INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

PERSONALIZED ENGLISH SPEECH SYNTHESIZER USING CONCATENATIVE SYNTHESIS

RAHUL SAWANT, H.G VIRANI, CHETAN DESAI

1. Electronics and Telecommunication, Goa college of Engineering, Goa, India.
2. Associate Professor, Electronics and Telecommunication, Goa college of Engineering, Goa, India.
3. Assistant Professor, Electronics and Telecommunication, Goa college of Engineering, Goa, India.

Accepted Date:

27/02/2013

Publish Date:

01/04/2013

Keywords

Concatenative Speech Synthesizer,

Syllable,

Demi Syllable,

Listeners Test,

Uniqueness,

Progressive Speech Disorder.

Corresponding Author

Mr. Rahul Sawant

Abstract

In this paper we discuss about personalized speech synthesizer using Concatenative speech synthesis. The personalized synthesizer speaks aloud written text in an individual's voice using previously stored database of vernacular sub words. Unlike conventional speech synthesizers, our database consists of combination of syllables and demi syllables as speech units. A unique classification for this database has been proposed for faster computation and execution. Besides intelligibility and naturalness we have introduced the parameter of uniqueness in the listeners test to verify individuality in synthesized speech. Our synthesizer was evaluated subjectively by 10 listeners based on the listeners test and produced 96% unique, 93% intelligible and 87% natural synthesized speech.

I. Introduction

Speech is a vital form of communication in day to day life; however progressive speech disorder disturbs body's natural speaking ability. These can be problems in different parts of the speech production system; hence patients can suffer with articulatory breakdown, phonemic breakdown (difficulties with sounds) and other problems. However, it is rare for patients to have just one of these problems and most people have more than one problem. As the disease develops, speaking ability decreases and patient might become mute. A personalized speech synthesizer will help these people to communicate with others in their own voice.

II. Speech Synthesis System

Speech synthesis is the automatic generation of speech waveform from input text [1]. This computer based system is able to read any text aloud, whether it was introduced in computer by an operator or scanned and submitted to an optical character recognition (OCR) system. TTS starts with previously stored database of speech units by analysis of training data. During concatenation stored units are placed in proper sequence at runtime. Thus TTS consist of two main tasks: text processing and speech generation. In text processing, the input text is transcribed into a phonetic or some other linguistic representation, and in speech generation speech waveforms are generated from linguistic representation and prosodic information [2]. Speech synthesis can be classified into three categories [3]:

1) Formant synthesis
2) Articulatory synthesis
3) Concatenative synthesis

Articulatory synthesis tries to model human vocal organs as perfectly as possible but computational load is considerably higher than other methods. Formant synthesis is based on source filter model of speech. It follows set of rules which is used to determine parameters necessary to synthesize desired utterance. Concatenative speech synthesis is based on concatenation prerecorded natural sounding utterances [2]. Thus intelligibility and naturalness in synthesized speech is higher as compared to formant and articulatory synthesis.

A. Concatenation Synthesis

Synthetic speech is generated by a concatenation of speech units stored in a reference database. Stored speech waveforms of various durations are concatenated during speech generation. Earlier approach for Concatenative synthesis was storing a spectral template for each short sound and retrieve template as required. Spectral effects of vocal tract are simulated using dynamic digital filter. Later stored speech waveforms of different duration were concatenated to produce synthesized speech. Thus unlike spectral template, waveform concatenation eliminates need of filtering [1].

B. Database selection

Most important aspect of Concatenative synthesis is choosing correct unit length. Choice of longer unit length has high naturalness, less concatenation points and a good control of coarticulation, but amount of units required and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedure become more difficult and complex [2]. Choices of units for TTS are phonemes, diphones, triphones, demi syllables, syllables and words.

English language has around 40 phonemes which gives great flexibility. The major disadvantage of concatenating such brief sounds is coarticulation, which causes large changes in the articulation and acoustics of phonemes, usually on a scale of one to three phonemes (i.e., coarticulation effects rarely extend beyond three phonemes; e.g., rounded lips during /s/ in “strew”). While diphones have the same duration as phonemes, their storage is much greater: a language with N phonemes has N^{2}diphones and diphones have inherently more dynamic behavior, requiring storage of more frames/unit than phonemes (many of the latter are steady-state units, needing only one frame to represent each). A database of diphone units is still very feasible, needing only a few thousand frames of spectral data [1].

