

A RESPONSE GENERATION IN THE MONGOLIAN SPOKEN LANGUAGE SYSTEM FOR ACCESSING TO MULTIMEDIA KNOWLEDGE BASE

Munkhtuya Davaatsagaan and Kuldip K. Paliwal

Signal Processing Laboratory
Griffith University, Nathan, Brisbane, Australia

ABSTRACT

By using automatic speech recognition (ASR) and text to speech (TTS) systems, which have been available in Mongolian for last few years, this research set out to implement a new version of the Mongolian Virtual Education Environment (VEE) that has not included a speech interface. The spoken language system aims to provide a natural interface between trainees and the environment by using simple and natural dialogues to enable the user to access the multimedia knowledge base of the VEE. We have worked on the response generation part of the system. This paper describes a TTS system for the VEE for university courses held in Mongolian. A concatenative speech synthesizer for Mongolian is applied for the TTS in response generation. A Festvox framework for unit selection speech synthesis was used to build the Mongolian voice. We discuss aspects of the voice development process and the results of a perceptual test of the synthesized voice.

Index Terms— Response generation, Spoken language application

1. INTRODUCTION

A research team based in Mongolia built a Virtual Education Environment for university courses several years ago. The VEE is a computer-based interactive environment that supports the process of learning and teaching university courses. It is an application of information technology, telecommunications and multimedia technologies to education. Key factors in the success of this environment are network efficiency and good multimedia communications. Education material is online and accessed by a wide variety of students with different backgrounds and expectations. So the material proposed should be polymorphous and of the highest possible quality. The knowledge base is designed using object-oriented methodology to represent concepts taught by lecturers. Within this environment, the lecturers use a Virtual Teaching Environment (VTE) to organize and produce lessons that are delivered by multimedia material. The

students access the multimedia lessons through a Virtual Learning Environment (VLE), which can be personalized or made flexible to suit the students' preferences [1].

Due to the lack of research on ASR and TTS for Mongolian, it was not available to integrate speech into the environment at the time. Therefore the VEE could not provide a natural interface between users and the environment, and work effectively and efficiently. In addition, the VEE used audio files stored in the media server as audio media for concepts. Consequently, the system needed a large amount of disk space in the media server for use of audio media alone. The system performance and runtime were not efficient and effective as well. Currently, a new version of the VEE, which is a spoken language system for accessing the multimedia knowledge base, is being implemented. The system has an ASR subsystem that converts students' and lecturers' queries in speech to words, and a TTS subsystem that conveys text responses to spoken information. In the queries section, users say to select a subject name, a lecturer name, a lesson, and its contents, and the media in which they want to learn. The responses would be control and feedback speech information and audio media being converted from text media using the TTS when users choose audio media for a concept.

For several years, research on TTS has been carried out in Mongolia. During the last few years great progress has been made in development of a quality Mongolian text-to-speech synthesizer. A unit selection concatenative speech synthesis system for Mongolian is used to generate speech response in the spoken language system. The work is done within the FestVox voice building framework [2], which offers general tools for building unit selection synthesizers in new languages. The unit selection paradigm is a cluster based technique where units of the same type (phones, diphones, syllables or whatever) are clustered based on their acoustic differences [3]. The clusters are then indexed based on high level features such as phonetic and prosodic context. Voices generated by this system run in the Festival Speech Synthesis System [4].

2. THE MONGOLIAN LANGUAGE

Mongolian is the best-known member of the Mongolic language family, and the primary language of the residents in Mongolia. Majority of speakers in Mongolia speak the Halh dialect. The Altaic theory proposes that the Mongolic family is a member of the larger Altaic family. As an Altaic language, Mongolian shares several characteristics with other languages in this family. Some of these include agglutination, the use of post-positions instead of prepositions, vowel harmony, the placement of modifiers before that which they modify, the absence of a relative pronoun, the absence of the verb meaning “to have”, the absence of grammatical gender, and the absence of articles. The basic word order of Mongolian is subject-object-verb. Modifiers (adjective, adverb, determiners, etc.) generally precede the word that they modify and do not show agreement, e.g. no matter what case (nominative, genitive, dative, etc.) or number (singular, plural) a noun is, the adjective will stay the same. Mongolian differs from other Altaic languages in verb ending. The Cyrillic alphabet is used for writing Mongolian. With few exceptions it is phonemic in the sense that the spelling of a word and its phonemic form are equivalent, so that each one can be derived from the other.

3. THE SPOKEN LANGUAGE SYSTEM

The overview of the spoken language system for accessing the multimedia knowledge base is given in Figure 3.1 this system consists of a virtual teaching environment and virtual learning environment. In addition, the VEE was built by 3-layer architecture – Media base, Knowledge base, and Lecture base. The ASR and TTS subsystems provide a human-computer speech interface for the spoken language system.

The Media base is the lowest layer of the VEE that contains a collection of media as text, graphics, audio, animation and video. The media illustrate a specific concept form, and a specific point of view at a specific level of details. The basic scenario in media base of VEE consists of

- a. When requesting to view a concept, the media base retrieves the concept’s media files; and
- b. When creating a concept, the concept’s media files are uploaded into the media base.

The knowledge base is built by the Concept model based on object oriented approach. In the knowledge base, a concept is a representation of an idea that can be taught, explained or communicated by the lecturer to the students. The knowledge base is a combination of concepts and relationships among them. The concept in the knowledge base is designed as follows. The three parts of the model are related by two relationships that state a concept which contains characteristics and the characteristics can be viewed through media.

The third layer is the Lecture base which contains a collection of lectures prepared by lecturers. The Lecturers

organize the concepts for representation within the learning process according to the specific needs of students.

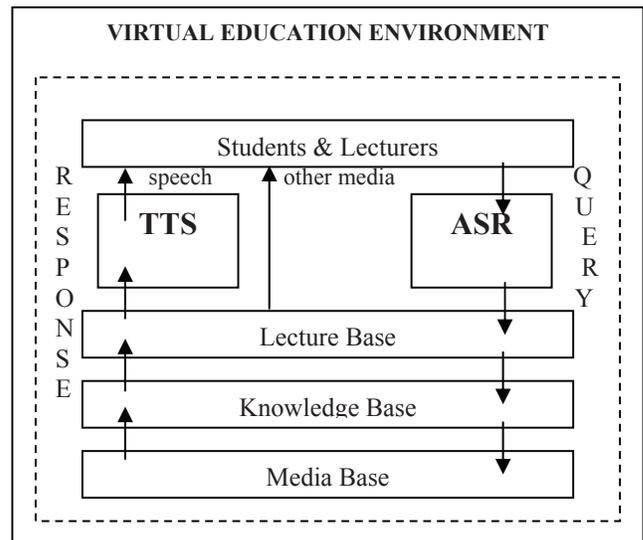


Figure 3.1: General architecture for VEE with ASR and TTS

Students and lecturers send their queries in speech using the ASR subsystem to the VEE and receive speech responses using the TTS from the VEE. This speech interface is a high degree of naturalness and intelligibility in the spoken presentation of relevant information.

4. RESPONSE GENERATION

The synthesis task of the system is performed here through the following three processes:

1. Creating speech database for unit selection: Storing most possible units with all possible prosodic variation for generating speech.
2. Analyzing text: Converting the input text into a phonemic internal representation with prosodic features.
3. Producing speech: Converting the internal representation into a waveform.

The architecture of the system shown in figure 4.1 has a layered structure and each layer consists of functional components. All required procedures and functions for the layers and their components will be defined in detail in the next sections.

4.1. Creating speech database

There are three issues concerning the creation of unit selection database.

- Developing text corpus
- Selecting optimal texts

- Recording the selected texts

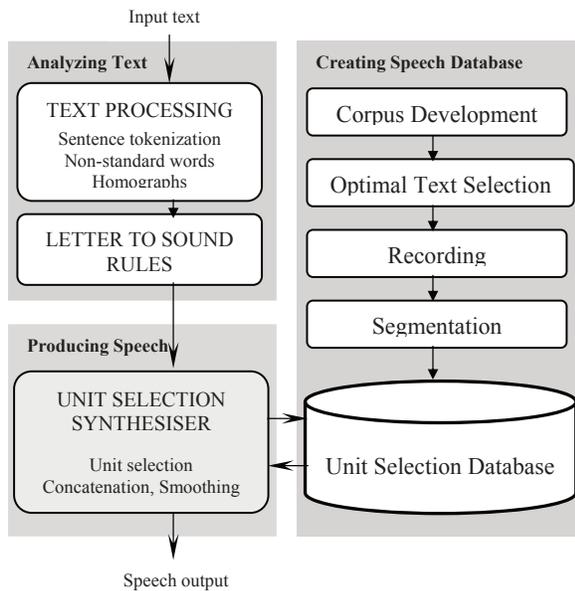


Figure 4.1: General system architecture for Mongolian TTS

4.1.1. Corpus development

In the first stage, we gathered a very large amount of text data of phonetically and prosodically varied speech from various sources comprising e-libraries, and news portals. The texts were all the same in UTF-8 encoding.

4.1.2. Optimal text selection and recording

Next, utterances which have the covered of most possible units with different prosody were automatically found from the corpus using optimal text selection algorithm. The greedy algorithm performs to select phonetically rich and balanced utterances from the text corpus.

The corpus consisted of a total of 731557 words and 1087 utterances were selected. There were 867 unique diphones and this covered 72.13% of the theoretically possible diphones in Mongolian. Furthermore, the corpus consisted of a total of 2465357 syllables and 11.35% of them were unique syllables. Among all the syllables instances: 2.73% of them were monosyllables, 53.88% of them were word-initial and word-final syllables, and rests of them were word-internal syllables. Most of the words consisted of approximately three syllables.

These utterances were recorded by native female speaker. The speaker had an experience of working in Mongolian National Radio for reading news and articles. The recording was done using noise cancellation microphone in a quite room environment.

4.1.3. Segmentation and clustering

We have been using the Festvox based phonetic recognizer for the task of automatic phonetic segmentation.

Given these segments, the unit selection algorithm in Festvox clustered the phones based on their acoustic differences. These clusters are then indexed based on higher level features such as phonetic and prosodic features. During synthesis, the appropriate clusters are sought using phonetic and prosodic features of the sentence. A search is then made to find a best path through the candidates of these clusters.

4.2. Analyzing text

4.2.1. Text processing

The text processing module performs sentence tokenization, handles the nonstandard words and carries out prosodic analysis.

In Mongolian, like English, whitespace (space, tab, newline, and carriage return) and punctuation can be separated from the tokens in the text. Each identified token is mapped to words, standard or non-standard.

Non-standard words are tokens like numbers or abbreviations, which need to be expanded into sequences of Mongolian words before they are pronounced. These non-standard words are often very ambiguous. Dealing with non-standard words requires three steps: tokenization to separate out and identify potential non-standard words, classification to label them with a type from a predefined table, and expansion to convert each type into a string of standard words [5].

An abstract representation of the prosodic prominence, structure and tune for the text is computed in this module.

4.2.2. Text to sound rules

There exists a well-defined mapping from the orthography to the pronunciation in Mongolian language. For well defined languages like Mongolian, writing rules by hand is simpler than training. We built hand-written rule sets for this system. Hand written letter to sound rules are context dependent re-write rules which are applied in sequence mapping strings of letters to strings of phones [5].

4.3. Producing speech

Text processing module is given a phone string together with features such as f0 information, stress value and so on. Then the synthesizer is to select from the database the best sequence of units that corresponds to the target representation. The best sequence would be one in which:

- Each unit we select exactly meets the specifications of the target unit (in terms of F0, stress level, phonetic neighbors, etc)

- Each unit concatenates smoothly with its neighboring units, with no perceptible break.

