Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review of Unit Selection Speech Synthesis

Sangramsing Kayte
Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

Monica Mundada
Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

Dr. Charansing Kayte
Assistant Professor
Department of Digital and Cyber Forensic, Aurangabad
Maharashtra, India

Abstract— Speech is used to express information, emotions, and feelings. Speech synthesis is the technique of converting given input text to synthetic speech. Speech synthesis can be used to read text as in SMS, newspapers, site information etc. and can be used by blind people. Speech synthesis has been widely researched in last four decades. The quality and intelligibility of the synthetic speech produced is remarkably good for most of the applications. This report intends to review four majorly researched methods of speech synthesis viz. Articulatory, Concatenated, Formant, and Quasi-articulatory Synthesis. Mainly in this paper focus is given on concatenate synthesis method and some issues of this method are discussed. Articulatory Synthesis is based on human speech production model. The synthetic speech produced by this model is most natural, but it is also the most difficult method. Concatenate Synthesis uses prerecorded speech words, phrases and concatenates them to produce sound. It is the simplest method and yields high-quality speech but is limited by its memory requirement to store beforehand all possible words, phrases to be produced. Formant Synthesis is based on the acoustic model of the human speech production system. It models the sound source and the resonance in the vocal tract, and is most common model used. Quasi-articulatory Synthesis is a hybrid of articulator acoustic model of speech production. Synthetic speech produced by this model sounds more natural and can be easily customized to meet different requirements of different applications and individual users.

Keywords— Unit selection Speech synthesis, articulatory synthesizer, formant synthesizer, concatenative synthesizer.

I. INTRODUCTION

Unit selection synthesis is also referred as corpus based synthesis. It uses large database. During database creation, each recorded utterance is segmented into some individual phones, syllables, morphemes, words, phrases, and sentences. An index of the units in the speech database is then made based on the segmentation and acoustic parameters such as fundamental frequency, pitch, duration, the status of the syllable and previous and next phones. This method provides naturalness in output speech as compared to other techniques. Speech synthesis is a process of automatic generation of speech by machines/computers. The goal of speech synthesis is to develop a machine having an intelligible, natural sounding voice for conveying information to a user in a desired accent, language, and voice. Unit selection synthesis shown in Fig.1 is a type of concatenative synthesis in which the largest matching sound file available in the speech corpus is concatenated for synthesis of target speech. It is capable of managing large number of units [1], also imparts prosody beyond the role of F0. It is quite necessary to make a clear distinction between role of F0 and Pitch: F0 is the actual frequency generated by the vocal cord or vocal fold, while Pitch is the perception of that frequency by the listener. Hence it not necessary that both are equal.This synthesis technique also retains the naturalness in the speech sounds being generated. Choosing unit length is an important task in Concatenative speech synthesis. A shorter unit length requires less spacebut sample collecting and labeling becomes more difficult and complex. A longer unit length gives more naturalness [2], better coarticulation effect and less concatenation points but requires more memory space. Choices of unitfor TTS are phonemes, diphones, triphones, demi syllables, syllables and words [3][4].

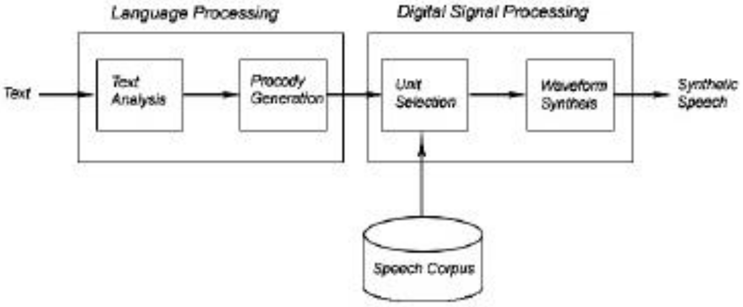

Fig. 1 Unit Selection Synthesis system

Unit-selection speech synthesis has become increasingly popular due to its enhanced prosodic quality and naturalness when compared to parametric or diphone synthesizers. The principle is based on the concatenation of naturally-produced speech units of variable length, avoiding signal manipulation at the concatenation points as far as possible. A sufficiently high quality can be reached when several instances of segments with different intonation are contained in the unit database; in that case, optimum units can be selected by minimizing intrinsic unit costs as well as concatenation costs

