ISSN:0975-9646

Audio Visual Speech Synthesis and Speech Recognition for Hindi Language

Kaveri Kamble, Ramesh Kagalkar

Department of Computer Engineering
Dr. D. Y. Patil School of Engineering & technology
Pune, India.

***Abstract*— Every person in the world want to share their information, thoughts from one person to another. So communication plays very important role into that. Speech is the primary means of communication. Hindi is very popular and well known language of India. Everybody understands and speak and write easily. Our System developed for Hindi Text to Speech and Speech to Text Conversion mainly into the Hindi Language. Expressions are the spirit of communication. Hindi TTS Engine plays very important role in TTS as well as STT conversion for Hindi language. TTS synthesizer will helpful for Text Processing and Speech Generation. This system is very helpful for the people who are having hearing impairment and blind people.**

***Keywords*— Text to Speech(TTS), Speech to Text(STT),Expressive Speech Synthesis, Boosting Gaussian Mixture Model (GMM), Mel Frequency Cepstral Coefficient(MFCC), Emotion, Speech Synthesis, Speech Generation, Signal Processing.**

I. INTRODUCTION

Language is a very fast and effective way of communicating. To use language means to express an unlimited amount of ideas, thoughts and practical information by combining a limited amount of words with the help of a known language like Marathi, Hindi etc.[5][19][12].The result of language production processes are series of words and structure. If people who are severely impaired in their hearing abilities want to take part in oral communication, they need a way to compensate their physical impairment. A text to speech (TTS) synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an Optical character recognition (OCR) engine. Speech is often based on concatenation of natural speech i.e units, that are taken from natural speech put together to form a word or sentence[20]. Naturalness expresses how intimately the output sounds like human speech, whereas intelligibility is the easiness with which the output is understood. This TTS system is able to read any written text for Hindi language.. The TTS can be a voice for those people who cannot speak. The TTS system can be useful for disabled person to make effective communications.[7][8]. The final stage of a TTS system is waveform generation which involves the production of an acoustic digital signal. Natural Language Processing (NLP) component, while the waveform generation stage is known as the Digital Signal Processing (DSP) component of a TTS System. This research has resulted in important advances with many systems being able to generate a close to a real natural speech for Hindi language. When we talk about speech-based interfaces for computer system, we refer to two basic technologies: Speech Recognition & Speech Synthesis. Speech Synthesis, *i.e.*, Text-to-Speech system and Speech Recognition, *i.e.*, Speech-to- Text system, together form a speech interface[8][9][12]. We are making an Application which will develop the text document from the simple voice for Hindi.. Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker. For Speech to Text conversion Feature Extraction refers to the process of conversion of sound signal to a form suitable for the following stages to use. Feature extraction may include extracting parameters such as amplitude of the signal, energy of frequencies, etc. Work on expressive speech synthesis has long focused on the expression of basic emotions.[6][7] Most research on expressive speech synthesis has been aimed at the prosodic expression of "basic" emotions such as sadness, fear, happiness, and anger. For emotional speech synthesis, many acoustic features, such as pitch variables, intensity, and speech rate have already been analysed[9]. One possibility for changing the sound of a unit selection voice is to apply voice transformation technology to the generated synthetic speech. Some prosody features, such as pitch variables (F0 level, range, contour, and jitter), and and speaking rate have already been analyzed. Then GMM and the Boosting-GMM are tried to predict the emotional speech acoustic features[6][7][8][9].

II. TEXT TO SPEECH SYSTEM

A text to speech (TTS) synthesizer is a computer based system that converts the given text into speech waveform[5]. Existing many TTS are developed for different languages like English, Telegu, and Punjabi etc. But now here we are developing for Hindi Language[18]. Text to speech should be made audibly communicate information to the user, when digital audio recordings are inadequate, for developing a user friendly speech synthesizer. Thus this system widely helps in developing a Computer-Human interaction like- voice annotations to files , Speech enabled applications, talking computer systems (GPS, Phone-based) etc. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it[17]. We have to create our own database to store Hindi words. We have to create our own dictionary which having all the Words, Text or Sentence. Hindi TTS Engine play very important role into the TTS system.

