IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Emotional Text to Speech Synthesis: A Review

Jayanta Kalita[1], Nabamita Deb[2]

Department of IT, Gauhati University[1,2]

Abstract: Several attempts have been done to add emotional effects to synthesized speech and several prototypes and fully operational systems have been built based on different synthesis techniques. Butfor Indian languages, there is still a lack of fully operational text to speech synthesis system with emotional effects. This paper aims to give an overview of what has been done in this field for some of the Indian Languages and highlights different issues faced during the development.

Keywords: Indian Language Scripts, IPA, MaryTTS, Emotional Speech Database

I. INTRODUCTION

Emotions are expressed in speech, face, gait and other body languages explicitly by human beings along with internal physiological signals such as muscle voltage, blood volume pressure, skin conductivity and respiration. The vocal expressions are harder to regulate than other explicit emotion signals. So, it is possible to know the actual affective state of the speaker from her/his voice without any physical contact. But exact identification of emotion from voice is very difficult due to several factors. The speech consists broadly of two components coded simultaneously: (i) "What is said" and (ii) "How it is said". The first component consists of the linguistic information pronounced as per the sounds of the language. The second component consists of non-linguistic or paralinguistic or supra-segmental component which includes the prosody of the language i.e. pitch, intensity and speaking-rate rules to give lexical and grammatical emphasis for the spoken messages; and the prosody of emotion to express the affective state of the speaker. The voice also contains information about the speaker's identity, age, gender, and body size. Thus, isolation of the affective information i.e. the emotion, from voice is not easy. [5].

J.M. Montero and et al. [1] found "Cold Anger" to be the most difficult emotion to implement. Felix Burkhardt and Nick Campbell [2] stated that the main challenge for emotional speech synthesis results from the discrepancy between natural but inflexible vs. artificial sounding but flexible synthesis approaches.
In this paper Section 2 describes the development of TTS system based on some Indian language scripts, Section 3 describes the issues faced in the development of MARY TTS system for German language and Section 4 finally describes different evaluation techniques of the synthetic voices.

II. REVIEWS ON INDIAN LANGUAGE SCRIPTS

The scripts in Indian languages have originated from the ancient Brahmi script [8]. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V.

HimangshuSarma and et al. [17] developed an Assamese speech corpus. They developed the speech corpus by recording seven hours of speech from twenty five different speakers from different regions of Assam. After identifying the phonemes used during transcription, they transcribed the words using International Phonetic Alphabet 2005 (IPA). During the transcription they found that from the IPA 2005 some of phonemes are absent in Assamese language. From the collected corpus they found that some of Assamese words are more frequently used than the others and sometimes the same word is pronounced differentially by speakers of different regions. They found that the pronunciation of many speakers of Assamese are influenced by other languages, such as English, Hindi and Bengali.

NavanathSaharia and et al. [19] created an Assamese tag-set that was not available before. They developed a well-defined tag-set of 172 tags in consultation with experts in linguistics. Using a manually tagged corpus of about 10000 words for training, they obtained a tagging accuracy of nearly 87% for test inputs.

Laba Kr. Thakuria and Prof. P.H. Talukdar [18] examined the syllabification rules for Assamese language and stated that Syllabification acts as a backbone for unit selection based text-to-speech system. Based on different structures of different languages syllabification rules are also varies. They presented an algorithm to break the Assamese words into

their component syllables based on Assamese grammar rules. To test the algorithm, they experimented it on 5000 distinct Assamese words chosen from Assamese Dictionary and had shown accuracy of 99% in their result giving about 20,665 syllables for the words.

S P Kishore and et al. [16] discussed the issues relevant to the development of voices for Indian languages. They observed that when the coverage of units is small, the synthesizer is likely to produce a low quality of speech, and there would be high variance among the scores given by different subjects. As the coverage of units increases, it increases the quality of synthesizer and there would be less variance in the scores given by different subjects.

SanghamitraMohanty [3] described the Text-To-Speech System for four of the Indian Languages namely, Hindi, Odiya, Bengali and Telegu. She observed that phone level concatenation is leading to complexity roughness of sound utterance while syllable level concatenation had smoothness in the utterance of the syllable.

While creating speech database for Hindi language, the author found that for the same part /ha/ in two different phones /bharat/ and /hai/, the wave forms are different and is completely depend on the prior and posterior phone or syllable. In a Bengali wave file the author noticed that for the text /bhAratAmarxdeshx/ the appearance of short vowel at the end of bhArat and Amarx is the place where syllable based concatenation is unavoidable to have smooth utterance as they appear very close. The onset point is confusing to identify for phone level concatenation.

III. THE MARY TEXT-TO-SPEECH SYSTEM

The text-to-speech system MARY (Modular Architecture for Research on Speech Synthesis) is a tool for research, development and teaching in the domain of text-to-speech synthesis.

Marc Schroder and JurgenTrouvain [4] presented the German TTS system using MARY. During tokenization the authors found some problems in German pronunciations of ordinal numbers due to inflections. The expansion of an ordinal number depends on its part-of-speech (adverb or adjective). On the other hand, for adjective ordinals, the inflection ending depends on gender, number and case of the noun phrase the ordinal belongs to.