A. Advantages of USS

1. By now, you've probably figured out that USS produces the most natural sounding speech.
2. Preservation of the Original Actor's Voice: The Text-to-Speech Engine chooses speech units that best fit the text you have typed in. USS involves pulling these speech units directly from the voice database, thus preserving the original voice of the actor at all times.
3. Higher quality audio files are produced: The TTS engine has at least 20 hours of recorded voice to choose from when it approaches the database to pull our speech units that match your text. This means that the audio files produced are much better quality than HTS audio files.
4. Sophisticated techniques can be used to smooth the joins between speech units to make each sentence sound as natural as possible.

B. Disadvantages of USS

1. Very long recording and development time: And I mean really long. Every new voice actor needs to record at least 20 hours of audio. Sometimes we need 60 hours of recording – that's a lot of talking! Then, the development of the TTS engine can take many months, even years.
2. Large database / footprint size: Because we need so many more hours of recording, the database that hosts all the audio files needs to be much bigger than the database for an HTS based voice.
3. Because USS does not tamper with the original actor's voice, it is impossible to change the emotion of the voice once the voice has been made. This means we cannot change the voice to be sad, happy or angry. Instead we would have to pre-record the actor saying each sentence with an angry or happy tone and then develop the voice using those audio files.

II. UNIT SELECTION PROCESS

The speech synthesis part accepts information from prosody generation part, retrieves the speech unit database to find a proper template for every target speech unit. During the selection process, the phonetic and prosodic constraints are applied. The smoothness of the concatenation is also concerned.

The unit selection process can be illustrated as Figure 2. In the figure, the target sentence is (it is very hot today)", which consists of 4 syllables (jin1, tian1, hen3, re4). Each syllable has a set of candidate units. The thick line and thick edge box indicate the selected unit sequence. In unit selection process, to get the best speech, we have to consider (17) the appropriateness of the candidate unit compared with target unit, (18) the smoothness between the selected units to be connected. Therefore, the selection process is to find a best path among all the possible paths in the connection lattice. The search process is guided by a cost function, which describes the degree of appropriateness of a unit and degree of smoothness between two units.

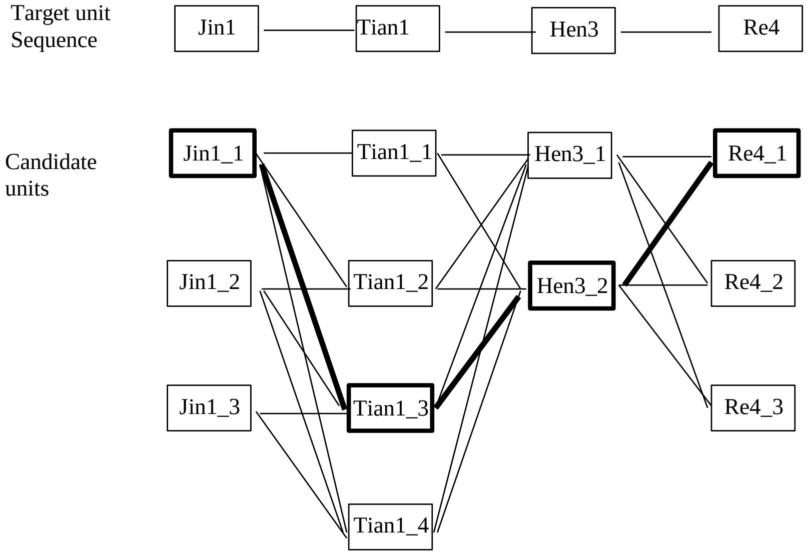

Figure 2. Illustration of unit selection

III. UNIT SPECIFICATIONS

In this work, we choose syllable as the synthesis unit. The reason to choose syllable is that syllable is a relatively stable unit. The coarticulation between syllables is relatively loose, while the coarticulation between sub-syllable units is very tight. Each unit is specified by a feature vector, which is used for matching in a unit selection process. Both the target units and units in inventory are described using this feature vector. The features describe the phonetic identity, phonetic context, break types around the unit, and prosody parameters of each unit. The detailed features are as the following:

1 Phonetic identity of the unit: Using the pronunciation of the unit is to ensure that the candidate unit will have the same sound as the expected one. The pronunciation includes the initial, final and tone. There are 22 initials, 38 finals, and 5 tones defined in this work.
2 Phonetic context: The co-articulation between two units is determined by the phonetic identity of its neighbours. The context of the unit will help find the unit with similar context of a unit. The phonetic context consists of the initials, finals, and tones of the previous and next units.
3 Breaks around the unit: The break types before and after the unit. The prosodic properties of a unit before a break and after a break are quite different. The break type information is an important index to evaluate the similarity of two units. The types we defined include syllable break, word break, and prosodic word break.
4 Prosody parameters: The prosody parameters are a collection of parameters that describe the duration, pitch contour and energy of a unit. We defined 10 parameters to describe the duration, pitch contour and energy of each unit.

IV. SYLLABLES AS THE BASIC UNIT

Indian languages are syllable centered, where pronunciations are mainly based on syllables. A Syllable can be the best unit for Indian language Speech synthesis systems. Intelligible speech synthesis is possible for Indian languages with syllable as the basic unit. Syllable units being larger in comparison to phones or diphones, can capture co-articulation better than phones. The number of concatenation points decreases when syllable is used as the basic unit. Syllable boundaries are characterized by regions of low energy, providing more prosodic information. A grapheme in Indian languages is close to a syllable. The general format of an Indian language syllable is C*VC*, where C is a consonant, V is a vowel and C* indicates the presence of 0 or more consonants. There are about 35 consonants and 18 vow els in Indian languages [5]. There are defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. A rule based graphe me to syllable converter was used for syllabification. Some of the rules used to perform grapheme to syllable conversion [6] are:

- Nucleus can be Vowel (V) or Consonant (C)
- If onset is C then nucleus is V to yield a syllable of type CV
- Coda can be empty or C
- If characters after CV pattern are of type CV then the syllables are split as CV and CV.
- If the CV pattern if followed by CCV then syllables are split as CVC and CV.
- If the CV pattern is followed by CCCV then the syllables are split as CVCC and CV
- If the VC pattern is followed by V then the syllables are split as V and CV.
- If the VC pattern is followed by CVC then the syllables are split as VC and CVC

V. POLYSYLLABLES AS THE BASIC UNIT

An attempt exploring a Speech synthesizing system using polysyllabic units has been made [8]. Since polysyllable units are formed using the monosyllable units already present in the database, the synthesis quality can be improved without augmenting any new set of units. The system uses a large database, which consists of syllables, bisyllables and trisyllables. While synthesizing, the first matching trisyllable is selected followed by the bisyllable and monosyllable units, as needed. Picking up the largest possible unit in the database improves the quality of speech, since the number of co-articulation points greatly reduce. Selection of an appropriate candidate unit set is carried out using search algorithms. Units which incur lowest total join cost for a word are preferred. Total join cost is the sum of selection cost of the candidate unit and the join cost of the selected candidate units [9].

$$Ctotal= Csel + CJoin \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

where, Ctotal is the total join cost, Csel is the selection cost and CJoin is the cost of join
Researches were conducted to find which order of syllables is best acceptable for synthesis[7].
The following are the recommended combinations:

a) Monosyllables at the beginning of a word and bisyllables at the end.
b) Bisyllables at the beginning of a word and monosyllables at the end.
c) Monosyllables at the beginning and trisyllables at the end of a word.
d) Trisyllables at the beginning and monosyllables at the end of a word.