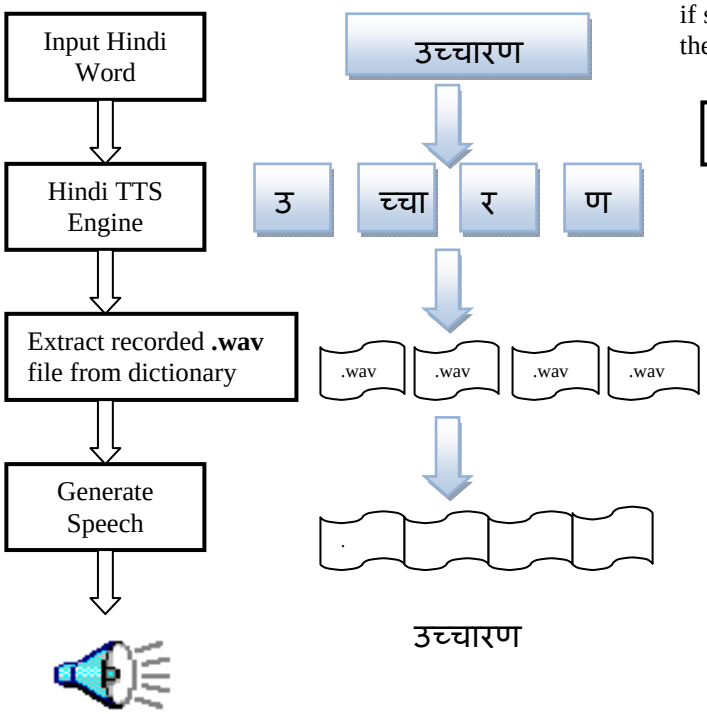


Fig.1.Flow chart of TTS system overview with Example.

III. Speech To Text System

Speech is the primary means of communication. Spoken words always play very important role in communication[7]. In this STT system we have to Speech as input to the system and output get in the form of Text[1][2]. Hindi TTS Engine play very important role for converting Speech to Text also[5]. Speech-to-text-translation (audio-visual translation) of spoken language into written text is an upcoming field since movies on DVDs are usually sold with subtitles in various languages like Hindi, English. In Speech to Text conversion process input that form into the form of .wav file[4]. Speech takes as a input to the Hindi TTS engine. Then play that .wav file extract features into that,. Match that feature to the training data set. When the training and the testing data set output should match then the Text has to be extracted for Hindi Language. Here we are developing STT system for Hindi Language. The real time application STT system is very useful[10]. The system acquires speech at run time through a microphone and processes the sampled speech to recognize the uttered Text. We used the hidden Markov model for speech recognition, which converts the speech to text[19]. Our speech-to-text system directly acquires and converts speech to text. Real-time speech-to-text-conversion aims at transferring spoken language into written text (almost) simultaneously.[14] This gives people with a hearing impairment, those who know the Hindi language has to access to the contents of spoken language in a way that they e.g. become able to take part in a conversation within the normal time frame of conversational turn taking . The main aim of speech-to-text transfer is to give people access to spoken words and auditory events almost simultaneously with the realization of the original sound event. If Hindi words are unknown or if sentences are too complex, the written form does not help their understanding .[13]

