IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 01-04
www.iosrjournals.org

Emotion classification and frequency domain parameters of speech signal for the expression of prosody in synthetic speech

Miss Ashwini S Shinde[1] ,Mr Sachin S Patil[2]
[1](Electronics and Tele commutation, ADCET,Ashta / Shivaji University Kolhapur, India)
[2](Electronics and Tele commutation, ADCET,Ashta / Shivaji University Kolhapur, India)

ABSTRACT : *In the past 10 years many speech synthesis systems has shown remarkable improvements in speech quality .Instead of monotonous and mechanical sounding utterances .For the better sounding pattern and more appealing sound synthesis few elements need to be taken care. Improvements in quality of synthetic speech and improvement in naturalness of the signal plays major contribution in the advancements in the synthetic speech technology. Improvement in signal quality can be technologically attended by concatenative synthesis. Naturalness is improved with the help of prosodic modelling. The proposed work mainly aims towards developing classifier for the input text with the help of linguistic model. Input to the system is pure text and which contains little or no explicit information about meaning, even though it may have punctuation marks in it. Its realization as a prosody is a major challenge and that has been undertaken by this project.*
Keywords *- Concatenative Speech Synthesis, prosodic modelling, sound synthesis, training set, vocal emotion.*

I INTRODUCTION

Speech is the principal mode of communication between humans, both for transfer of information and for social interaction. Consequently, learning the mechanisms of speech has been of interest to scientific research, leading to a wealth of knowledge about the production of human speech, and hence to technological system to simulate and to recognize speech electronically [1].Speech synthesis is the name given to techniques used to create artificial human speech. Earlier mechanical device called a Vocoder could be played like a keyboard instrument to produce synthetic speech .Today, we are able to use computers to produce synthetic speech from entered text, and keyboard virtuosity. There have been a number of computer-based approaches for producing synthetic speech, including articulatory approaches, formant-based approaches, and concatenative approaches.

1.1 Articulatory Speech Synthesis

Articulatory approaches have traditionally been based on vocal tract models that treat it as the concatenation of tubes of differing diameters. A notation consisting of electrical circuit components is often used to represent such models due to the similarity of the underlying mathematics and the ability to build the corresponding electrical circuits .Eventually, computers could be used to simulate these circuits to perform speech synthesis .

1.2 Formant-Based Speech Synthesis

Formant-based approaches focus on the resonant frequencies in the sound wave instead of the underlying physical activity that produces the sound. These approaches produce intelligible, though quite unnatural sounding, speech. Due to their greater tractability, these approaches have been used to create practical applications of speech synthesis. One system that exemplifies this approach is MITalk.

1.3 Concatenative Speech Synthesis

Concatenative speech synthesis techniques are currently very popular. They involve taking recordings of speakers, analyzing the recordings into units that are separated, stored, and concatenated (i.e. joined sequentially) to create new utterances. These techniques have the advantage of capturing some of the Character of the original speaker's style, and thus tend to sound more natural than articulator or formant-based approaches.

II. EMOTIONAL TEXT TO SPEECH

The goal of next generation speech synthesizers is to express the variability typical to human speech in a natural way or, in other words, to reproduce different speaking styles and particularly the emotional ones in a reliable way [4]. The quality of synthetic speech has been greatly improved by the continuous research of the speech scientists. Nevertheless, most of these improvements were aimed at simulating natural speech as that uttered by a professional announcer reading a neutral text in a neutral speaking style. Because of mimicking this style, the synthetic voice results to be rather monotonous, suitable for some man-machine applications, but not for a vocal prosthesis device such as the communicators used by disabled people [4]. In the last years, progress in speech synthesis has largely overcome the milestone of intelligibility, driving the research efforts to the area

