Designing and creating Punjabi Speech Synthesis System Using Hidden Markov Model

Harsimarjeet Kaur[1], Dr. Parminder Singh[2]

[1]Master of technology in Computer Science and Engineering, GNDEC Ludhiana, Punjab, India
[2]Professor & Head of Department in CSE, GNDEC Ludhiana, Punjab, India

Abstract - *In this research discussed the various method to collect the Punjabi text and method to record that Punjabi text into speech. To generate the text to speech system most important is to build speech corpus. Large collection of data consist Punjabi text. We have collect Punjabi text from various domains like financial, government, current news etc. along with pre-build dictionaries. Whenever whole data related to Punjabi text had collected, then built a system that converting the text data into speech. We have extracted all related features that needed to generate speech. Database has built that contain whole features which will further used to produce a complete utterance. The system is based on phonemes so also called phonemes based text to speech generation system for convert Punjabi text into speech. The system contains offline and online phases. The database creation process is covered under offline phases while system that converts text into speech covered under online phases. The wave file will be generated when we write any Punjabi word into text work space in designing phase. Hidden Markov Model used to design TTS system which is statistical parametric speech synthesis system.*

Key Words: **TTS, SPSS, HMM, GRSS, NLP, AI**

1. INTRODUCTION

Natural Language processing (NLP) is branch of Artificial Intelligence (AI) that encourages machines to understand, interpret and control the language spoken by human being. There is gap between communications with computer by human language so to fill that gap; NLP draws many rules including computational linguistics and computer science. NLP has number of applications are: Speech synthesis, Speech recognition, Information retrieval, Machine translation and Text summarization, Sentiment analysis, Machine transliteration, Automatic summarization [1].

The idea of Natural language processing is to build and design a computer machine that will understand, analyze and produce natural human languages.

There are five components of NLP which are listed below and described every component by step by step in Fig. 1.1.

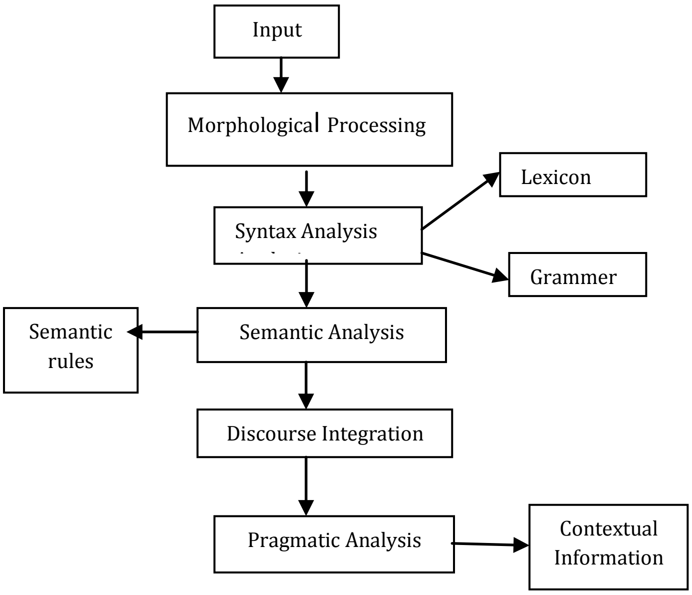

Fig. 1.1 Components of NLP

1.1 Speech Synthesis

The computer system that produces human voice is known as Speech Synthesis Machine. Because it convert the normal text into sounds or speech that is why it is also called TTS means Text to Speech System or TTS machine. A TTS synthesis system converts written orthographic text into corresponding artificial speech signals. A speech synthesis system is divided into two parts. The first part accepts the information in the form of text and gives output as a symbolic representation. The second part takes the output from the first part as input and converts into a synthesized waveform as output. The speech synthesis systems generate the output that usually refers to how much similarity and naturalness of output sounds with speech of real person. TTS based on three parameters that are Intelligibility, Naturalness and Accuracy. Intelligibility means normalizing of style, Naturalness means real that not influenced by other and Accuracy means correctness and reliability [3].