VI. UNIT SELECTION PROBLEM

Unit Selection involves finding the best sequence of unit instances which is closely matching with a given target specification of the required unit sequence in terms of features. The best instance is decided by minimizing target cost between features specification of desired unit and available instances of the unit and the joining costs between the selected instances in the sequence of units [11]. In order to synthesize a sequence of 15 units (nearly 5 words) by selecting units from a typical database having on an average 10 occurrences of each basic unit, we have 10 possible sequences. As we move towards bigger and bigger databases for natural sounding and emotional synthesis, efficient algorithms for unit selection need to be explored. Essentially the Unit Selection problem is a heuristic search problem involving optimization of selection costs of each unit across the sequence of units to reach a minimum. We can either reach the optimization by doing a local optimization or by attempting a global optimization. A deeper analysis of the nature of the Unit Selection problem can let us understand which approach is better in reaching optimal results efficiently. We have experimented with local optimization in [12]. Global optimization is a search problem in very large high dimensionality search space. Enumerative/Brute force search is not suggested considering that Speech Synthesis would often want a real time speed performance. Genetic Algorithms (GA) have been widely successful for solving global optimization problems in huge search spaces [13]. Through this paper, we describe a GA implemented by us for the Unit Selection problem. Before we proceed, we look at some of the unit selection algorithms being used in popular systems. [11] and [14] describes a unit clustering and a pruned viterbi search based unit selection that chooses a best path through a state transition network so as to minimize unit distortion (i.e. target cost) as well as concatenation distortion (i.e. join cost). This algorithm is used in Festival and CHATR systems. A 3-Tier Non-Uniform Unit Selection Algorithm used in the Microsoft China Mandarin TTS is described in [15] which reduces the choices for each unit in the sequence at each tier based on feature distances and concatenation costs. Further this paper is organized as follows: Section II introduces the basic Genetic Algorithm. In section III we describe the implementation of the various genetic operators for the Unit Selection problem. Section IV presents the perceptibility comparison between local optimization approach and the GA approach.

VII. UNIT SELECTION SPEECH DATABASE

The speech database we are using is developed from the Hindi, Marathi, English databases described in [11]. It consists of basic units of varying sizes at syllable and phone levels. Each instance of the units is stored along with prosodic and linguistic features like pitch, duration, energy, phonetic context and syllable position in the word. The instances of each unit are further stored in the increasing order of their global prosodic mismatch function (GPMF) [16] value. The GPMF is an objective function that we use to describe the suitability of an instance of a unit in the most probable situations the unit might be used.

VIII. UNIT SELECTION FRAMEWORK

Different unit selection algorithms are implemented as different synthesis options in the SPACE synthesizer. The following options are currently available: diphone synthesis [8], "standard" unit selection synthesis (explained below), and our unit selection synthesis algorithm (experimental option) which is explained later. The diphone synthesis option synthesizes an input text by combining single diphone candidates as required and there is no selection involved. The standard unit selection synthesis option evaluates possible combinations of candidate units which are either diphones orphones and selects the best combination using a cost function based on both target and join costs. Within this framework,the different synthesis options can share part of or the whole speech database, and also the selection cost function and the associated implementation if necessary.

IX. CONCLUSION

For Text to speech conversion the Unit selection speech synthesis is the simplest method where phonemes which are called units. The unit plays important role. However, unit selection synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. The most important aspects in unit selection synthesis is to find correct unit length. With longer unit’s high naturalness, less concatenation points and good control of unit selection are achieved, but the amount of required units and memory is increased. This method also has a problem of pitch differences of units and also spectral discontinuities.

REFERENCES

[1] Rahul Sawant, H.G Virani, and Chetan Desai, “Database selection for Concatenative speech synthesis With novel endpoint detection Algorithm”,IJAIEM, Volume 2, Issue 5, May 2013, pp.173-180.