The number of different syllables in each language is considerably smaller than the number of words, but the size of unit database is usually still too large for TTS systems. For example, there are about 10,000 syllables in English. Unlike with words, the coarticulation effect is not included in stored units, so using syllables as a basic unit is not very reasonable. Demi syllables represent the initial and final parts of syllables. One advantage of demi syllables is that only about 1,000 of them are needed to construct the 10,000 syllables of English. Using demi syllables, instead of phonemes and diphones, requires considerably less concatenation points. Demi syllables take into account of most transitions and large number of coarticulation effects. It also covers a large number of allophonic variations due to separation of initial and final consonant clusters. However, the memory requirements are still quite high, but tolerable. Compared to phonemes and diphones, the exact number of demi syllables in a language cannot be defined. With purely demi syllable based system, all possible words cannot be synthesized properly [2]. With use of both syllable and demi syllable as database unit, size of the database is considerably reduced.

C. Database classification

Considering Konkani/Marathi/Hindi as the vernacular language of choice, Aksharasform the basic unit of writing system and also an orthographic representation of speech sounds [3]. These Aksharas can be classified as C, V, CV, CC, VC, CCV, VCC, CVC, CCCV, CVCC, CCCVC, CCVCC and CCCVVCC where C is consonant and V is a vowel.

III. Speech Synthesis Process

System consists of three main parts: the parser, database and concatenation unit as shown in figure 1.

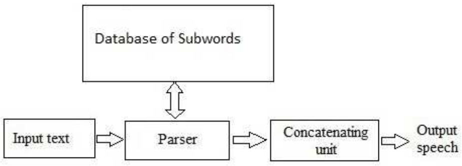

Fig. 1. Speech synthesis system

Courtesy: Chetan D. [4] and further edited

Synthesizer works as follows shown in figure 2

Step 1: Input text is given either by operator or OCR in standard form.

Step 2: Parser transcribes the text into the form of stored speech units.

Step 3: Checks for transcribed text in syllable or demi syllable list.

Step 4: Retrieves corresponding sound file from stored database

Step 5: Concatenate in sequence with addition of suitable silence between words.

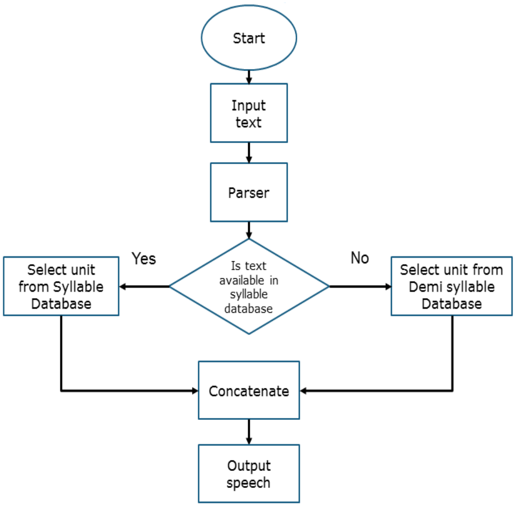
Start
Input text
Parser
Is text available in syllable database
Yes
No
Select unit from Syllable Database
Select unit from Demi syllable Database
Concatenate
Output speech

Fig. 2. Speech synthesis flow

A. Database creation

Database is created by reading training speech consisting of all possible sounds. Depending on type of speech recording user has to select whether to store in syllable or demi syllable database. Care should be taken to finish recording in one sitting to maintain prosody. Figure 3 shows flowchart of database creation.

