

Speech Synthesizer Using Concatenative Synthesis Strategy for Marathi language (Spoken in Maharashtra, India)

Mr.S.D.Shirbahadurkar*. Dr.D.S.Bormane**

*Asst. Prof, E&TC dept JSPM's Rajarshi Shahu College of Engineering, Pune, India,

**Professor & Principal JSPM's Rajarshi Shahu College of Engineering, Pune, India

E-mail: s_shir00@yahoo.co.in, bdattatraya@yahoo.com, shirsd@gmail.com

ABSTRACT

In this paper, we present the concatenative text-to-speech system and discuss the issues relevant to the development of a Marathi speech synthesizer using different choice of units: words, phonemes as a database. Quality of the synthesizer with different unit size indicates that the word synthesizer performs better than the phoneme synthesizer. The most important qualities of a speech synthesis system are naturalness and intelligibility. We synthesize the Marathi text and perform the subjective evaluations of the synthesized speech. As a result, (1) 81% of speech synthesized by the proposed method was preferred to that by the conventional method, (2) The error rate of TTS synthesizer is around 8.22%, (3) Speech synthesis runtime was reduced for proposed method. The results show the effectiveness of the proposed method.

Keywords: Speech synthesis, concatenation, unit size, syllabification.

1. INTRODUCTION

Speech is one of the most vital forms of communication in everyday life. On the contrary the dependence of human computer interaction on written text and images makes the use of computers impossible for visually and physically impaired and illiterate masses. Automatic speech generation from natural language sentences can overcome these obstacles. In the present era of human computer interaction, the educationally under privileged and the rural communities of any country are being deprived of technologies that pervade the growing interconnected web of computers and communications.

II. SPEECH SYNTHESIS SYSTEM

A Text-to-Speech (TTS) Synthesizer is a computer based system that should be able to read any text aloud, whether it was introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. The objective of a text to speech system is to convert an arbitrary given text into a spoken waveform. Main components of text to speech system are: Text processing and Speech generation.

A. SCRIPTS OF INDIAN LANGUAGES

The basic units of the writing system in Indian languages are Aksharas, which are an orthographic representation of speech sounds. An Akshara in Indian language scripts is close to a syllable and can be typically of the following form: C, V, CV, CCV, VC and CVC where C is a Consonant and V is a vowel

B. FORMAT OF INPUT TEXT

The scripts of Indian language are stored in digital computers in ISCII, UNICODE and in transliteration

scheme of various fonts. The input text could be available in any of these formats could be conveniently separated from the synthesis engine. An Indian language have a common phonetic base, the engines could be built for one transliteration scheme that can represent the script of all Indian language.

C. MAPPING OF NON-STANDARD WORDS TO STANDARD WORDS

In practice, an input text such as news article consists of standard words and non-standard words such as initials, digits, symbols and abbreviations. Mapping of non-standard words to a set of standard words depends on the context, and it is a non-trivial problem.

D. STANDARD WORDS TO PHONEME SEQUENCE

Generation of sequence of phoneme units for a given standard word is referred to as letter to sound rules. The complexity of these rules and their derivation depends on the nature of the language.

E. SPEECH GENERATION COMPONENT

Given the sequence of phonemes, the objective of the speech generation component is to synthesize the acoustic wave form. Speech generation has been attempted by concatenating the recorded speech segments. Current state-of-art speech synthesis generates natural sounding speech by using large number of speech units. Storage of large number of units and their retrieval in real time is feasible due to availability of cheap memory and computation power. The approach of using an inventory of speech units is referred to as unit selection approach. It can also be referred to as data-driven approach or example based approach for speech synthesis. The issues related to the unit selection speech synthesis system are: 1) Choice of unit size, 2) Generation of speech database, 3) Criteria for selection of a unit.

F. CHOICE OF UNIT SIZE

An inventory of larger size of units such as sentences, phrases and words could constitute an ideal speech database for speech generation. However, if the size of the units is large, the coverage of all possible words, phrases, proper nouns, and other foreign words may not be ensured.

G. GENERATION OF UNIT SELECTION DATABASES

There are two issues concerning the generation of unit selection databases. They are: 1) Selection of utterances which has the coverage of all possible units, 2) Recording of these utterances by a good voice talent.

III. CONCATENATIVE SYNTHESIS

In this approach synthesis is done by using natural speech. This methodology has the advantage in its simplicity, i.e. there is no mathematical model involved. Speech is

Table 2: Naturalness in terms of error

Subject	No. of sentences & words	No. of NSN words	% Error
S1	20 and 446	42	9.4
S2	20 and 446	35	7.84
S3	20 and 446	36	8.07
S4	20 and 446	38	8.52
S5	20 and 446	34	7.62
S6	20 and 446	28	6.27
S7	20 and 446	37	8.29
S8	20 and 446	43	9.64
S9	20 and 446	41	9.19
S10	20 and 446	33	7.39
Overall		367/4460	8.22

The error rate of TTS synthesizer is around 8.22%.

