Text to Speech Synthesis System for Punjabi language using Statistical Parametric Speech Synthesis Technique

Harsimarjeet Kaur, Parminder Singh

***Abstract*: Statistical Parametric Speech Synthesis has been most growing technique rather than the traditional approaches that we are used to synthesizing the speech. The shortcoming of traditional approaches will be overcome with latest statistical techniques. 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 ` and robustness. It generates high quality of speech from small training database. Deep Neural network and Hidden Morkov model are basic statistical parametric speech synthesis techniques. Gaussian mixture model, sinusoidal model are also under this categories. Features were extracted in two type spectral features like spectral bandwidth, spectral centroid etc. and excitation features like F0 frequencies etc. We are using 722 Punjabi phonemes. Using sound forge software we extracted the 200 wave file from 1 hour pre-recording wave file related to those phonemes. Each and every phonemes feature was extracted and saved in database. We were extracting 28 features of each phoneme. TTS text-to-speech system generates sounds or speech as a output when provided the text of Punjabi language. There were already many TTS are developed on different Indian languages. The system that we are trying to build is based only on Punjabi language.**

Index Terms*: SPSS, *TTS, Phonemes, HMM.

I. INTRODUCTION

Artificial creation of speech is called Speech Synthesis. Speech synthesizer or Speech generated Computer systems are used for synthesizing the speech from text. A text to speech arrangement is also identified as TTS System that converted the normal text which is written in any language into speech. There are also other systems that only represent the symbolic linguistic description like phonetic transcription into speech. Numerous processing stages in text to speech system are used. Analyses and normalizes the received text, generates proper pronunciations of each word in context and also produces prosody of the sentence to be vocalized. All this work is done in front end. The term prosody means rhythm, emotions, melody and intonation. The TTS arrangement reliant on three parameters: accuracy of generating sound, intelligibility and naturalness in speech that will be generating.

Revised Manuscript Received on July 08, 2019
Harsimarjeet Kaur, Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India.
Parminder Singh, Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India.

General steps of a synthesizer:

There are three main parts of processing text into speech.

- **Text analysis**: Identify basic utterance and word from raw data.
- **Linguistic analysis**: Proper pronunciation of particular word, allocating prosodic structure to them, parsing, intonation and duration.
- **Waveform generation**: Generate wave form

II. RELATED WORK

A. F. Jalin and J. Jayakumari (2017) TTS (Text-To-Speech) systems generate speech for the text given as input. Though generation of speech is with moderate complexity the aspect of introducing the naturalist with the expression of the speaker is a big challenge faced in TTS. The intelligibility and the natural sounding speech have arrived at the suitable level. MFCC is calculated. This work will be continued by modifying using festival framework [1].

N. Adiga et al (2017) described the different features have been used for voicing decision. They can be outline into time domain and frequency-domain features. Ordinarily, time area features measure the acoustic nature of voiced sound for example energy, periodicity, zero crossing rate, and short-term relationship. In addition, several improvements have been presented for voicing result from break down speech recorded in realistic situations. Few of the current algorithms like YIN, Get_F0, SWIPE etc. provided very good voicing estimate. Voicing decisions are taken by some threshold which was chosen empirically. The implementation of these procedures rests critically on the threshold. To expand the precision and to evade threshold, statistical means such as HMM, Gaussian mixture model (GMM-HMM), neural network model, and deep neural network also used for voicing decision [2].

E. Gerbier et al (2017) defined that in a communal DNN based SPSS framework, text features at the input are made out of categorical, numerical features and duration features. The categorical features are ordinarily 1-hot-k encoded depictions. Learning inserting for the phones/ words/ utterances can also be well-thought-out as one alternative tactic. Usually a post-filtering method is adopted to enhance formants so that the sounds are sharper, and the over smoothing effect is compensated [3].

R. Kaur et al (2016) converted text into speech for language like Punjabi via eSpeak. The work for Formant Synthesis

in Punjabi Language can be used into many applications like screen readers for people with visual impairment, entertainment productions etc. increase naturalness and intelligibility [4].

F. Araújo et al (2016) described the evolution of an arrangement based on a genetic algorithm (GA) to automatically guess the input parameters of the Klatt and HLSyn formant synthesizers. GA algorithm for utterance duplicate procedure begins by dividing the input voice file into frames and configuration file [5].

D. Mahanta et al (2016) describes that how Grapheme are converted into phonemes that are used to build TTS system. Such type of conversion is difficult for non-native language like Indian English. Using North American English accent a CMU dictionary was prepared which contain 39 phonemes and 12500 words. Those are useful for synthesis of speech and recognition of speech. The main motive of this research had to improve intelligibility and naturalness to pronunciation of Indian English [6].

III. SPEECH SYNTHESIS TECHNIQUE

A. Concatenate speech synthesis

It is technique that uses to synthesizing the sounds by concatenation short sample of recorded sounds. These recorded sounds are called units. It is tough to find the duration of particular unit and it might vary according to the implementation, roughly its range is in between 10millisecond to 1second. Concatenate Text to speech synthesis system is extremely prohibitive because of huge data requirements and development time. So a more statistical method was developed instead of using tradition methods. The speech is built by combining and processing the parameters like fundamental frequency and magnitude spectrum. A good quality of sounds is generated by combining them and processing them [7].

B. Articulation speech synthesis

It is way for generating or synthesis the speech which is reliant on model of human vocal. To control the shape of vocal tract there are numerous ways which generally comprises by changing the location of articulator such as tongue, lips, and jaw etc. speech is formed digitally simulating when the air is flowing over the depiction of the local tract. It helps to model the human speech making system like vocal tract system and articulatory procedures straightly. If someone having deficiency of knowledge of the complex human articulation organs for them it is also very difficult to implement such method [8].

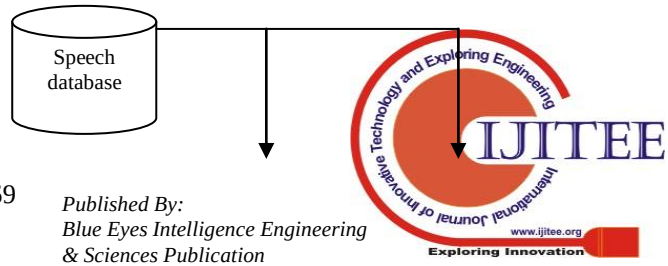
C. Formant speech synthesis

It uses the model of speech which is known as source filtering model, where speech is modeled by constraints of the filter model. Formant synthesis that is based on rule can yield that quality of speech which sounds unnatural, since it is very tough can yield quality of speech which sounds unnatural, to approximate the vocal tract and source parameters. The demerits of these approaches are that large data requirements and development time is needed. Cascade level structure means resonators are connected with each other as a series and parallel formant system structure means separately find formant frequencies and then concatenating all outputs to generate sounds or speech [9].

IV. SYSTEM ARCHITECTURE

Statistical Parametric Model for Speech Synthesis, when we talk about a model-based method to speech synthesis, mostly when we wish to learn this model from data, we generally mean a statistical parametric model. The model is parametric as it defines the speech using parameters, rather than stored exemplars. It is statistical for the reason that it defines those parameters using statistics like means and variances of probability density functions which capture the distribution of parameter values found in the training data. HMM and Deep Neural Network is two statistical parametric speech synthesis techniques.

HMM-dependent synthesis is a statistical parametric speech synthesis process reliant on HMM. In this scheme, the frequency spectrum means vocal tract, vital frequency means vocal source, and duration means prosody of speech are demonstrated at the same time by HMMs. Speech waveforms are delivered from HMMs themselves dependent on the most extreme probability. It comprises of two phases, the training phase and the synthesis phase. The spectrum and excitation parameters are take out from speech database and modeled by circumstance dependent HMMs in the training phase. A model usually consists of three states that represent the beginning, the middle and the end of the phone. The synthesis phase deals with generation of speech signals by concatenating the situation dependent HMM according to the text to be synthesized. The training part is alike to those used in speech recognition systems. The main difference is that both spectrum and excitation (e.g., log F0 and its dynamic features) parameters are taken out from a speech database and modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are occupied into account). As a result, the system models spectrum, an arbitrarily given text conforming to an utterance to be synthesized is changed to a context dependent label arrangement and lately the utterance HMM is built by concatenating the context dependent HMMs conferring to the label sequence. Secondly, state intermissions of the HMM are decided based on the state interval possibility density functions. Thirdly, the speech parameter generation algorithm creates the sequence of mel-cepstral coefficients and log F0 values that maximize their output possibilities. Lastly, a speech waveform is synthesized straight from the created mel-cepstral coefficients and F0 values using the MLSA filter with binary pulse or noise excitation.