2. Statistical Parametric Speech Synthesis (SPSS)

The main advantages of SPSS from traditional synthesis technique are that it has more flexibility to change the characteristics of voice and support more multiple languages i.e. multilingual, has good coverage of acoustic space and robustness. It generates high quality of speech from small training database. The disadvantage of SPSS is degradation in quality of speech sounds. The advantage of this technique is that it has capability to produce intelligible speech with small footprints. Deep Neural Network (DNN) and Hidden Markov Model (HMM), Gaussian Process regression (GPR) based speech synthesis are main speech synthesis systems in Statistical Parametric Model. The speech that generated is belongs to parameters rather than exemplar so, these models are also called parametric models. It is statistical because using statistics like means and variances of probability density functions it defined the parameters and those parameters which capture was found by the distribution of parameter values in the training dataset [7].

2.1 Hidden Markov Model

The main purpose of Synthesis system is to generate synthetic speech with having high quality. The HMM system contains two phase:

i) Training Phase
ii) Synthesis Phase.

The fundamental frequency such as vocal source, duration such as prosody and Frequency Spectrum such as Vocal tract all features are modeled simultaneously by HMMs. The Maximum likelihood Creation Algorithm (MLCA) used by HMM to produce speech waveform. In training phase, the features like spectrum and excitation parameters are extracted in training phase. These features are take out from speech database then modeled those features to convert HMM into context dependent HMM. A model usually consists of three states that represent the beginning, the middle and the end of the phone. In Synthesis phase, according to give text the speech signals are concatenated then waveform is generated as output. Speech synthesis filters are used to generate the speech waveform. MLSA etc. filter is used to produce speech sounds. Both the training and synthesis part which shows how statistical parametric speech synthesis system works. Training part contains features like spectral and excitation. All are stored in database and next step is train the model. In synthesis the speech will be generated using that features that already stored in database. The merit of HMM based Speech Synthesis techniques that it is very useful for many languages with very small modification and voice characteristics can also be modified very easily, as compare to other traditional speech synthesis system HMM is better. Another advantage is requirement of less amount of database to produce good quality of emotional speech or different type of speaking styles [8].