5. PERCEPTUAL EVALUATION

For evaluating the speech synthesis system, we synthesized texts from the Knowledge base. A set of 12 concepts selected from different lectures which made up 133 sentences were synthesized and 8 students who are native speakers of Mongolia were asked to evaluate the quality of the Mongolian synthesizer. Each listener was asked to evaluate 10-15 sentences chosen randomly from the set of sentences. In our first experiment, intelligibility of synthesised speech was evaluated on two levels: word level and sentence level. Subjects, participating in the test were asked to write down everything they heard. The percentage of correctly understood sentences is around 80%, and word intelligibility rate is close to 91%. In our second experiment, degree of acceptability of the synthesised speech was assessed by the following steps:

- The synthesized wave file was played to the listener once or twice only, without the sentence being displayed.
- The listener was asked to rate the naturalness of the synthesized speech waveform between 0-5 (0 for poor quality and 5 for excellent quality)
- The sentence was displayed and the listener was asked to select on the words which were Not Sounding Natural (NSN) to him/her. Here the listener was allowed to listen to the synthesized speech any number of times [6].

Table 5.1. *The evaluation results*

Listeners	Number of sentences	Mean score	Number of NSN words/ Total words	Errors in %
1	13	3,69	11/145	7,58
2	15	3,67	13/156	8,33
3	12	3,66	9/123	7,31
4	11	3,81	12/136	8,82
5	10	4,2	8/118	6,77
6	13	3,92	10/137	7,29
7	14	3,14	14/152	9,21
8	11	4,09	10/129	7,75
Overall		3,9		7,88

The mean of the scores given by each listener and the number of NSN words found by each listener are shown in Table 5.1. We can see that the mean score across all the listeners is 3,9 and the percentage of NSN words across all listeners is 7,88 %. The full list of NSN words was analyzed and the following observations were made.

- The text processing components were not handling all the combination of special characters (scientific symbols) with digits and letters.
- Around 17% of NSN words were new scientific loan words from English and other languages.

6. CONCLUSION

The work was the first attempt to use a unit selection voice for Mongolian for response generation of a spoken language system. We have discussed the issues to be considered in developing a unit selection speech synthesizer for Mongolian language. We conducted an evaluation on the Mongolian speech synthesis system. The subjects were asked to identify the words which are not sounding natural. The observation of the list of words indicated the effect of new scientific loan words, and special scientific characters should be considered while building the speech corpus and analyzing texts. Hence further work will mean improving the highlighted issues of the system.

7. ACKNOWLEDGEMENTS

The authors wish to thank the speaker for recording her voice to create the speech database and the students for volunteering for the evaluation process.

8. REFERENCES

- [1] Davaatsagaan, M., Yeoh, E.T. (2003) A Virtual Education Environment for Engineering Courses. *Cyberscape Journal*, Volume 1. ISSN 1675-9281
- [2] Black, A. W. and Kevin, A.L., "Building Synthetic voices", Language Technologies Institute, Carnegie Mellon University. Amer. Pittsburgh, PA. Online: <http://festvox.org/bsv/>, accessed on 12 Mar 2008.
- [3] Black, A. W. and Paul, A. T., "Automatically clustering similar units for unit selection in speech synthesis", *Eurospeech97 Proc.*, vol.2, 601-604, 1997.
- [4] Black, A. W. Paul, A.T. and Richard, C., "The Festival speech synthesis system", CSTR, University of Edinburgh. UK. Edinburgh. Online: <http://festvox.org/festival/>, accessed on 17 May 2007.
- [5] Munkhtuya, D., and Kuldip, K. P., "Diphone-based concatenative speech synthesis system for Mongolian", *ICEMG2008 Proc.*, vol.2, 317-320, 2008.
- [6] Kishore, S., Black, A., Kumar, R., and Sangal, R. "Experiments with Unit Selection Speech Databases for Indian Languages" Presented at National seminar on Language Technology Tools, India. Hyderabad, 2003.