Retrieval Number: I10420789S19/19©BEIESP
DOI: 10.35940/ijitee.I1042.0789S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

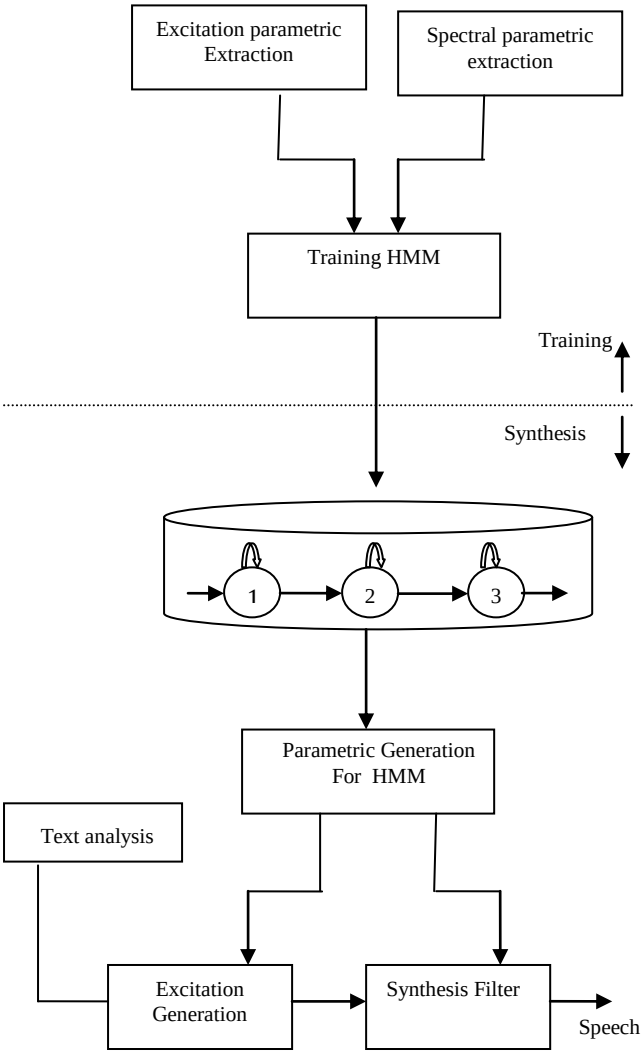


Fig. 1: Statistical parametric speech synthesis system

V. METHODOLOGY

a. Generating features from Phonemes

1. **Speech input**: We extracted the phonemes from wave sound file using sound forge software. Wave file is detached into minor frames.

2. **Feature extraction:** The extraction of best parametric illustration of acoustic signal is a chief task to produce best recognition performance. The mfcc contain seven computation steps:
   - Pre-emphasis
   - Framing
   - Windowing
   - DFT
   - Mel filter bank
   - Discrete cosine transformation
   - Energy data and Spectrum.

These all are pre-processing steps of wave file. Features will extract from audio wave files. These features are mfcc, spectral centroid, spectral bandwidth, zero crossing rate etc .The results will stored into database as Excel sheet. Python is used to extract features.

3. **Train model:** Features that were extracted from every wave file, first stored in database. All features will be modeled using Deep Neural Network technique. All data should store in database future used in testing or synthesis part. We can also train model using HMM-GMM. Train process based EM algorithm that done following:
   - Assign each frame to HMM state.
   - Re-estimate model parameter for GMMs (means and variance) and HMMs (Transition probability).
   - Repeat the convergence.