of naturalness and fluency. These features become more and more necessary as the synthesis tasks get larger and more complex: natural sound and good fluency and intonation are mandatory for understanding a long synthesized text [5]. A vital part of speech technology application in modern voice application platforms is a text-to-speech engine. Text-to-speech synthesis (TTS) enables automatic converts any available textual information into spoken form. With the evolution of small portable devices has made possible the porting of high quality text-to-speech engines to embedded platforms [2] [3]. It is well known that speech contains acoustic features that vary with the speaker's emotional state. The effects of emotion in speech tend to alter pitch, timing, voice quality and articulation of the speech signal [6] [7]. Expressive speech synthesis from tagged text requires the automatic generation of prosodic parameters related to the emotion/style and a synthesis module able to generate high quality speech with the appropriate prosody and the voice quality [8]. Furthermore, adding vocal emotions to synthetic speech improves its naturalness and acceptability, and makes it more 'human'. We provide the user with the ability to generate and author vocal emotions in synthetic speech, using a limited number of prosodic parameters with the concatenative speech synthesizer [9]. The voice plays an important role for conveying emotions. For example, rhythm and intonation of the voice seem to be important features for the expression of emotions [10] [11]. Adding emotions to a synthesized speech means that the latter can verbalize language with the kind of emotion appropriate for a particular occasion (e.g. announcing bad news in a sad voice). Speech articulated with the appropriate prosodic cues can sound more convincing and may catch the listener's attention, and in extreme cases, it can even avoid tragedies . An improved synthesized speech can also benefit from other speech-based human-machine interaction systems that perform specific tasks like read-aloud texts (especially materials from the newspaper) for the blind, weather information over the telephone, auditory presentation of instructions for complex hand free tasks

2.1 Prosodic synthesis

In order to evaluate the results of various prosodic TTS techniques, it is necessary to ask what constitutes good prosody in general. it is useful to first ask what constitutes good synthetic speech. The goal of synthetic speech is to provide sounds that are easily understandable by a human listener. The speech should be intelligible in the sense that a person can recognize the exact words that are being synthesized. Furthermore, the speech should sound as human as possible. It is also necessary to ask what additional considerations should be added for synthetic speech. Because prosodic synthetic speech has the additional goal of creating synthetic speech that sounds like it was produced with a particular emotion, prosodic identity is an important component of prosodic TTS quality. It should also be noted that prosody is an area that needs to be investigated more broadly with respect to synthetic speech generation. In some applications, it may be important for transformed speech to be identifiable as coming from a specific emotional response. for example when a person reads a children's story it may be sufficient to have the different voices sound like they come from the same speaker, yet appear to represent different emotions, and thus be distinct.

III. RECENT RELATED RESEARCHES: A REVIEW

Mumtaz Begum *et al*. have presented the findings of their research which aims to develop an emotions filter that can be added to an existing Malay Text-to-Speech system to produce an output expressing happiness, anger, sadness and fear. The emotions filter has been developed by manipulating pitch and duration of the output using a rule-based approach. The emotional speech output has undergone several acceptance tests. The results have shown that the emotions filter developed has been compatible with FASIH and other TTS systems using the rule-based approach of prosodic manipulation. However, further work needs to be done to enhance the naturalness of the output.

Zeynep Inanoglu *et al*. have described the system that combines independent transformation techniques to provide a neutral utterance with some required target emotion. The system consists of three modules that are each trained on a limited amount of speech data and act on differing temporal layers. F0 contours have been modeled and generated using context-sensitive syllable HMMs, while durations are transformed using phone-based relative decision trees. For spectral conversion which is applied at the segmental level, two methods have been investigated: a GMM-based voice conversion approach and a codebook selection approach. Converted test data have been evaluated for three emotions using an independent emotion classifier as well as perceptual listening tests. The listening test results have shown that perception of sadness output by their system has been comparable with the perception of human sad speech while the perception of surprise and anger has been around 5% worse than that of a human speaker. Now concatenative speech synthesis techniques allow the creation of voices that have more distinct identities, there is a desire to create many different voices.

IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 01-04
www.iosrjournals.org

IV MOTIVATION OF THE RESEARCH

The recent research works have been briefly reviewed in the previous section. From the review, it can be seen that the previous research works have explained adding emotions in text to speech synthesizing systems. The existing research works have achieved the emotion adding process by extracting speech features like speed, energy or modeling the duration of the speech sounds and some works perform speech synthesizing for obtaining the emotions. The existing works elucidate the emotion adding process by using any one of the speech features and not by using a combination of all the features. Some drawbacks are present here, i.e. using any one of the speech feature cannot always exactly and correctly identify user's emotions. For example the works based on speed/energy feature are mainly focused on the specified feature and does not consider the speech synthesizing and duration modeling. Above all, most of the previous works are done in the English language and very little works are presented in the Hindi language. A dedicated speech synthesizing technique is needed for Hindi as Hindi linguistically varies substantially from English or other foreign languages. All the aforesaid issues and the necessities have motivated me to do the research work in this topic.