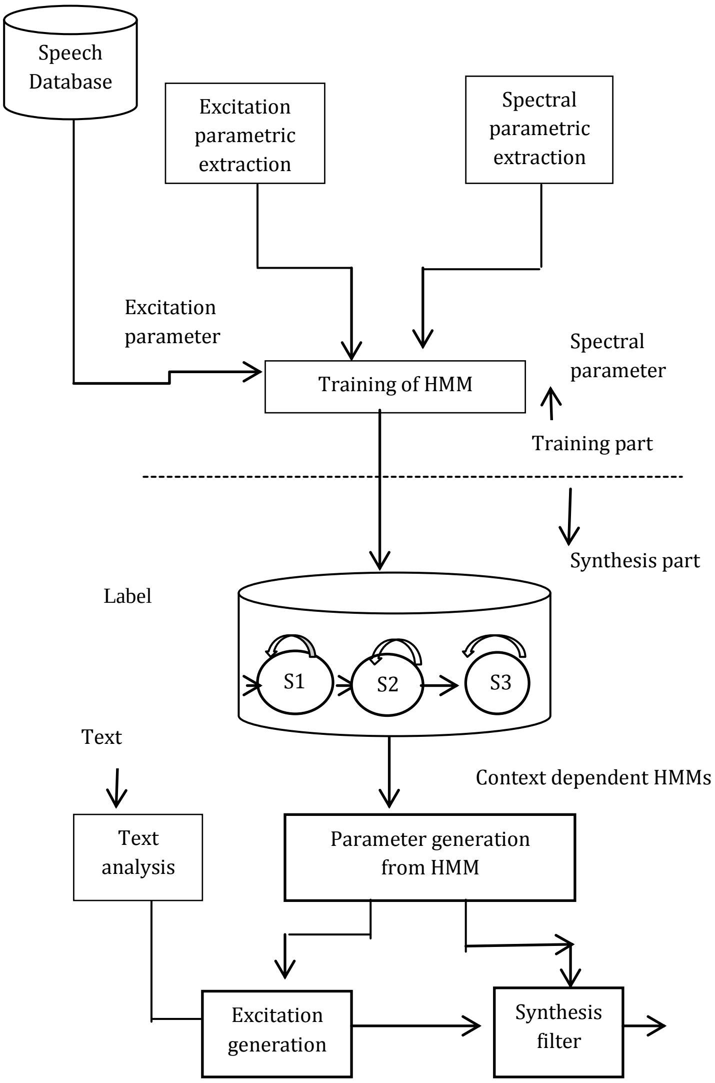

Fig 1.2 Statistical parametric Speech Synthesis

3. Speech Corpus for Punjabi Language

3.1 Phoneme

The phoneme is act as smallest unit of any language. On the basis of some rules phonemes are combined to form the valid phonemes combination. Phoneme sound is the sound of any language ordinals. For Punjabi language phonemes are divided as: segmental and supra-segmental phonemes. As similar with other languages Punjabi language also include only segmental phoneme in its alphabets but not segmental phonemes. The main base of any language is phoneme. Phonemes are act as isolated unit in any language. When we use the basic unit as phoneme then very small amount of data is needed that will cover the all possible phonemes of language. But in case when phoneme is elected as basic unit for synthesis, then there is more chance of discontinuities at point of concatenation of phoneme and also improper energy distribution. The segmental phonemes have its independent presence and are divided into vowels and consonants, in which there are total 20 vowels and total 38 consonants. Segmental phonemes units have their own independent sound, which can be differentiate easily.

3.2 Diphone

The adjacent combine of phones, in phonetics are called diphone. Now for the calculation of Diphone the basic formula is that if z is total number of phonemes in language, then z^2 is the total number of Diphone in language. But some languages have constraints due to which Diphone for each language is mainly smaller than z^2. The speech synthesized by the combination of Diphone is similar to natural human sound as compare to the synthesized by the combination of phonemes. Diphone is bigger unit as compare to phoneme The justification behind using the combination of adjacent phones known as diphone, is that the accent apprehension at ``center'' is the most steady section, whereas the shift from one fragment to an additional contains the phenomena and it's very inflexible to model. So, the diphone do the cutting of the units at the points of comparative stability, instead of phone-phone unstable changeover.

3.3 Syllable Units

Syllable is verbalization unit which having single vowel sound used with or without any consonants. The depiction of syllable for any language is very hard chore. Many languages are syllabic languages, Punjabi is one of them. For Punjabi language the seven form of syllable are known: V, VC, CV, VCC, CVC, CVCC and CCVC. Here V is vowel and C is consonants. The syllable is grouped into monosyllable, trisyllable, disyllable and polysyllable. Monosyllable which means word includes the one syllable unit only. Disyllable means two units of syllable are included in word. Likewise, trisyllable means three syllables form the single word.

4. LITERATURE REVIEW

D. Siddhi et al. (2019) survey on different method of Text to speech Synthesis. This paper provides general overview of various speech synthesis methods. It categorized the three techniques in paper that are formant, Concatenate and Articulatory based speech synthesis. Cascade, parallel, Klatt and PASCAS techniques covered under Formant based synthesis. Vocal-tract, Acoustic model, Noise source model covered under Articulator speech synthesis. Diaphone, corpus and hybrid covered under concatenate based. It also described various unit selection techniques like SPSS. Unit selection method depends on target cost and concatenated cost. Pitch synchronous Overlap Add PSOLA Linear Prediction Pitch synchronous overlap Add LP-PSOLA etc. described briefly under Di-phone synthesis. Prosodic, preprocessing and pronunciation is still a field having needs to work much and need more improvement to generate speech that is more natural [6].

D. Qinsheng et al. (2019) made a survey on Articulatory Speech Synthesis. It will provide complete knowledge about Articulatory system. The main articulators are tongue, jaw and lips. Rather than that there are also some vocal tract parts. Pulmonic voices are generated when breathing and muscles act as source of energy. Lungs will store oxygen; vocal folds separated the vocal tracts from lungs. The process is called vocal chords. Then speech will generated from vocal folds, filtered by vocal tract then finally spoken words coming through mouth or nostrils. Vocal tract and organs that produce speech will describe in this research and introduced different method to generate Articulatory Speech. Synthesis methods are divided into two types that are static and dynamic methods. The three goal of generating Articulatory System synthesis accuracy compared with number of speakers, naturalist and intelligibility. It is used for speech synthesis and speech reorganization both purpose. Vocal tract geometry depends on Base geometry, vocal tract movement, movement generation, collision handling. Non-linear sound generation model is to be built in acoustic synthesis [5].

J. Parker et al. (2018) described speech synthesis system for adaptation of an expressive single speaker using technique named as deep neural network. This research pursues to attain adjust an expressive single speaker DNN text to speech model to generate a novel speaker using only neutral speech. Generating good quality speech can be challenging and costly task, so it is useful to adapt an expressive single speaker model. The main advantage of SPSS, that it has ability to speakers and expression very easily. The linguistic features are normalized using DNN regression and then converted into normalized acoustic model.it work on single speaker so it's very challenging to add another speakers and single speaker generate speech that are unclear. Experiment was done under the quantitative evaluation and qualitative evaluation. Mean scoring, speaker similarity are checked under qualitative evaluation. Quantitative evaluation checks the distortion on some test data set between synthesis speech and natural speech [7].

X. Wang et al. (2018) purposed F0 model for SPSS. The F0 is an important acoustic feature of the speech waveform. It is supposed as pitch and carries the tone and intonation in word utterance. This model condenses the F0 data of earlier frames in the F0 series and modifies the mean of the present F0 distribution with the help of a linear function. Text to speech system built from F0 frequency using recurrent neural network. It described the statistical acoustic model .i.e. SAD which is equivalent to combinational trainable linear filter and RNN. Dependencies of F0 contours are check with deep acoustic model (DAM). Result in this research give DAR model having good accuracy and less over-smoothing. Evolution test .i.e. mean opinion score is calculated as subjective testing. This paper will compare DAR, SAR and Baseline models [8].

K. M. Khalil and C. Adnan (2018) mentioned HMM model to generate TTS system to process the Arabic language. The system was based on phonemes that uses as HMM synthesis unit. Arabic language has five syllable arrangements: CV, CW, CVC, CWC and CCV, where C signifies a consonant, V signifies a vowel and W signifies a long vowel. The main motive of this system is to build TTS that maintain coherence to combine text using concatenating HMM phonemes. It helps to improve quality of waveform that will be generated. The main advantage of using these techniques is that it contains less error rate when training and testing the dataset. STRAIGHT vocoder is used for this purpose. For the manufacture of waveforms equivalent software is the HTS engine version 1.05 was used. This software is self-determining from the rest of the tools and permits to creating waveforms from HMMs [9].

5. Problem Formulation

To establish any TTS system there is need to deep study the various techniques that are made already. There are so many approaches that are used to generate speech from text input based on different Indian or Foreign Languages and these approaches also contain some kind of limitation due to which performance get degraded. To overcome those problems some news techniques are available to generate more natural speech. So, statistical parametric speech synthesis approaches uses to overcome these problems because this approaches work for multi-language only for some changes.

The basic techniques of speech synthesis were:

Concatenative TTS is very restrictive, because of large data requirements and development time. So a more statistical method was established rather than the traditional brute force techniques. This technique generates sound by joining the various parameters like fundamental frequency i.e. F0, magnitude of spectrum etc. and with the help of those parameters generate speech.

Articulatory synthesis was based on the Human Articulation system. It built to generate the speech same as human speech production system. Human Articulation system contains vocal tract system, various articulatory organs like lip, tongue, jaw etc. and articulatory processes generate the speech directly. This technique is the toughest technique to implement because of having limited knowledge about the complex human articulation system. So, to generate such system needs to understand the whole biological working of human articulatory system.