To train HMM model four parameters are used. X, Y, a, b, Where X is numbers of states, Y is the possible observation, 'a' is state transition probabilities and 'b' is the output probabilities. Forward and backward algorithm is used to model Hmm. Some of the cases Viterbi Algorithm are also used. All the implementation is done by python program.

b. Text to speech

1. **Text as input:** Input written in text box is as a Punjabi language. Another way is to browser a file in which Punjabi content was written and the text inside the file will show in text box.
2. **Tokenization** it is also called pre-processing step. In this step, the unwanted text written in text box will be removed. Punctuation, special characters etc. will remove from text and every sentence is converted into the form of tokens. Tokenization is of two type sentence tokenization and word tokenization. Text is input and stream of tokens are output. If sentence generates will act as token then is called sentence tokenization otherwise if in a particular sentence each word is act as a single token then under the word tokenization.
3. **Part of speech:** Input is as a word or sentence and result will be that every work has to assign a pos. then, result will store in database.
4. **Phonetic synthesizer:** text input first converted into token and token will be further converted into phonemes. The phonemes features will extract from database and wave sounds will generate. And then save the sounds.
5. **Concatenation:** Text will convert into phonemes. Concatenate process is used to join each phoneme to make a word or sentence.
6. **Speech synthesis:** using filter generate the speech sounds or waveform.

VI. RESULT

By experimenting towards the purpose of trying to generate mfcc coefficient and text to speech sounds. During training phase, features are extracted from sound files that are kept in database. The phonemes that were extracted will show in table below:

Table I: List of phonemes

Phoneme	Spec_cen	Zcr	S_band	Mfcc1
ਬਾ	1118.92	0.03772	1466.65	134.795
ਬ	807.378	0.02605	1538.95	127.505
ਬੀ	1596.301	0.02456	2173.80	69.4022
ਬੇ	3222.111	0.02605	1538.95	101.223
ਚਾ	1796.201	0.02805	433.332	123.957
ਦਾ	234.2232	0.02325	1538.95	956.871
ਗਾ	871.7834	0.02235	912.333	293.654
ਦੂ	1956.300	0.021005	213.119	171.505
ਤੇ	807.378	0.02555	412.334	657.541
ਚਾ	1446.104	0.02705	1132.90	127.547

In table 1 list of phonemes with their features are mentioned. Each phoneme contains 28 parameters i.e. chroma_stft, rmse, spectral_centroid, spectral_bandwidth, Zero-crossing rate, mfcc etc., but we mentioned 4 inside the paper. All these features were extracted using Librosa python library and results of every phoneme stored in database manually. Punjabi consonants are handled with the help of sound forge software. Model will trained using these features.After extraction of mfcc, it contains steps like pre-emphasis, framing and windowing etc.In fig. 2 described the waveform representation of given text input.

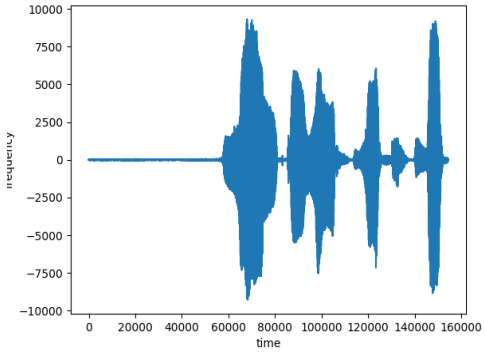


Fig. 2: Plot the wave representation of sound file

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, we assume the frequencies over time very carefully or securely. After pre-emphasis the shape of waveform is strictly be little bit compared as to actual representation of waveform which shown in fig. 3.

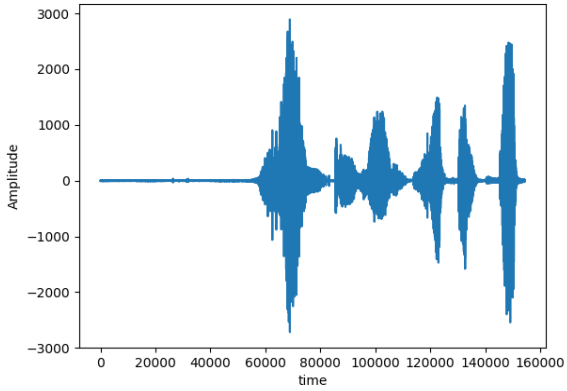


Fig. 3: Plot the pre-emphasis wave representation of sound

Signals are concatenating into accent frames and decent approximation of frequencies is generated. This process is called framing. In fig 4 show that speech signals are divided into frames .

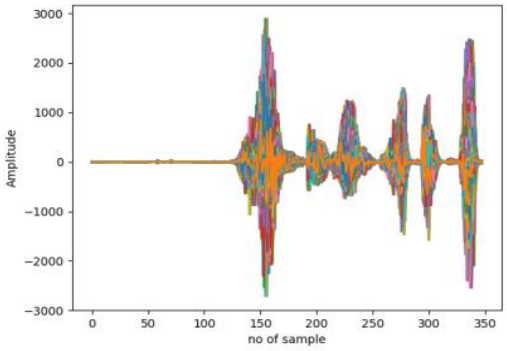


Fig. 4: Plot the frame representation of phoneme 'ਦਾ'

In fig. 2, 3 and 4 mentioned the pre-emphasis and framing the waveform of given text.

In table 2 described the word tokenization and sentence tokenization. In word tokenization, every word act as single token while in sentence tokenization the whole sentence consider as single token. Tokenization is pre-processing steps. Using python program builds a TTS system which text is input and audio generated as output. Input as Punjabi text, speech will be generated into wave format. In text box we can write Punjabi text or another way, browse a text file and open it. The content inside file will show in text box

Table II: Tokenization of sentence.

Sentence tokenization	Word Tokenization

Retrieval Number: I10420789S19/19©BEIESP
DOI: 10.35940/ijitee.I1042.0789S19
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

['ਕੋਰੀਆ ਵੀਅਤਨਾਮ ਕੰਬੋਡੀਆ ਅੰਗੋਲਾ ਬੁਸਨੀਆਂ ਕੋਸੁਵੋ ਰਵਾਂਡਾਂ ਅਲਜੀਰੀਆ ਸੂਡਾਨ ਤਿੱਬਤ ਅਫਗਾਨਿਸਤਾਨ ਸ਼ਿਰੀ ਲੰਕਾਂ ਫਲਸਤੀਨ ਖਾਲਿਸਤਾਨ ਕਸ਼ਮੀਰ ਨਾਗਾਲੈਡ ਅਸਾਮ ਨਾਰਦਰਨ ਆਇਰਲੈਂਡ ਐਲ ਸੈਲਵਾਡੋਰ ਨਿਕਾਰਾਗੂਆ ਚੈਚਨੀਆ ਆਦਿ ਖਿੱਤਿਆਂ ਵਿੱਚ ਮਨੁੱਖੀ ਲਹੂ ਦੇ ਦਰਿਆ ਵਗੇ ਹਨ ਜਾਂ ਵਗ ਰਹੇ ਹਨ ।']	['ਕੋਰੀਆ', 'ਵੀਅਤਨਾਮ', 'ਕੰਬੋਡੀਆ', 'ਅੰਗੋਲਾ', 'ਬੁਸਨੀਆਂ', 'ਕੋਸੁਵੋ', 'ਰਵਾਂਡਾਂ', 'ਅਲਜੀਰੀਆ', 'ਸੂਡਾਨ', 'ਤਿੱਬਤ', 'ਅਫਗਾਨਿਸਤਾਨ', 'ਸ਼ਿਰੀ', 'ਲੰਕਾਂ', 'ਫਲਸਤੀਨ', 'ਖਾਲਿਸਤਾਨ', 'ਕਸ਼ਮੀਰ', 'ਨਾਗਾਲੈਡ', 'ਅਸਾਮ', 'ਨਾਰਦਰਨ', 'ਆਇਰਲੈਂਡ', 'ਐਲ', 'ਸੈਲਵਾਡੋਰ', 'ਨਿਕਾਰਾਗੂਆ', 'ਚੈਚਨੀਆ', 'ਆਦਿ', 'ਖਿੱਤਿਆਂ', 'ਵਿੱਚ', 'ਮਨੁੱਖੀ', 'ਲਹੂ', 'ਦੇ', 'ਦਰਿਆ', 'ਵਗੇ', 'ਹਨ', 'ਜਾਂ', 'ਵਗ', 'ਰਹੇ', 'ਹਨ', '।']