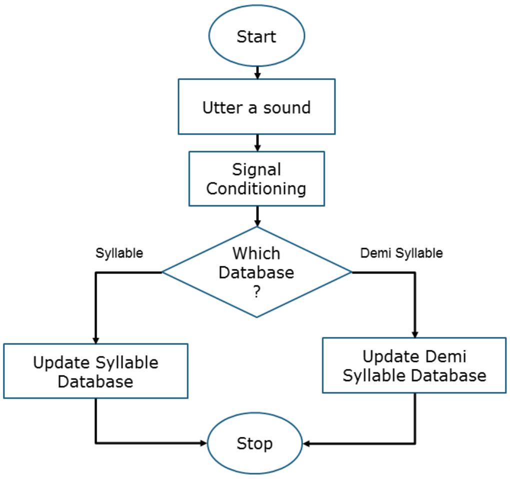

Fig. 3. Database creation

B. Boundary detection of syllable and demi syllable

Recorded speech units contain silence which needs to be removed before saving in database. A unique end point detection method is used to identify syllable boundaries and retain samples between boundaries [4]. Each syllable of a word is composed of an initial demi syllable, which comprises the initial consonant and the first part of the following vowel, plus a final demi syllable, which includes the remaining portions of the vowel and any following consonants. A modified version of above end point detection algorithm along with certain set of rules is used to extract initial and final demi syllable from the syllable and stored in demi syllable database.

C. Concatenation unit

Once the parser identifies link to the corresponding text, speech units are extracted from database and stored in sequence in cache memory of concatenation unit to be played at output. Synthesized output can be further processed and smoothed to get refined output which closely resembles target word.

IV. Performance Evaluation

Listener's test

Sentence directly recorded and one with Concatenative synthesis was played to 10 listeners. A typical multimedia PC with desktop speakers was used to play these utterances to 10 native speakers in calm environment room. The test subjects were around 20 to 25 years of age with no prior experience in speech synthesis experiments. Test was carried on three parameters: naturalness, intelligibility and uniqueness. Listeners were told to score each category from the scale of 1 to 5, where 5 is perfect and 1 is not acceptable. In order provide wider scale listener was given freedom to score in between like 1.5, 2.5, 3.5 and 4.5. Table 1 shows listeners score for the developed system.

TABLE I. LISTENERS TEST

Listeners Test			
Listener	Uniqueness	Naturalness	Intelligibility
1	5	4	4.5
2	4	3	4
3	4.5	4.5	5
4	5	5	4
5	5	5	4.5
6	5	4	4.5
7	5	5	5
8	5	5	5
9	5	4	5
10	4.5	4	5

Result shows that the synthesized output to be 96% unique, 93% intelligible and 87% natural when compared with direct recorded sound. Naturalness is acceptable but needs some improvement.

Word recognition test

Word recognition test was also conducted on words formed using demi syllables. 10 listeners were told to hear concatenated demi syllable words and predict the heard sound. Total of 9 words were played and evaluated. Table 2 shows listeners' score

TABLE II. WORD RECOGNITION TEST

Word recognition test	
Listener	Words recognized
1	4
2	5
3	3
4	3
5	3
6	4
7	6
8	3
9	3
10	4

Word recognition test shows on an average 4 words were able to be identified from 9 played words using demi syllable concatenation without using signal processing technique.

V. CONCLUSION

In this paper the proposed personalized TTS will be helpful for people with progressive speech disorder to retain their voice in communication. Use of syllable and demi syllable in database has reduced number of speech units retaining the naturalness and intelligibility of synthesized speech. With classification of database it has become easier to retrieve stored units and reduced computational time as compared to time taken by linear search. Presently system is tested on few sentences and requires spectral smoothing to increase naturalness of synthesized speech. Using PSOLA methods or hybrid model method the quality of synthesized speech can be increased [5].

References

1. O' Shaughnessy D, Speech Communication - Human and Machine Addison Wesley New York J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

2. Sami Lemmetty, Thesis-Review of Speech Synthesis Technology, Helsinki University of Technology

3. S.D Shirbahadurkar, D.S Bormane, Speech synthesizer using Concatenative synthesis strategy for Marathi language

(spoken in Maharashtra, India), International journal of recent trends in engineering, Vol 2, No. 4, November 2009

4. Chetan D., Vanessa C., “Development of a personalized integrated speech-to-text cum text-to-speech generator for English language”, Goa college of Engineering June 2008

5. Fabio Violaro, Olivier Boeffard, “A Hybrid model for Text-to-Speech synthesis”, IEEE Transactions on speech and Audio Processing, vol. 6, No 5, September 1998