The unit selection database does require additional phonetic coverage for nouns and words which could occur more frequently. The studies also highlight need for good text processing component for Indian language synthesizers.

Paired Comparison Test

In order to evaluate the performance, the speech samples synthesized by the proposed method were compared with those made by the conventional method using phonemes as a database. The input Marathi texts for the conventional method are the same sentences as those for the proposed method.

The listeners were five males and three females without any known hearing problems. The speech samples were presented through loud-speakers in a sound-proof room. The listeners were asked to listen to the speech samples only once because the mean length of one sentence was very long (about ten seconds). The listeners were asked to judge which of the two samples of the same target sentence they considered to be more natural. They were not allowed to judge both samples of the pair equally good. Each speech sample of a pair was arranged in random order, and the order of the sentence pairs was randomized, too. The listeners took a rest intermittently.

The result of the paired comparison test is presented in table 3

Experimental Results

81% of synthesized speech by proposed method was evaluated as more natural speech than synthesized speech by the conventional method. Speech synthesis runtime was reduced

VI. CONCLUSION

In this paper, we discussed the issues relevant to the development of unit selection speech systems for Marathi language. It was observed that when the coverage of units is small, the synthesizer is likely to produce a low quality speech. As the coverage of units increases, it increases the quality of the synthesizer.

Table 3: Results of a paired comparison test

Listeners	Preference score (Naturalness) [%]	
	Proposed Method	Conventional Method
Male A	83	17
Male B	79	21
Male C	86	14
Male D	78	22
Male E	77	23
Female F	84	16
Female G	76	24
Female H	85	15
Total	81	19

We conducted subjective tests to evaluate the performance of speech synthesizer. From the perceptual results, it was observed that the word unit performs better than the phoneme units, and seems to be a better representation for languages such as Marathi.

In order to achieve greatest naturalness, the areas that need more attention are text analysis, prosody and creation of big speech database for concatenation synthesis.

VII. REFERENCES

[1] S.P. Kishore, A. W Black, Rohit Kumar, and Rajiv Sangal, "Experiments with unit selection Speech Databases for Indian Languages."

[2] Aniruddha Sen, "Speech Synthesis in India", IETE Technical Review, Vol 24, No 5, Sep-Oct 2007, pp 343-350.

[3] S. P Kishore and A. W. Black, "Unit size in Unit selection Speech Synthesis", Proceedings of EUROSPEECH, Geneva, Switzerland, 2003.

[4] S. P. Kishore, Rohit Kumar, and Rajeev Sangal, "A data – driven synthesis approach for Indian Languages using syllable as basic unit," in Proceedings of International Conference on National Language Processing (ICON), 2002.

[5] S. P. Kawachale and J. S. Chitode, "An Optimized Soft Cutting Approach to Derive Syllables from Words in Text to Speech Synthesizer", in proceedings Signal and Image Processing, 2006, pp 534.

[6] E. Veera Raghavendra, Srinivas Desai, B. Yegnanarayana, Alan W Black, Kishore Prahallad, "Experiments on Unit Size for Unit selection Speech Synthesis", Blizzard 2008.

[7] Hiroyuki Segi, Tohru Takagi and Takayuki Ito, "A Concatenative Speech Synthesis Method using Context Dependent Phoneme Sequences with variable length as a Search Units, Fifth ISCA Speech Synthesis Workshop- Pittsburgh

[8] Eric Lewis and Mark Tatham, "Word and Syllable Concatenation in Text to Speech Synthesis", Proceedings of sixth

[9] Eric Lewis, Mark Tatham and K Morton, "Syllable Reconstruction in Concatenated waveform Speech Synthesis", Proceedings of International Congress of Phonetic Sciences, pp2303-2306, ESCA, 1999.

[10] Hiroyuki Segi, Tohru Takagi and Takayuki Ito "A concatenative speech synthesis method using Context dependent phoneme sequences with Variable length as a search units", Fifth ISCA Speech Synthesis Workshop- Pittsburgh.

[11] Kalika Bali, Partha Pratim Talukdar, N. Sridhar Krishna, A.G.Ramakrishnan, "Tools for the development of a Hindi Speech Synthesis System", Fifth ISCA Speech Synthesis Workshop- Pittsburgh.

[12] Jerneja Zganec Gros and Mario Zganec, "An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis.