Removal of Empty Bands from Speech Signals

Sukhdeep Kaur[1], Dr. Dharamveer Sharma[2]
[1]Department of Computer Science, [2]Head of Department of Computer Science
[12]Punjabi University, Patiala

Abstract- In order to provide speech analysis, speech production and perception can be exploited with the help of various speech processing applications. Few properties or features can be extracted from speech signal S(n) with the help of speech analysis. With the objective of simplifying the speech signal and for removing the redundancy present within the speech signal, the transformation of S(n) is done into another signal. In this research work, the novel technique is proposed to remove empty bands from the speech signal. The proposed technique is based on the threshold value for the removal of empty bands. In the proposed technique, the frequency of each band is calculated and a band which has frequency below threshold value is removed from the signal. The performance of the proposed technique is tested in MATLAB and it is analyzed that proposed technique performs well in terms speech enhancement.

Keywords- Empty bands, LDA

I. INTRODUCTION

The easiest and most commonly used mode by human beings to perform communication is speech. With the help of speech, the information can be exchanged in the most natural and efficient manner. Thus, natural language speech recognition is considered to be an important area of research today. The mechanism through which the speech signal is converted into a sequence of words using an algorithm is known as speech recognition [1]. Within signal processing, speech processing is known to be an interesting field. Designing an approach in which speech can be given as input to the machine and statistical modeling can be used to perform various operations is known as speech recognition. Within the applications that need human machine interface, automatic speech recognition has been used widely. For recording, interpreting and understanding the human speech in clear manner, several research approaches have been proposed by the researchers over the time. The conversion of text into speech in an automatic manner is known as Text-to-Speech synthesis (TTS). The computer is allowed to speak through this TTS technology. This process is very similar to the native speaker that reads the text in a particular language. The text that is to be converted is given as input to the TTS system and this text is further analyzed by a computer algorithm which is commonly known as TTS engine. The text is then pre-processed and few mathematical models are used to synthesize the speech. An output is generated in the form of sound data in an audio format by the TTS engine [2]. There are two major phases involved within the TTS process. The input text is converted into a phonetic or any other linguistic representation through text analysis which is the initial phase. Further, from this phonetic and prosodic information, output is generated which is known as generation of speech waveforms and is the secondary phase. Usually, high and low-level syntheses are the alternative names of these two phases. The data can be extracted from a word processor, a mobile text, or the newspaper and given as input text. A string of phonemes that has some additional information for providing right stress and duration is known as the character string and it is pre-processed and analyzed in the form of phonetic representation [3]. The information from high-level is used to generate the speech sound at the end using the low-level synthesizer. Over the past many years, the artificial generation of speech-like sounds is being done. A front-end and a back-end are the two parts that collectively generate a TTS system. Equally written-out words are generated from the raw text that also includes symbols such as numbers and abbreviations in the initial stage which is also known as text pre-processing or normalization. Further, for each word, the phonetic transcriptions are assigned at the front-end and the prosodic units such as sentences and phrases are generated by dividing and parking the text into prosodic units. The text-to-phoneme or grapheme-to-phoneme conversion is known as the process through which the phonetic transcriptions are assigned to the words. At the front end, the output is generated which is a symbolic linguistic representation that is made up by the phonetic transcriptions and prosody information. The symbolic linguistic representation is then converted into sound by the back-end which is commonly known as the synthesizer [4]. The computation of target prosody is involved in this part within particular systems. Speech synthesis can be performed in various manners depending upon the task to be performed. However, Concatenative Synthesis is known as the most commonly used mechanism since the most natural-sounding synthesized speech is generated generally by it. On the basis of concatenation of segments of the recorded speech, the concatenative synthesis is performed.There are certain steps followed within the TTS synthesis process. This text-to-speech synthesis will takes place in various steps. It gets text as input, which is firstly analyzed and then converted into phonetic description. Then it generates a a prosody. From the provided information, it produces a speech signal [5].The text is broken into token. The conversion of token to word creates the orthographic form. The token “Mr”, “Mister” is the orthographic form which is formed by the expansion of token. In same way the orthographic firm of “12” and “1997” is “twelve” and “nineteen ninety seven” respectively. When the text analysis is completed, pronunciation is applied. Letter cannot be converted into phonemes as the corresponding is not parallel always. In many cases, single letter can only corresponds to one phoneme or several phonemes. Additionally, many letters can correspond to a particular phoneme. In this dictionary based solution, as many morphemes or words can be stored in the dictionary. Full forms are made with the help of inflections, derivations and composition rules [6]. Sometimes, a full form based dictionary can be used in all the full forms of word are stored. It determines the pronunciation of those words which are not found in the dictionary.The application differs according to the size of their dictionaries. The dictionary based solution is much more effective than rules-based solution dictionary. Moreover, the dictionary-based solution is more accurate than the rule-based solution only if they have enough number of phonetic dictionaries available.

II. LITERATURE REVIEW

Long Zhang et.al [7] presented a two-stage processing scheme for single-channel speech disturbance and denoising the enhanced spectrum of the noisy signals. Several dereverberation and denoising methods have been proposed. These methods are based either on the reverse filtering or speech enhancement. In this, firstly the sound is recorded using noise power estimation method then the amount of disturbance is estimated using reverberant time. However, these methods often ignore the spectral structure of the speech signals, which is used to improve the performance. The main objective of this paper is to propose a two-stage processing scheme for single-channel deverberation and denoising using convolutive transfer function. Therefore, this concludes a well defined two- stage processing scheme for speech deverberation and denoising based on the CTF model.

Yash VardhanVarshney et.al [8] presents Non-Negative matrix factorization approach for the denoising of the speech signals. The existence of noise distortion in the present scenarios is almost unavoidable. These distortions are created because of the noise present in our surroundings and background noises. The main motive is to remove the stiffness related issues in the speech improvement signals. So many approaches have proposed to remove the speech distortions. During noise reduction, normally invalid models fail to explain the audio signals due to the moving nature of noise and the speech signal due to which supervised models handles this situation very efficiently.

Mehmet AlperOktar et.al [9] proposed a new speech enhancement technique using combination of wavelet notch filter and thresh holding. A speech recognition application has been developed and handles the undesired background sounds. Wavelet based methods are implemented for denoising the speech signals which are being corrupted. These noises can be removed by low passing filtering, which enhance the sharp signals. This wavelet based technique has been applied in many fields. There are so many problems related noise removal have been successfully solved by this technique. This method focuses on the improvement of speech quality of human listener or for the other speech processing algorithm. So, the study summarizes that, this new approach improves the speech signals corrupted with distorted background sounds and noises.

Jie Wei, et.al [10] proposed a low SNR signal which is combined with the wavelet thresh hold and Minimum Mean Square Error Short Time Log Spectral Amplitude Estimation (MMSE-LSA). The speech will we pre-processed by using MMSE-LSA and then the noise will be denoised in the domain of wavelet. This particular algorithm gives us better performances on the speech intelligibility and time domain, which is verified by the performed simulations. Adaptive Filter method need source noise in the form of reference which implements it in real world whereas; Wiener Filtering technique is not applicable to non-stationary noise environment, which will produce distortion and disturbance in the rest of the noise. Hence, the author concludes that the proposed methods enhances the signal through the MMSE-LSA ideas and shows better results by performing simulations.

Ryoji Miyahara, et.al [11] proposed a gain relaxation within the signal enhancement that has been designed for speech recognition along with a local noise source that is unknown. Softer enhancement of a target signal is done such that the potential degradation within the speech recognition can be removed by applying gain relaxation. The small undesirable distortion within the target signal components is responsible to distort the performance of this system. The signal enhancement over the input is achieved without causing any effects on the performance of clean speech. A commercial PC records the directional interference suppression with signals that are evaluated here.

Fahad Sohrab, et.al[12] reviewed an algorithm for denoising the single-channel speech signals. A non-negative matrix factorization is used for the speech and multiple noise dictionaries. There are different types of noises present in our environment degrade the quality and communication of the sound. MMSE technique is employed for the estimation of speech spectral amplitudes. The researchers’ have proposed a Recognize and Separate methods to clear the noisy and distorted signals. Finally, non-negative matrix factorization is used as the basic algorithm. The main objective is to separate the required speech signals from background disturbed noisy signals. Therefore, the author concludes that, in specific environments of noise, NMF gives better results than the algorithm modified for the generic noises.

III. RESEARCH METHODOLOGY

The MFC application is developed by using concatenation method for implementing speech synthesis. Within visual Studio ultimate edition, the MFC application is generated. Generally there are standard Windows applications, dialog boxes, forms-based applications, Explorer-style applications, and Web browser–style applications which are the five different types within which MFC executables fall basically. For its implementation, a dialog box is utilized. The Microsoft Foundation Class (MFC) Library is used as a base to generate an MFC application for Windows. The MFC application wizard is used for generating an MFC application in a very simple manner. Within the Visual Studio Express editions, the support of MFC projects is denied.

1. Pre-processing of text: The input text can be entered into a text box by the user. For entering the text, a dialog box was opened.

2. Segment entered text: Further, the entered text is segmented into words and then into graphemes within the next step. For any given language, the smallest unit of a writing system is the grapheme. Thus, the division of each word into its grapheme units is done here. For instance, the text “ ਇਹ ਪੰਜਾਬੀ ਦਾ ਵਾਕ ਹੈ ” is entered by the user.

Further, following words are generated by segmenting the text:

ਇਹ , ਪੰਜਾਬੀ, ਦਾ, ਵਾਕ, ਹੈ ,

Further, the segmentation of each word is done into graphemes as follows:

ਇਹ = ਇ , ਹ
ਪੰਜਾਬੀ = ਪੰ , ਜਾ , ਬੀ
ਦਾ
ਵਾਕ = ਵਾ , ਕ
ਹੈ

ਦਾ , ਹਾ is left to be similar within the graphemes list.

3. Database Development: The prerecorded sound units that are stored within the database are utilized by the concatenation synthesis of speech. Thus, the development of database is the most important requirement for implementing the concatenation method. Phoneme sounds are involved for the development of database. Each grapheme is recorded here to generate phoneme since for any language phoneme is considered to be the smallest speech unit.

- Selection of graphemes: Unique grapheme units of Punjabi language are identified initially for the development of database. For large number of unique words present within the Unicode format, the Punjabi corpus was formed. With the help of this, the graphemes of Punjabi were identified. Through the analysis it is seen that the database was developed by converting 830 graphemes to phoneme.
- Recording of phonemes: A native male or female Punjabi speaker records each grapheme. Also, the recording can be done within the studio at specific pitch, bit rate or at other prosody properties. Any kind of noise present within the recorded speech is eliminated with the help of Audacity which is software that helps in improving the quality of recorded speech. Each phoneme that consists of the following properties is recorded:

Project rate: 44100 Hz,

Channels: mono, 32-bit float
Labeling of phoneme: Careful and correct labeling of phonemes is important for performing concatenation of phonemes. A grapheme symbol is used to label each phoneme. For instance, the labeling of sound “ ਦੀ ” is to be labeled by the name itself, if it is recorded, such as “ ਦੀ .wav”. This is done because within the wave file format that includes .wav extension, this each recorded sound is saved.

4. Concatenation of Phonemes: For synthesizing the speech, the concatenation of phoneme units is done with the help of concatenation algorithm. C++ programming language s used to write the code. For instance, the phonemes of “ ਇਹ ਪੰਜਾਬੀ ਦਾ ਵਾਕ ਹੈ ” are found from the database. Further, for generating the speech through the concatenation of phoneme units, all the phoneme sounds are concatenated.

5. Remove Empty Band and Re-Generate Signal: The signal which is generated by the text to speech convert has noise or contain empty band which affect quality of the signal. The threshold based technique is applied which can remove empty band from the signal. When the empty bands get removed from the signal it leads to increase quality of the speech signal.