IV. EVALUATION OF THE SYNTHETIC VOICE

MOS score was considered to evaluate intelligibility and emotion content of the sentences.
R.K. Bhakat, N.P.Narendra and K.S.Rao [11] carried out four phase's subjective evaluation on sad and neutral system for Hindi language. They observed that some neutral sentences had unwanted prosodic variations and there were some pitch and energy fluctuations that caused the sentences to sound unnatural. To get better MOS score they applied prosody modification algorithms on the unsatisfactory sentences.

Milos Cernak and Milan Rusko [20] present the experiments on the use of the perceptual objective measure – ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ) for the automatic evaluation of synthetic speech. They compare PESQ values with MOS values by calculating Pearson's correlation coefficients. They stated that PESQ measure can be used in the automatic evaluation of new versions of synthetic voices, without a need of subjective evaluation tests. They also found the evident that PESQ cannot be used for the evaluation of diaphone voice on small sample size.

V. CONCLUSIONS

The Indian languages are still lagging behind in the development of emotional text to speech synthesis system. This is may be due to insufficient resources or insufficient researches. We hope this paper will bring about some understanding and inspiration for further research in this area.

REFERENCES

[1] J.M. Montero, J. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, J.M. Pardo "Emotional Speech Synthesis: From Speech Database To TTS" Available at: https://www.researchgate.net/publication/221487874
[2] Felix Burkhardt and Nick Campbell "Emotional Speech Synthesis" 2009
[3] SanghamitraMohanty," Syllable Based Indian Language Text To Speech System" International Journal of Advances in Engineering & Technology, May 2011.©IJAET ISSN: 2231-1963
[4] Marc Schroder, JurgenTrouvain " The German Text-to-Speech Synthesis System MARY:A Tool for Research, Development and Teaching" Available at : http://mary.dfki.de
[5] Aditya Bihar Kandali,AurobindaRoutray ,Tapan Kumar Basu "Vocal emotion recognition in five native languages of Assam using new wavelet

features" Int J Speech Technol (2009) 12: 1–13 DOI 10.1007/s10772-009-9046-4
[6] "The Social and Emotional Voice". Retrieved 29 March 2012.
[7] Sauter, Disa A.; Eisner, Frank; Calder, Andrew J.; Scott, Sophie K. (1 November 2010). "Perceptual cues in nonverbal vocal expressions of emotion". The Quarterly Journal of Experimental Psychology. 63 (11): 2251–2272. doi:10.1080/17470211003721642
[8] AnandArokia Raj, TanujaSarkar, Satish Chandra Pammi,SanthoshYuvaraj, MohitBansal, Kishore Prahallad, Alan W Black "Text Processing for Text-to-Speech Systems in Indian Languages" Available at :http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.9263
[9] Bitopi Sharma, PurnenduBikash Acharjee2, Prof. P.H. Talukdar, "An Improved Grapheme to Phoneme rules for Assamese Language" International Journal of Advanced Research (2015), Volume 3, Issue 12, 1460 – 1464
[10] V. Tamilselvi, Dr.P.Visu "Development of Syllable Based Unit Selection Text-To-Speech Synthesis System for Tamil Using Three Level Fall Back Technique" 2015 IJEDR | NC3N 2015 | ISSN: 2321-9939
[11] R.K. Bhakat, N.P.Narendra and K.S.Rao "Corpus Based Emotional Speech Synthesis in Hindi" DOI- 10.1007/978-3-642-45062-4_53
[12] B.Ramani, S.L.Christina, G.A.Rachel, V.S.Solomi, M.K.Nandwana, A.Prakash, Aswin.S.S, R.Krishnan, S.P.Kishore ,K.Samudravijaya, P.Vijayalakshmi, T.Nagarajan, H.A.Murthy," A Common Attribute based Unified HTS framework for Speech Synthesis in Indian Languages" 8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain
[13] Van Santen, J. (April 1994). "Assignment of segmental duration in text-to-speech synthesis".Computer Speech & Language. 8 (2): 95–128. doi:10.1006/csla.1994.1005.
[14] AnindyaJyoti Roy, Arnab Bhattacharya," Emotional Text to Speech Synthesis in Indian Language"
[15] ArchanaBalyan, S.S. Agarwal and AmitaDev, "Speech Synthesis: A Review," in International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 2 Issue 6,June-2013
[16] S P Kishore, Alan W Black, Rohit Kumar, Rajeev Sangal "Experiments With Unit Selection Speech Databases for Indian languages" Available at https://www.cs.cmu.edu/~awb/papers/LTTIT-HCU-Exps.pdf
[17] HimangshuSarma, NavanathSaharia, Utpal Sharma, Smriti Kumar Sinha, Mancha JyotiMalakar, "Development and Transcription of Assamese Speech Corpus". National seminar cum Conference on Recent threads and Techniques in Computer Sciences.
[18] Laba Kr. Thakuria , Prof. P.H. Talukdar," Automatic Syllabification Rules for ASSAMESE Language" Laba Kr. Thakuria et al Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 2(Version 1), February 2014, pp.446-450
[19] NavanathSaharia, Dhrubajyoti Das, Utpal Sharma and JugalKalita "Part of Speech Tagger for Assamese Text" Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 33–36, Suntec, Singapore, 4 August 2009. c 2009 ACL and AFNLP
[20] MilosCernak, Milan Rusko "An Evaluation of Synthetic Speech Using the PESQ Measure" Forum Acusticum 2005 Budapest