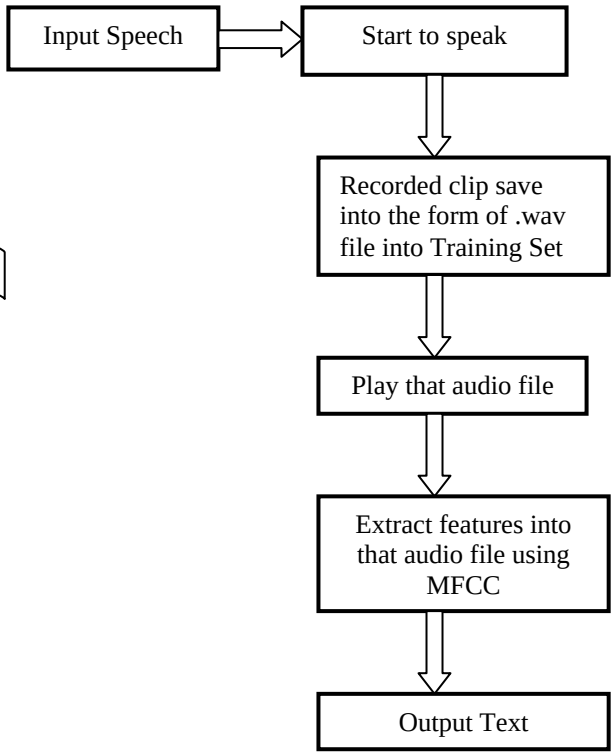


Fig.2. System overview of STT conversion

IV. Expressive Speech Synthesis

The synthesis of expressive speech is a target application with potential relevance in several areas, including the dynamic generation of multimodal media content and naturalistic human–machine interaction[8]. The expression is defined as the indicator of various emotional states that reflect in the speech waveform[15]. Expressive speech synthesis deals with synthesizing speech and adding various expressions related to different emotions and speaking styles to the synthesized speech.[9] The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. Emotion which is a very important element in expressive speech synthesis. Expressions that are Happiness, Sadness, Fear, Surprise etc. have to be very important in storytelling application. Expressions play very important role into the Communication. In our system we have to detect Emotions according to the expressions. Our approach primarily consists of three steps: first we take the text and the target PAD values as input, and employ text-to-speech (TTS) engine to generate neutral speeches. Boosting-GMM is used to convert natural speech to emotional speech. The GMM method is more suitable for a small training set[6]. TD-PSOLA algorithm is useful for the Pitch duration. acoustic features of emotional speech has calculated by using MFCC.

V. System Architecture

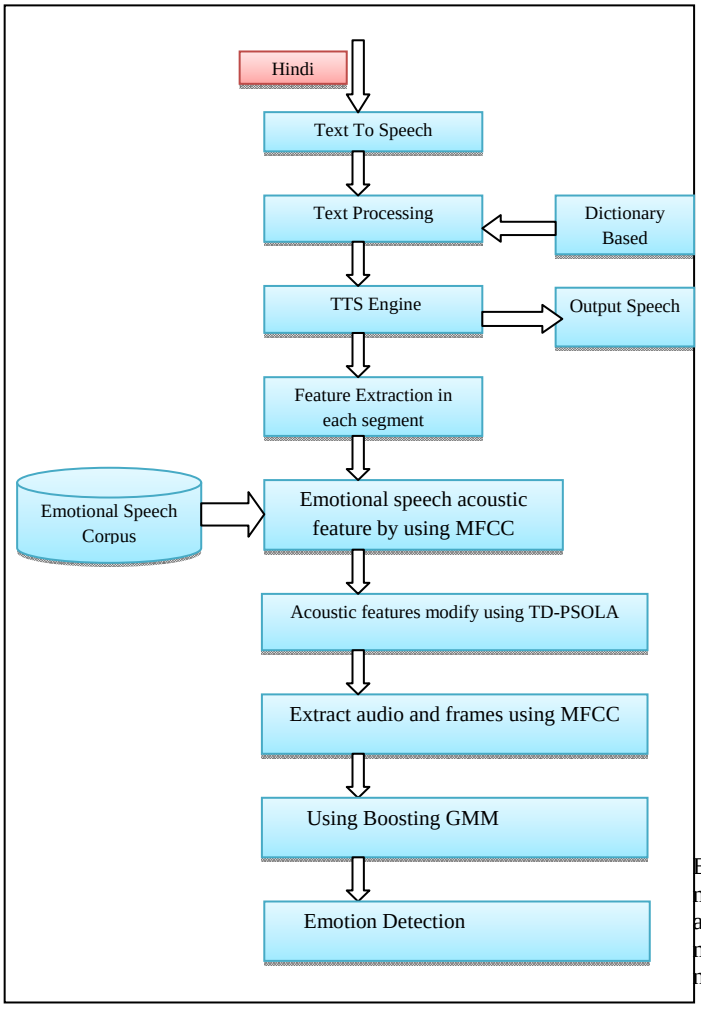


Fig.3. System Architecture of overall system

VI. Proposed System

Many Text to Speech systems are already developed for Different languages existing but they are not so user friendly. Hence, to full fill this requirement we develop our system for Automatic Translation of Text to Speech and Vice Versa for Hindi Language. The aim of this system is to develop a desktop application that can combine between three functions, converting Hindi Text to Speech, translating the Hindi Speech to Text and Expressive Audio visual speech synthesis based on Emotions. The System is very user friendly so user can use this system at any time and any place. The user also can use each function separately in order to provide more benefits and efficiency.