[2] JernejaZganecGros and Mario Zganec, ”An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis Systems”,Journal of Computing and Information Technology, 2008, pp. 69 – 78.

[3] Hiroyuki Segi, Tohru Takagi and Takayuki Ito, ” A concatenative speech synthesis method Using context dependent phoneme sequences With variable length as search units”,5th ISCA Speech Synthesis Workshop Pittsburgh, PA, USA June 14-16, 2004, pp.116-120.

[4] MunkhtuyaDavaatsagaan, and Kuldip K. Paliwal, “ Diphone-Based Concatenative Speech Synthesis System for Mongolian”, IMECS, March, 2008, Hong Kong, pp. 19-21

[5] G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and R Prathibha, “A Complete Text-To-Speech Synthesis System In Tamil”, in 0-7803-7395-2/02, IEEE proceedings 2002.

[6] S. Saraswathi and T.V. Geetha, “Design of language models at various phases of Tamil speech recognition system”, International Journal of Engineering, Science and Technology Vol. 2, No. 5, 2010, pp. 244-257.

[7] T.Jayasankar, R.Thangarajan, J.Arputha Vijaya Selvi, “Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis”, in International Journal of Computer Applications (0975 – 8887), Volume 25– No.1, July 2011.

[8] Vinodh M Vishwanath, Ashwin Bellur, Badri Narayan K, Deepali M Thakare, Anila Susan, Suthakar N M and Hema A Murthy,“Using Polysyllabic units for Text to Speech Synthesis in Indian languages,” Proceedings of National Conference on Communication (NCC),pp.1-5, 29-31 Jan. 2010

[9] Alan W. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis”, Proc. EUROSPEECH 97, Rhodes, Greece, 1997, Vol. 2, pp. 601-604. K. Elissa, “Title of paper if known,” unpublished.

[10] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, "Text-to-speech synthesis using syllable-like units," in National Conference on Communication, Kharagpur, India, Jan 2005, pp 277-280

[11] Alan W. Black and Nick Campbell, "Optimizing selection of Units from Speech Databases for Concatenative Synthesis", In Proceedings of Eurospeech 95, vol 1., pp. 581 – 584, Madrid, Spain, 1995

[12] S P Kishore, Rohit Kumar and Rajeev Sangal, "A Data Driven Synthesis Approach For Indian Languages using Syllables as Basic Unit", in Proceedings of Intl. Conf. on NLP (ICON) 2002, pp. 311-316, Mumbai, India, 200

[13] Lawrence Davis ed., "Handbook of Genetic Algorithms", Van Nostrand Reinhold, New York, 1991
Alistair Conkie, "A robust unit selection system for speech synthesis", in Joint Meeting of ASA/EAA/DAGA, Berlin, Germany, March 1999

[14] Min Chu, Hu Peng, Hong-yun Yang, Eric Chang, "Selecting Non-Uniform Units from a very large corpus for Concatenative Speech Synthesizer", in Proceedings of ICASSP, Salt Lake City, 2001

[15] Rohit Kumar, S. P. Kishore, "Automatic Pruning of Unit Selection Speech Databases for Synthesis without loss of Naturalness", submitted to ICSLP, Jeju Island, Korea, 2004

[16] Bigorgne, D., Boe.ard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J.L., Metayer, I., Sorin, C., and White, S., 1993. Multilingual PSOLA text-to-speech system. In: Proc. ICASSP, pp. II.187-190.

[17] Chan, N. C. and Chan, C. Prosodic Rules for Connected Mandarin Synthesis. J. Inform. Sci. Eng. 8, 261-281. 1992

[18] [Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015 Impact Factor: 1.492

[19] Sangramsing Kayte, Dr. Bharti Gawali "Marathi Speech Synthesis: A review" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

[20] Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT) (IMPACT FACTOR: 3.32)

[21] Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014 (IMPACT FACTOR: 2.080)