Hossan, M.A. et al. "A novel approach for MFCC feature extraction",IEEE Conference 2010,ICSPCS, 10.1109/ICSPCS.2010.5709752

[7] H.W. Park, A.R. Khil, and M.J. Bae, "Pitch detection based on signal- to-noise-ratio estimation and compensation for continuous speech signal," in Convergence and Hybrid Information Technology. Springer, 2012, pp. 767–774.

[8] Huy Toan Nguyen, "Implementation of Noise Reduction Methods for Rear-View Monitoring Wearable Devices", Journal of Software Networking, 113–136. doi: 10.13052/jsn2445-9739.2016.007 c 2016 River Publishers.

[9] Jasha Droppo and Alex Acero, Noise robust speech recognition with a switching linear dynamic model, 2004

[10] J. Morales-Cordovilla, N. Ma, V. Sa´nchez, J. L. Carmona, A. M. Peinado, J. Barker et al. "A pitch based noise estimation technique for robust speech recognition with missing data," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 4808– 4811.

[11] Jesper Rindom Jensen, "Noise Reduction with Optimal Variable Span Linear Filters", IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 24 , Issue: 4 , April 2016)

[12] M. Gales et al, " The Application of Hidden Markov Models in Speech Recognition Foundations and Trends in Signal Processing", Vol. 1, No. 3 (2007) 195–304 2008

[13] P. Taylor, A. Black and R. Caley "Heterogeneous relation graphs as a formalism for representing linguistic annotations," Speech Communication, Vol. 33, pp. 153-174, 2001.

[14] S. Cassidy and J. Harrington, "Multi-level annotation in the Emu speech database management system", Speech Communication, Vol. 33, pp. 61-77, 2001. http://emu.sourceforge.net/

[15] T. Moazzeni, A. Amei, J. Ma, and Y. Jiang, "Statistical model based SNR estimation method for speech signals," Electronics letters, vol. 48, no. 12, pp. 727–729, 2012.

[16] Urmila Shrawankar, Vilas Thakare. Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment. Zhongzhi Shi; Sunil Vadera; Agnar Aamodt; David Leake. 6th IFIP TC 12 International Conference on Intelligent Information Processing (IIP), Oct 2010, Manchester, United Kingdom. Springer, IFIP Advances in Information and Communication Technology, AICT-340, pp.336-342, 2010, Intelligent Information Processing.