Formant synthesis is based on those rules that label the frequencies of the vocal tract. This technique contains the source filtration model to generate speech. Filters were used to model the speech. Rule based formant synthesis generates speech which quality of sounds was unnatural. Hence, it is tough to take the estimate source parameters. The demerit of these techniques are that to build such system large amount of data is needed and to implement system need more time.

To address these problems we are using hidden Markov model or deep neural network because it does not contain large amount of database. To implement such a model and develop a search algorithm that finds similar documents written in Punjabi text (Gurumukhi script).

6. Results of Features Extraction

In speech synthesis and speech recognition feature extraction is considered as most important part to separate the one speech from another. Because of individual characteristics of every speech utterances and an extensive range of feature extraction approaches are used to extract those characteristics. These characteristics depend on some criteria i.e. feature extraction in speech should be easily measurable, should not be mimicry susceptible, do not get affected due to the change in speech environment and must be stable over the time. The main agenda of extracting features means gathering some important information in form of feature vector that depends on the behavioral tracts, Vocal tract and excitation source. There are possible procedures exist to represent the speech signal parametrically. These are Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), Cepstral Coefficients using DCT and many more. One of the best common and popular methods is to using MFCC. Before extracting the MFCC coefficient, there are few initial steps taken that are Frame- blocking, windowing, FFT, Mel frequency wrapping, Cepstrum etc. The original wave file time domain representation is in Fig. 1.2.

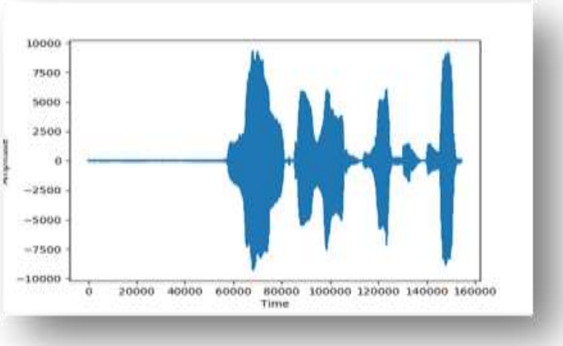

Fig. 1.2 Plot the Wave Representation of Sound Speech

6.1 Pre-emphasis

For amplify the frequencies in higher level, the filter is used called pre-emphasis filter. Lower frequencies usually have larger magnitude as compare of higher frequency. Fourier transformation operation used to overcome some numerical issues. The ratio between signals to noise can be improved using pre-emphasis.

When pre-emphasis is completed, converted the signal in short term frequency frames. Reason behind this method is that the frequencies are always vary after some time in signals, so in various cases find the Fourier transformation over time is not making any sense. To overcome this problem, assume the frequencies over time carefully or securely.

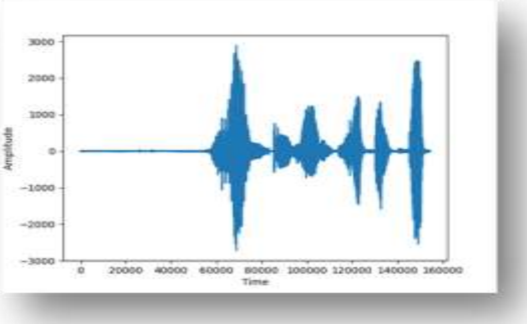

Fig. 1.3 Plot the Pre-emphasis Wave Representation of Sound

6.2 Framing

Signals are concatenating into accent frames and decent approximation of frequencies is generated. Speech signals are divided into frames and every frame contain some length between amplitude. If frame length is too large then all spectral parameters will not be captured but if it is too small then resolution should be degraded. First frame is connected to next, next frame connected to its previous and next so on. So frames are overlapped with each other.

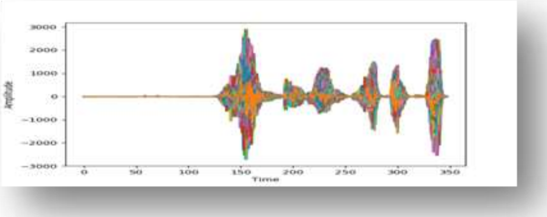

Fig. 1.4 Plot the Frame Representation of Sound

6.3 Windowing

The next step is hamming window function will applied to slicing the signals into frame. A narrow main lobe and a low side lobe are contained by window function. The basic idea is to getting the smooth edges with minimum discontinuities. Hamming window is used in speech processing field. In order to pretend the continuity of first and endpoint of frame, every frame will be multiplied with hamming window. In each frame signals are represented as s(n) where n=0, 1, 2……N-1. Then signals are multiplied with hamming window i.e. s(n) * w(n), where w(n) is hamming window which is expressed in equation 4.1.

$$w(n,\alpha) = (1-\alpha) - \alpha\cos\left(\frac{2\pi n}{N-1}\right),$$

$$0 \le n \le N-1 \qquad \ldots(4.1)$$

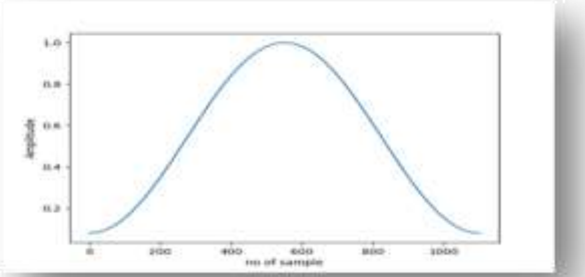

Fig. 1.5 Plot the Windowing Representation of Sound

6.4 Mel frequency wrapping

The way in which human can perceives frequencies is based on Mel Scale and 1000 mel scale will be equal to 1000hz.

The mel scale expressed in equation 4.2.

$$mel(f) = 2595 \times \log_{10}(1 + f(100)) \qquad \ldots.(1.1)$$

Where f be actual frequency and mel(f) be perceived frequency.

Filter band have triangular band pass frequency.

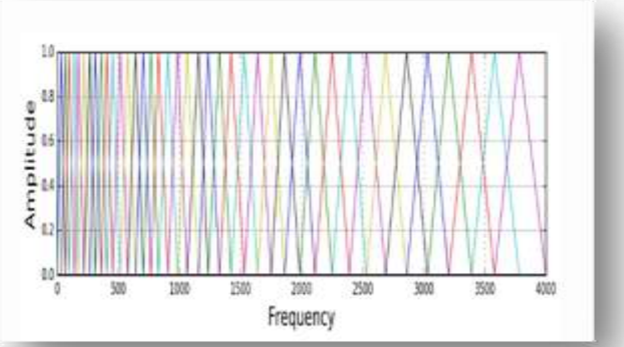

Fig. 1.6 Plot the Mel Filter Bank

The system will be tested to verify and validate it. We selected the 10 persons from my college and type some words in system and taken their feedback. Feedback back is depends on 5 factors: 1: wrose, 2: Poor, 3: Fair, 4: Good and 5: Excellent. Most of them rated it Good position. Because TTS system generated for punjabi language is somewhere difficult, synthesized speech still having limitation like naturalness.

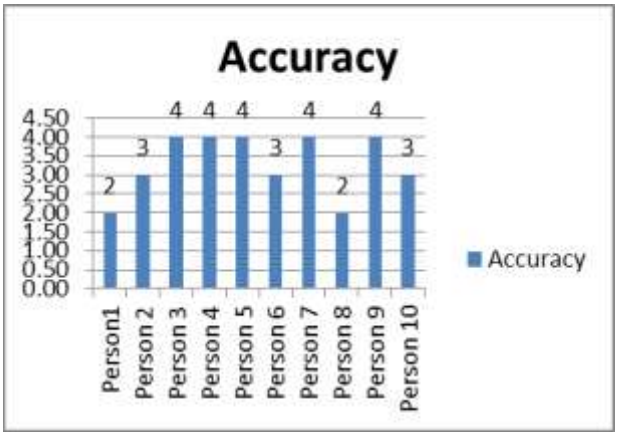

Fig. 1.7 Accuracy graph

So to find the accuracy of system used a formula asexpressed in equation 4.3.

$$\text{Accuracy}=\frac{\text{sum of ranking rate of no. of persons}}{\text{total number of ranking factor}} \quad ...(1.2)$$

The accuracy is defined as the ratio between sum of ranking rate of number of person to total number of rating factor.

7. Conclusion

In this thesis work, discussed the different techniques used in speech synthesis system. Some were basic techniques like concatenating speech synthesis; articulatory speech synthesis etc. and other were Statistical Parametric speech synthesis (SPSS) like HMM and DNN. Also, mentioned various steps that used to generate TTS system. Two types of formant based synthesis discussed, cascade base synthesis and parallel based formant synthesis. Concatenative speech sounds were traditional approach which produces sounds after concatenating the utterances. Robots are best fit for Articulatory Speech Synthesis. Various types of limitation in these speech synthesis techniques as needs a huge amount of database to build such system need to have complete knowledge about articulatory system of human being. Naturalness and intelligibility was also limited. Because of these limitations the system was less effective to overcome these problems Statistical Techniques are also popular from few decades. SPSS contains contain HMM, DNN, LRM, GMM etc. The main merit of this system is that it contains fewer amounts of data as compare to traditional techniques. Naturalness and intelligibility better as compare to previous synthesis techniques. In this research we prepared our own database. We have 722 phonemes sheet. We have audio file with wave format. For 722 phonemes we extracted wave file for each phonemes, collected 200 wave files and each their features and stored in database. Database prepared from wave file and 28 feature coefficients were extracted of each phoneme. After extracting MFCC, various steps applied to amplifying the signals. The generated system also contain some limitations, the sounds that generated is not natural as much as spoken by humans. There are hybrids techniques on which there is least work in research area and will be need of higher consideration in future.

REFERENCES

[1] D. Jurafsky and J. H. Martin, "Speech and Language Processing 18 BT - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," in An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2009.
[2] A. Balyan, S. S. Agrawal, and A. Dev, "Speech Synthesis: A Review."
[3] G. Kaur, "Formant Text To Speech Synthesis Using Artificial Neural Networks," 2019.
[4] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2007.
[5] D. Qinsheng, Z. Jian, W. Lirong, and S. Lijuan, "Articulatory speech synthesis: A survey," in Proc. - 14th IEEE Int. Conf. on Computational Science and Engineering, CSE 2011 and 11th Int. Symp. on Pervasive Systems, Algorithms, and Networks, I-SPA 2011 and 10th IEEE Int. Conf. on IUCC 2011, 2011.
[6] D. Siddhi, J. M., and D. Bhavik, "Survey on Various Methods of Text to Speech Synthesis," Int. J. Comput. Appl., vol. 165, no. 6, pp. 26–30, 2019.
[7] J. Parker, Y. Stylianou and R. Corolla "DNN-BASED SPEAKER-ADAPTIVE POSTFILTERING WITH LIMITED ADAPTATION DATA FOR STATISTICAL SPEECH SYNTHESIS SYSTEMS Mirac ¸ G ¨ oksu Ozt ¨ Bo ˘ gin University Computer Engineering , Electrical & Electronics Engineering," pp. 7030–7034, 2019.
[8] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 26, no. 8, pp. 1406–1419, 2018.
[9] K. M. Khalil and C. Adnan, "Arabic speech synthesis based on HMM," 2018 15th Int. Multi-Conference Syst. Signals Devices, SSD 2018, pp. 1091–1095, 2018.

BIOGRAPHIES

1. Harsimarjeet Kaur, Assistant Professor, GGI, Khanna (Punjab), India.

2. Dr. Parminder Kaur, Professor and Head of Department in CSE, GNDEC, Ludhiana (Punjab)/India