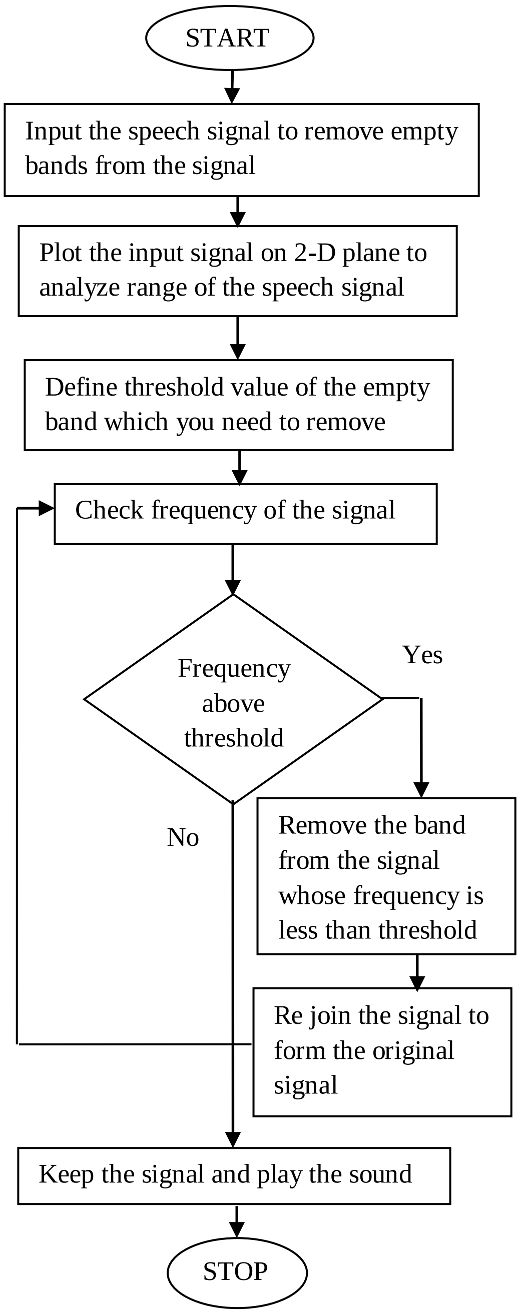
START
Input the speech signal to remove empty bands from the signal
Plot the input signal on 2-D plane to analyze range of the speech signal
Define threshold value of the empty band which you need to remove
Check frequency of the signal
Frequency above threshold
Yes
No
Remove the band from the signal whose frequency is less than threshold
Re join the signal to form the original signal
Keep the signal and play the sound
STOP

Fig.1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The proposed work is implemented in MATLAB and the results are evaluated as shown below.

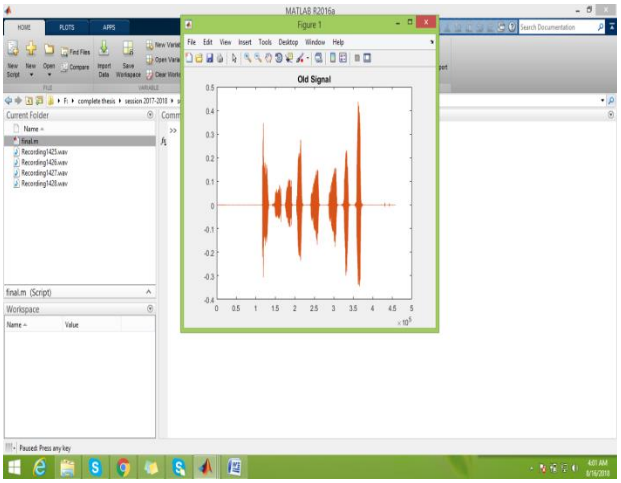

Fig.2: Signal with empty bands

As shown in figure 2, the speech signal which is generated with the text to speech converted is displayed in the figure. The generated speech signal has the empty spaces and we have analyze the threshold frequency of the empty bands

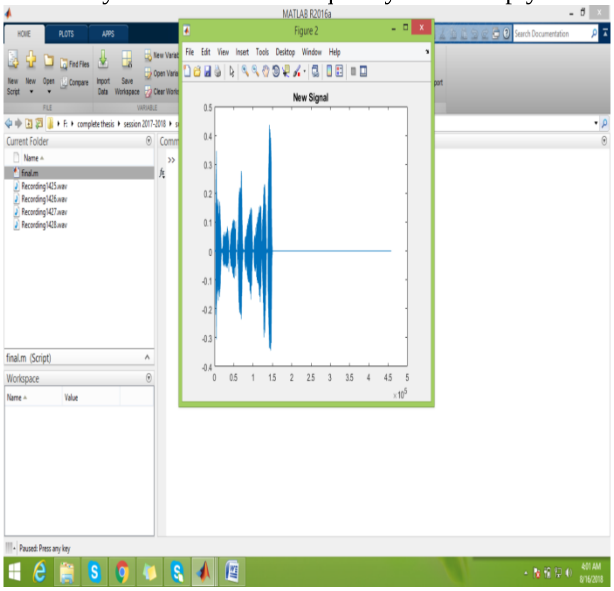

Fig.3: Signal without white spaces

As shown in figure 3, the speech signal which is generated with the text to speech converted is displayed in the figure. The generated speech signal has the empty spaces and we have analyzed the threshold frequency of the empty bands. The frequency which is above the threshold value is kept and all other frequency bands will be removed from the speech signal.