VII. Process Execution

1.Text To Speech Conversion:-

i. Get Text from user.

ii. Get English similar words related to text entered.

English Word	Related Hindi Word
aap	आप
arthapurna	अर्थपूर्ण

iii. Apply Hindi text to TTS engine.

iv. Extract audio from database.
For natural speech extract audio related to words given at input

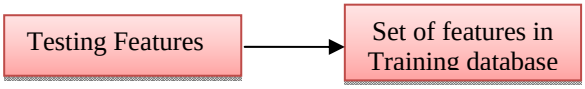
v. Recognize audio.

2. Speech To Text Conversion:-

i. Get Audio from user.
Input = .wav file (speech audio file)

ii. Extract Features of audio signals using MFCC.

iii: Match Features from training data (stored into database).



iv Extract text from database according to features.

3. Emotion Detection from Video:-

i: Get Audio from user.

ii: Extract features of Audio and Video.

iii: Apply Boosting GMM.

Boosting-GMM algorithm contains several weak prediction models. One of them is the basic prediction model and the others are assistant prediction models. The basic model is the regression model for predicting feature differences between emotional and neutral speeches using GMM.

iv: Search from database matches with expressions of testing clip.

vi. Result type of expressions in clip.

VIII. Screen Shots of TTS and STT System:

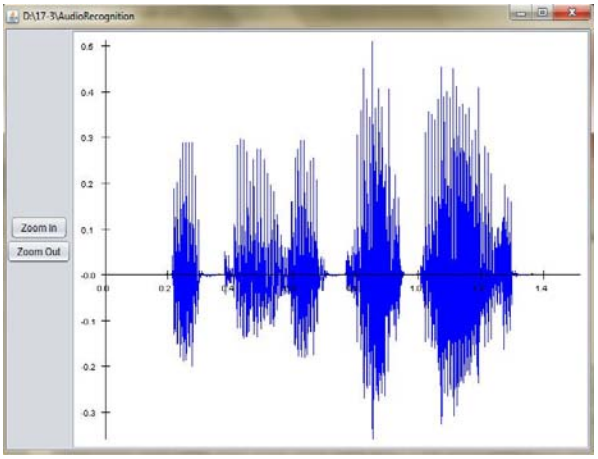
A] Process execution overview of TTS system:

1. Enter any word as input to the system. Suppose we have to enter उच्चारण as a word input.

2.

3.We have to hear the उच्चारण word clearly.

4.



5.then display the signal for the word उच्चारण.

B] Process execution overview of STT system:

1.In STT system firstly give input as a .wav file which has to be stored into the our own dictionary.

2.

3.then select the file like I have to give Abhipray.wav file as input.

4.

5.Then play that audio file.

6.By using MFCC we have to extract feature from audio file.

7.Extract the Text.

IX. Conclusion

The text to speech conversion may seem effective and efficient to its users if it produces Natural speech and by making several modifications to it. This system is useful for deaf and dumb people to Interact with the other peoples from society. The system which is a desktop application that works primarily as a converter between text and speech, for the Hindi language in both directions. Text to speech synthesis is a critical research and application area in the field of multimedia inter-faces. The proposed work presents an algorithm for converting text to speech by using natural language processing for Hindi language. There are many text to speech systems (TTS) available in the market and also much improvisation is going on in the research area to make the speech more effective, natural with stress and emotions. At present, with inadequate prosodic models in place, the quality of synthetic speech generated by the synthesizers is poor. So efforts can be done for the development of prosodic models. The further work can be done to improve the natural-ness and intelligibility of TTS. A Web based application can also be designed which can convert text in any Indian languages into speech.

Acknowledgment

The authors would like to thank Chairman Groups and Management and the Director/Principal Dr. Uttam Kalwane, Colleague of the Department of Computer Engineering and Colleagues of the Department the D. Y. Patil School of Engineering and Technology, Pune Dist. Pune Maharashtra, India, for their support, suggestions and encouragement.