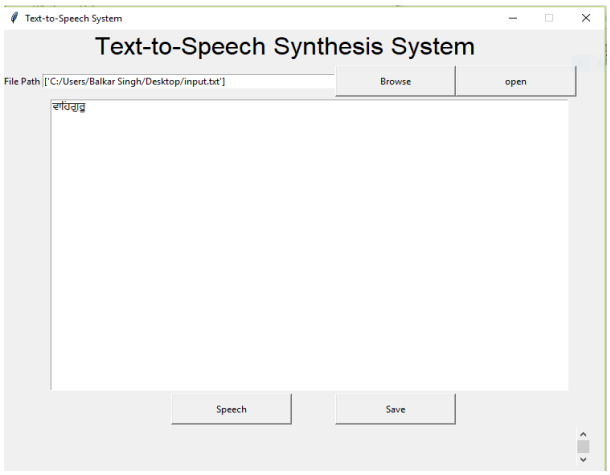


Fig. 5: TTS system

VII. CONCLUSION

In this paper we 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 like hmm and Deep Neural Network. Also, mentioned various steps that used to generate TTS system. Database prepared from wave file and 28 features were extracted of each phoneme. After extracting mfcc, various steps applied to amplifying the signals. The numerical value in table 1 are features of wave file that are extracted manually using python Librosa feature extraction. There are also some limitations of this TTS system as the quality of output depends on length of string in input text, shorter the string better will be the result. Naturalness of artificial speech is still not as much as similar to natural speech. To overcome these problems there are hybrids techniques on which there is least work in research area and will be need of higher consideration in future.

REFERENCES

1. A. F. Jalin and J. Jayakumari, "Text to speech synthesis system for Tamil using HMM," in *IEEE International Conference on Circuits and Systems, ICCS 2017*.
2. N. Adiga, B. K. Khonglah, and S. R. Mahadeva Prasanna, "Improved voicing decision using glottal activity features for statistical parametric speech synthesis," *Digit. Signal Process. A Rev. J.*, 2017.
3. E. Gerbier *et al.*, "Deep Elman recurrent neural networks for statistical parametric speech synthesis," *Speech Commun.*, 2017.
4. R. Kaur, R. K. Sharma, and P. Kumar, "Building a Text-to-Speech System For Punjabi Language," *IT-CSCP*, 2016.
5. F. Araújo, J. Filho, and A. Klautau, "Genetic algorithm to estimate the parameters of Klatt and HLSyn formant-based speech synthesizers," *BioSystems*, 2016.
6. D. Mahanta, B. Sharma, P. Sarmah, S R Mahadeva Prasanna, "Text to Speech Synthesis System in Indian English" in *IEEE Region 10 Conference (TENCON) — Proceedings of the International Conference*, 2016.
7. 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 Introdction to Natural Language Processing, Computational Lingustics, and Speech Recognition*, 2009.
8. S. Lukose and S. S. Upadhya, "Text to speech synthesizer-formant synthesis," in *IEEE International Conference on Nascent Technologies in Engineering, ICNTE 2017 - Proceedings*, 2017.
9. G. Kaur and P. Singh "Formant Text To Speech Synthesis using Artificical Neural Network," in *IEEE Second internal conference on advanced computational and communication paradigms,ICACCP2019*.

AUTHORS PROFILE

Harsimarjeet Kaur, Mtech. Student in Guru Nanak Dev Engineering College, Ludhiana (India)., her area of interest is Speech Processing. For any query or help contact on email harsumar1994@gmail.com.

Dr. Parminder Singh, Head of Department and professor in CSE in Guru Nanak Dev Engineering College, Ludhiana (India). He has published various research papers in the field of speech processing. Email parminder2u@gmail.com