The iSolar Blizzard Challenge 2013 Entry

Anh-Tuan Dinh[1], *Thanh-Son Phan*[2], *Dang-Hung Phan*[1], *Tung-Lam Phi*[1], *Tat-Thang Vu*[1], *Chi-Mai Luong*[1]

[1]Institute of Information Technology, Vietnam Academy of Science and Technology, Vietnam

[2]Faculty of Information Technology, Le Qui Don Technical University, Hanoi, Vietnam

{anhtuan, danghung, tunglam, vtthang, lcmai}@ioit.ac.vn, sonphan.hts@gmail.com

Abstract

The paper describes the iSolar TTS system architecture and how the Blizzard database was integrated into our system. The core of iSolar was based on HTS engine and Flite. An analysis of the mean opinion scores for test sentences shows where our system can be improved.

1. Introduction

The Blizzard Challenge is an evaluation of speech synthesis systems [1]. The year's evaluation provided a large speech database. Participants have to use the provided audio data to build up voices in 2 tasks: The main English audio book tasks and the pilot Indian task.

The iSolar TTS system is designed for research purpose. The system is designed to be transparent to non-expert user, without limiting its potential for experts to achieve optimal performance. Our system supports common lexica, Unicode and the International Phonetic Alphabet (IPA).

In recent TTS system, one problem is that the quality of synthetic voice is often proportional to the quantity of manual input required in training process. Much of the work in iSolar system is to determine the automatic alignment methods. HTK and Sphinx are supported in the alignment phase to achieve the best possible quality. Manual input is optionally supported after automatic phase with a convenient graphic user interface.

This is our first time to participate in the Blizzard Challenge. The participation is hoped to highlight the differences between the iSolar system and other research systems and assisted in identifying our future work. The iSolar entry for the Blizzard Challenge 2013 is a prototype based on HMM approach [2][5].

Our work focuses on EH2 task in English audio book challenge. We did not submit other tasks. The section 2 describes System architecture, Section 3 describes building blizzard voice process and Section 4 discusses the experimental results. Final section is conclusion and describes our future work.

2. System architecture

The iSolar is a flexible and modular TTS system. The modules consist of text preprocessor, sentence analyzer, word disambiguation and HMM based synthesizer. The iSolar system uses the object oriented programming paradigm throughout and was developed in C and Python. The current version can run on both UNIX style platforms such as Linux and Window systems.

The synthetic component used for the evaluation is a HMMM-based synthesizer, which creates speech by concatenating the HMM-model of each phoneme in input sentence. [4]

The modular structure allows the system to support new voice and new language by replacing the appropriate phoneme set, natural language processing modules, language model and acoustic model. The flexible modular structure is generally used in multi-language TTS systems.

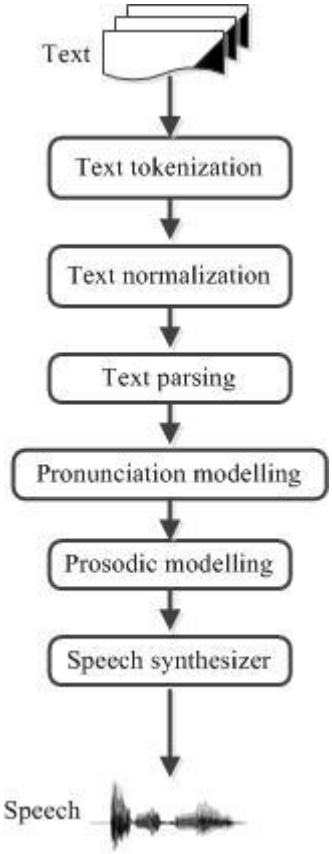

Figure 1: System architecture

3. Building Blizzard Voice

In the year's challenge, we submit only the EH2 task. In the task, a new voice is built using segmented audio data sampled at 16 kHz. The transcriptions provided with the Blizzard database did not have any guarantees of correctness. So we didn't use the provided transcription.

3.1.Phonetic transcription phase

The transcriptions were generated with the Flite. English text analyzer was based on the phoneme set of HTK. A manual revision needed to correct the errors due to the mismatches between the recorded voices and the orthographic text. Because the time for integrating Blizzard database in our system is limited so we only manually check the out of vocabulary words.

To perform the manual transcription, a python tool was developed. It displays the orthographic text of each sentence at the top and the phonetic transcription at the bottom. The out of vocabulary words are highlighted. A play button helps the user listen to the speech data of the sentence and do the transcription.

In the manual phase, not only out of vocabulary words are transcribed precisely but also the orthographic text can be corrected.

3.2.The phonetic Segmentation phase

The segmentation phase is fully HMM based. HTK was used to segment the 2000 sentences from Blizzard database. The segmentation for speech data took about 1 hour. No automated post-processing of the HMM labels or further manual segmentation corrections were undertaken.

After the voice building phase, model files are produced containing all data needed for synthesis using the voice. The synthesis phase is the concatenation of each appropriate HMM model of each phoneme in input sentence. 2000 files are used to build voice A and 500 files are used as test sentences.

3.3.Phonetic feature extraction

Beside the phonetic features such as the current phoneme and its context, the iSolar system uses a variety of prosodic features for HMM-based synthesis. Stanford POS tagger[3] was used for POS tagging. The high level prosodic such as phrase breaks were predicted based on a syntactic analysis of the input text.