V PROPOSED METHODOLOGY

The analysis of existing research works asserts that lack of dealing and drawbacks are present in the existing text to speech synthesizing systems. In this work, I intend to develop a text to speech synthesizing technique to synthesize Hindi speech with emotions. To accomplish the above task, I have categorized the entire research into three modules; they are i) Feature Extraction ii) Modeling iii) Synthesizing.

The first research module developed will be an emotion-based speech classification technique. For this, different speech signals with different emotions will be collected. Though there are numerous types we will consider only significant emotions like sadness, anger as well as normal speech in our technique. As it will be a supervised technique, training dataset will be generated and features will be extracted from it and finally it will be used for training. The features will be selected according to the emotions and emotional speeches will be classified based on the features. Most probably, the extracted features will be strength, speed, time instance and high, low pitches of the speech.

In the second research module, the classified speeches will be subjected to analysis so as to identify the properties in every emotion. Using the nature of emotion obtained from the features, two analytical models will be developed for representing the emotions. The first analytical model will be for representing sadness in speech whereas the second model will be for representing angriness in speech both with reference to the normal speech

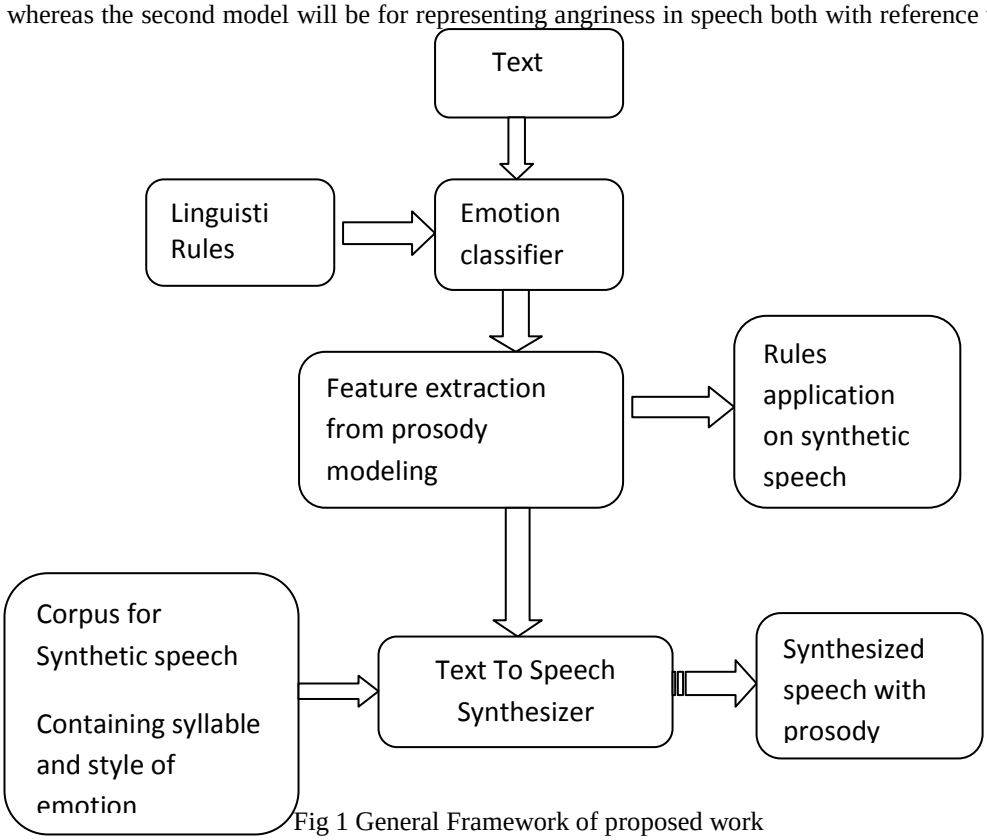

Fig 1 General Framework of proposed work

In the final research module, the major work of speech synthesizing will be performed by developing a Text to Speech Synthesizer for Hindi. Here, the analytical model that will be developed in the previous section will play a vital role. In the synthesis process, when a Hindi text is given to the synthesizer along with the required type of emotion, the synthesizer will synthesize Hindi speech for the specified type of emotion. In synthesizing such speech, the analytical models that are developed for the concerned emotion will be utilized.