REFERENCES

[1] J. Tao ,Y. Kang,and A. Li ,Prosody Conversion From Neutral Speech to Emotional Speech, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 14, No. 4, July 2006.

[2] M. Theune, K. Meijs, D. Heylen, and R. Ordelman ,Generating Expressive Speech for storytelling Applications' , IEEE Transaction on Audio, Speech and Language Processing,Vol.14, No.4, July 2006.

[3] D. Govind , S. Mahadeva Prasanna , Expressive speech synthesis: a review , Springer Science Business Media New York 2012.

[4]B. Yegnarayana and K. Sri Rama Murty , Event-Based Instantaneous Fundamental Frequency Estimation From Speech signals, IEEE Transactions on Audio ,Speech. And Language Processing, Vol.17. No. 4, May 2009.

[5] O. Turk and M. Schroder , Evaluation of Expressive Speech Synthesis with Voice Conversion and Copy Resynthesis Techniques, IEEE Transactions on Audio, Speech and Language Processing, Vol.18, No.5 ,July 2010.

[6] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, Member, IEEE, Emotional Audio-Visual Speech Synthesis Based on PAD,IEEE Transactions on Audio, Speech and on Audio, Speech and Language Processing, Vol.19, No.3 , march 2011.

[7] J.Sangeetha,S.Jothilakshmi , S. Sindhuja , V. Ramalingam ,Text to Speech synthesis system for Tamil, International Conferenceon Information Systems and Computing (ICISC-2013),India.

[8] K. Kamble and R. Kagalkar,A Review:Translation of Text toSpeech Conversion for Hindi language, International Journal of Science and Research (IJSR) Volume 3 Issue 11, November,2014.

[9] M. Singh, K. Verma , Text to Speech Synthesis for numerals into Punjabi language, International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 7 July 2013 ISSN 2279 0756.

[10] N. Swetha, K. Anuradha ,Text-to-speech conversion, International Journal of Advanced Trends in Computer Scienc and Engineering ,Vol .2,No.6, Pages (2013).

[11] S. Ahlawat, R. Dahiya , A Novel Approach of Text to Speech Conversion Under Android Environment, (IJCSMS) International Journal of Computer Science Management Studies, Vol. 13, Issue 05, July 2013.

[12] P. Shetake, A. Patil, P. Jadhav , Review Of Text To Speech Conversion MethodS, International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-2, Issue-8, Aug.-2014.

[13] S. Suryawanshi, R. Itkarkar, D. Mane , High Quality Text to Speech Synthesizer using Phonetic Integration, International Journal , Advanced Research in Electronics and Communication Engineering (IJARECE) Volume3, Issue 2, February 2014.

[14] D. Sasirekha, E. Chandra ,Text to Speech: A Simple Tutorial,International Journal of Soft Computing and Engineering (IJCSE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[15] S. Hertz, J. Kadin, And K. Karplus, Member, IEEE,The Delta Rule Development System for Speech Synthesis from Text,Proceedings of the IEEE ,Vol.73, No.11, November 1985.

[16] R. San-Segundo, J. Montero, R. Barra-Chicote, J. Lorenzo, Architecture for Text Normalization using Statistical Machine translation techniques, Springer-verlag Berlin Heidelberg 2011.

[17] A. Chauhan, V. Chauhan, S. Singh, A. Tomar, and H. Chauhan, A Text to Speech System for Hindi using English Language,IJCST Vol 2.Issue 3,September 2011.

[18] S. Padmavathi, K. Reddy , Conversion Of Braille To Text in English,Hindi and Tamil Languages International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.3, No.3, June 2013.

[19] S. Suryawanshi, R. Itkarkar, D. Mane , High Quality Text to Speech Synthesizer using phonetic Integration, International Journal of Advenced Research in Electronics and Communication Engineering (IJARECE) Volume 3, Issue 2,February 2014.

[20] O. Trk, O. Byk, A. Haznedaroglu, and L. Arslan ,Application of conversion for cross language rap singing transformation in proc.IEEE ICASSP, Taipei, Taiwan, April 2009.

[21] Z. Zeng, P. Maja, G. Roisman, and S. Thomas ,A survey of affect recognition methods: Audio, visual and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell., vol. 31 , no. 1, pp. 3958, Jan.2009.