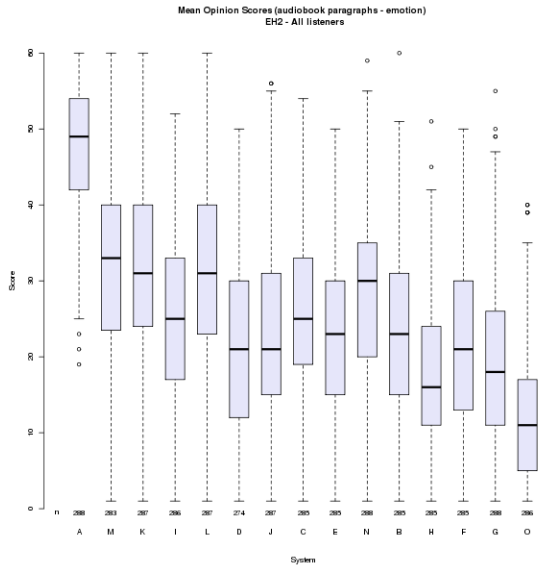

Figure 2: MOS results in emotion test

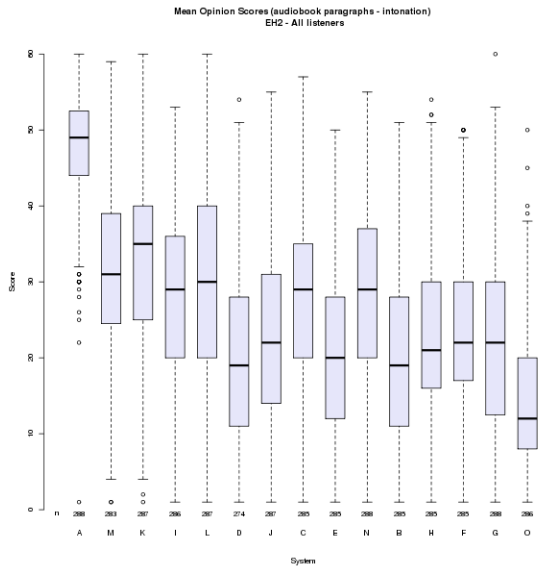

Figure 3: MOS results in emotion test

4. Experimental Results

The test result in EH2 consists of 9 sections. The listeners include of paid participants, volunteers and speech experts. Our system (O) applied simple knowledge about prosody in tagging phase. So the full context labels lack of sufficient prosodic data such as intonation. So the emotion and intonation couldn't be reconstructed well in synthetic voice. The results of emotion and intonation all test are

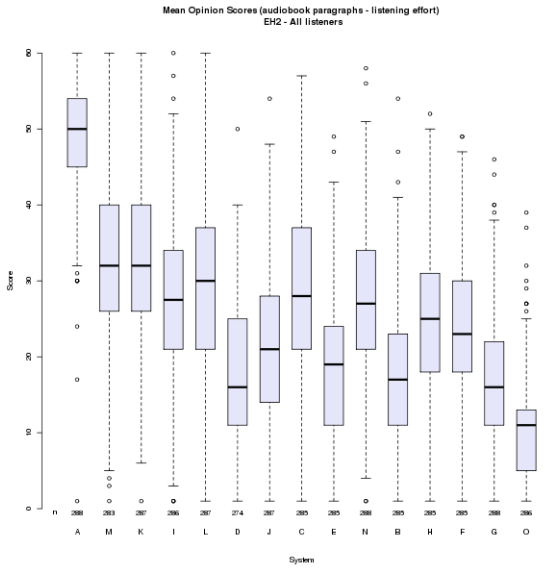

Figure 4: MOS results in listening effort test

shown in Figure2 and 3.

The use of Flite for generating full-context model and HTK in force alignment phase could affect negatively to the quality of training model because the phonetic system of HTK is different from those of Flite. The intelligibility can be improved with a more carefully generated full-context model.

5. Conclusion

The paper describes the iSolar entry in the Blizzard Challenge 2007. This is our first prototype to take part in the year's challenge. Speech was built from the audio data and orthographic transcription only. All other input data was generated automatically by the system.

The participation on Blizzard Challenge has helped highlight where we should focus on the future for example the improvements to the extracting of acoustic parameters.

6. Acknowledgments

This work was partially supported by ICT National Project KC.01.03/11-15 "Development of Vietnamese - English and

English - Vietnamese Speech Translation on specific domain". Authors would like to thank all staff members of Department of Pattern Recognition and Knowledge Engineering, Institute of Information Technology (IOIT) - Vietnam Academy of Science and Technology (VAST) for their support to complete this work.

7. Reference

[1] "The Blizzard Challenge 2012", Simon King and Vasilis Karaiskos (CSTR, University of Edinburgh, UK).

[2] "Speaker-Independent HMM-based Speech Synthesis System –HTS 2007 System for Blizzard Challenge 2007", Junichi Yamagishi, Heiga Zen, Tomoki Toda, Keichi Tokuda.

[3] "Part-of-Speech Tagging from 97% to 100%: Is it Time for some Linguistics". Christopher D. Manning

[4] "An Introduction to HMM-based Speech Synthesis". Junichi Yamagishi

[5] "A tutorial on Hidden Markov Model and Selected Applications in Speech Recognition". Lawrence R. Rabiner