5.1 Methodology

The main objective of the research can be achieved by following methodology

5.1 Design of complete framework for speech synthesis by concatenation.
5.2 Feature extraction from natural speech for prosidical modeling manually.
5.3 Formulization of typical values as a set of rule.
5.4 Transplantation of extracted feature values on synthetic speech.
5.5 Iterative testing and acceptability of individual variants.
5.6 Follow-up correction according to test results.
5.7 Selection and transplantation of modeled prosodic pattern for given sentence type.

The proposed research modules will be implemented in MATLAB and its performance will be evaluated.

VI. CONCLUSION

Linguistic classification and prosody modeling combine enforcing the prosody in neutral synthetic speech. Around the world it is recently viewed as best practices to attempt application of linguistic theory for prosody modeling. An iterative testing of acceptability of individual variants is be done by subjective tests. This is followed by follow-up correction according to the test results The expected contributions of this work are: A demonstration that using suggested methodologies aids prosodic synthesis. which will be a better quality speech synthesis.

REFERENCES

PAPERS

[1] Iain R. Murray and John L. Arnott, " Synthesizing Emotions In Speech: Is It Time To Get Excited?", In Proceedings of the Fourth International Conference on ICSLP, Philadelphia, PA , USA, Vol. 3, pp. 1816-1819, 1996

[2] Jerneja Zganec Gros, Ales Mihelic, Nikola Pavesic, Mario Zganec and Stanislav Gruden, "Slovenian Text-to-Speech Synthesis for Speech User Interfaces", World Academy of Science, Engineering and Technology, Vol.11, No.1, pp.1-5, 2005

[3] M.L. Tomokoyo, W.A. Black and K.A. Lenzo, "Arabic in my hand: small footprint synthesis of Egyptian Arabic," In Proceedings of the Eurospeech'03, Geneva, Switzerland, pp. 2049-2052, 2003

[4] Montero, Gutierrez-Arriola, Palazuelos, Enriquez, Aguilera and Pardo, "Emotional Speech Synthesis: From Speech Database to TTS", In Proceedings of the 5th international conference on spoken language processing, Sidney, pp. 923-926, 1998

[5] Ibon Saratxaga, Eva Navas, Inmaculada Hernaez and Iker Luengo, "Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque", In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC), pp. 2126–2129, 2006

[6] Selma Yilmazyildiz, Wesley Mattheyses, Yorgos Patsis and Werner Verhelst, "Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication", In Proceedings of 7th Pacific Rim Conference on Multimedia, Springer Lecture Notes in Computer Science, Hangzhou, China, Vol. 4261, pp. 1-8, 2006

[7] Cynthia Breazeal and Lijin Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech", Journal Autonomous Robots, Vol. 12, No. 1, pp. 83-104, January 2002

[8] Ignasi Iriondo, Joan Claudi Socoro and Francesc Alias, "Prosody Modelling of Spanish for Expressive Speech Synthesis", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, pp. 821-824, 2007

[9] Caroline Henton and Peter Litwinowicz, "Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech", In Proceedings of the 2nd ESCA/IEEE workshop on Speech Synthesis, pp. 73-76, 1994

[10] Enrico Zovato and Jan Romportl, "Speech synthesis and emotions: a compromise between flexibility and believability", In Proceedings of Fourth International Workshop on Human-Computer Conversation, Bellagio, Italy, 2008

[11] Klaus R. Scherer, "Vocal communication of emotion: a review of research paradigms", Journal Speech Communication - Special issue on speech and emotion, Vol. 40, No. 1-2, pp. 227–256, April 2003

Books/websites

[1] E.Keller,G.Belly Improvements in speech synthesis cost 258

[2] Dutoit, T. (1977) An Introduction to Text-to-Speech Synthesis. Dordrecht: Kluwer Academic

[3] Ben Gold ,Nelson Morgan speech and audio signal processing wiley India edition

[4]Lawrence rabiner,being hwang juang Fundamentals of speech recognition pearson publication

[5]http://hts.ics.nitech.ac.jp/