V. CONCLUSION

Natural language speech recognition is considered to be an important area of research today. The mechanism through which the speech signal is converted into a sequence of words using an algorithm is known as speech recognition. In this research work, the text to speech converter is designed which can convert the Punjabi text into Punjabi signal. It is analyzed that in some text to speech converters the empty bands are present which affect its efficiency. In this work, the module is added which can remove empty bands from the signal. To remove empty bands from the signal, the threshold based technique is proposed. In that technique, the bands which have high frequency than the threshold value is kept and all other will be removed from the signal. The Proposed algorithm is implemented in C sharp and MATLAB. The MATLAB is used to remove empty bands from the input signal.

VI. REFERENCES

[1]. Stolbov, M., Aleinik, S., “Speech enhancement with microphone array using frequency-domain alignment technique”, 2014, Proceedings of 54-th International Conference on AES Audio Forensics, pp. 101–107

[2]. Prudnikov, A., Korenevsky, M., Aleinik, S., “Adaptive beamforming and adaptive training of dnn acoustic models for enhanced multichannel noisy speech recognition”, 2015, Proceedings of 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015), pp. 401–408

[3]. Deng, L., Droppo, J., Acero, A., “Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise”, 2004, IEEE Trans. Speech Audio Process. 12(2), 133–143

[4]. Korenevsky, M., Romanenko, A., “Feature space VTS with phase term modeling”, 2016, Speech Comput. Lect. Notes Comput. Sci. 9811, 312–320

[5]. Hirsch, H., Pearce, D., “The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions”, 2000, Proceedings of ISCA ITRWASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium

[6]. Tomashenko, N., Khokhlov, Y., “Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing”,2014, Proceedings of Interspeech, pp. 2997–3001

[7]. Long Zhang, Jiaxuehen, Y ou Luo, Jiafei Fu, Zhongfu Ye, “Supervised Single-Channel Speech Dereverberation and Denoising Using a Two Stage Processing”, 2017, IEEE

[8]. Yash Vardhan Varshney, Z.A. Abbasi, M.R. Abidi, Omar Farooq and Prashant Upadhyaya, SNMF Based Speech Denoising With Wavelet Decomposed Signal Recognition”, 2017, IEEE

[9]. Mehmet AlperOktar , Mokhtar Nibouche , Yusuf Baltaci, “Denoising Speech by Notch Filter and Wavelet Thresholding in Real Time” , 2016, IEEE

[10]. Wang “MMSE-LSA based Wavelet Threshold Denoising Algorithm for Low SNR Speech”, 2016, JieWei , Ming IEEE

[11]. Ryoji Miyahara† and Akihiko Sugiyama, “Gain Relaxation: A Useful Technique for Signal Environment with an Unaware Local Noise Source Targeted At Speech Recognition”, 2016, IEEE

[12]. Fahad Sohrab, Hakan Erdogan, “Recognize And Separate Approach For Speech Denoising Using Non Negative Matrix Factorization”, 2015 , IEEE

Sukhdeep Kauris a student of M. Tech (CSE) in a department of Computer Science, Punjabi University Patiala carrying out her research work under the guidance of Dr. Dharamveer Sharma. Her main research interest is Speech Smoothing for Punjabi Language.

Dr.DharamveerSharma Ph.D. (Computer Science), MCA.Presently serving as Head of Department of Computer Science,Punjabi University,Patiala.His key research areas are Optical character recognition, Natural language processing, general computing. He has published more than 100 research papers in journals and conferences of repute including IEEE, ACM, Springer etc. He has developed various software and websites, prominent among those are his own websites, Unicode Based Punjabi Language Utilities, Result Processor (OMR), Offline On campus Admission counselling system, Encyclopedia of Sikhism and many more.