ISSN 2319 – 2518 www.ijeetc.com
Vol. 3, No. 2, April 2014

Research Paper

A COMPREHENSIVE REVIEW ON CONCATENATION BASED TEXT TO SPEECH SYNTHESIS FOR INDIAN LANGUAGE

Arun Kumar C[1]* and Shreekanth T[1]

*Corresponding Author: Arun Kumar C, ✉ arun07837@yahoo.com

The goal of Text-To-Speech (TTS) synthesis system is to convert an arbitrary input text to intelligible and natural sounding speech so as to transmit information from a machine to a person. In the present world of human computer interaction the visually impaired community in India and other developing countries are deprived of technologies that could help them to communicate with the sighted world. In this view many Text-To-Speech (TTS) systems have been developed. This review traces the earlier works on the development of TTS system using Concatenation based speech synthesis system. Concatenative speech synthesis systems form utterances by concatenating pre-recorded speech samples of different unit length. The quality of synthesized speech obtained from approximate matching of syllables and direct waveform concatenation will be of better quality and natural, when compared to Pitch Synchronous Overlap and Add (TD-PSOLA) and Harmonic plus Noise Model (HNM) technique.

Keywords: Speech synthesis, Concatenative synthesis, Dynamic Time Wrapping (DTW), Frequency Domain (FD), Harmonic plus Noise Model (HNM), Mean Opinion Score (MOS), Pitch Synchronous Overlap and Add (PSOLA), Time Domain (TD)

INTRODUCTION

The ultimate goal of Text-To-Speech (TTS) synthesis is to convert an ordinary orthographic text into an acoustic signal that is indistinguishable from human speech (Marian Macchi, 1993). This generally involves two steps:

1. Text processing.
2. Speech generation.

The objective of the text processing component is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus (Kishore *et al.*, 2002; and Klatt, 1987). For natural sounding speech synthesis, it is essential that the text processing component

[1] Department of E&C, SJCE, Mysore, Karnataka, India.

produce an appropriate sequence of phonemic units corresponding to an arbitrary input text (Paul Taylor, 2009).

The conversion of words in written form into speech is nontrivial. Moreover, in order to sound natural, the intonation of the sentences must be appropriately generated. Synthesis of speech cannot be accomplished by cutting and pasting smaller units together. Attention has to be paid to Smoothing out the discontinuities in such a process so that the resulting signal approximates natural speech (Ravi and Sudarshan Patilkulkarni, 2011). According to the speech generation model used, speech synthesis can be classified into three categories as Articulatory synthesis, Formant synthesis and Concatenative synthesis (Lemmety, 1999).

- Articulatory synthesis, which attempts to model the human speech production system directly.
- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
- Concatenative synthesis, which uses different length prerecorded samples derived from natural speech (Lemmety, 1999).

The articulatory method is still too complicated for high quality implementation as it is very difficult to track and simulate human vocal organ i.e. vocal tract and vocal fold. Vocal tract modulates air flow by varying position and shape of mouth and vocal fold excites vocal tract when air expelled out of lungs (Lemmety, 1999). The formant and concatenative methods are the most commonly used speech synthesis techniques. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular (Thomas, 2007).

This paper discuss the issues in selecting the best algorithm among the concatenative speech synthesis technique for developing Text-To-Speech (TTS) synthesis system for Indian language. The best algorithm will be chosen using comparative Mean Opinion Score (MOS) results obtained from the audience.

CONCATENATIVE SYNTHESIS

Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods (Lemmety, 1999). In present system units used are usually words, syllables, demisyllables, phonemes, diphones and sometimes even triphones. Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary (Kishore and Black, 2003).

Concatenation of words is relative easy to perform and coarticulation effects within a word are captured in the stored units. However, there is a great difference with words spoken in isolation and in continuous sentence which makes the continuous speech to sound very unnatural. Because there are hundreds and thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system. 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 (Lemmety, 1999).

There is also no way to control prosodic contours over the sentence. The current synthesis systems are mostly based on using phonemes, diphones, demisyllables or combination of these units. Demisyllable represents the initial and final parts of syllables. One advantage of demisyllable is that only about 1,000 of them are needed to construct 10,000 syllables for English. However the memory requirements are still quite high, but tolerable. One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade of between longer and shorter units (Lemmety, 1999). With longer units high naturalness, coarticulation effect and less concatenation points are achieved this will remove glitches between each character as much as possible and increases naturalness but the amount of memory required to store database increases compared to other synthesis technique. With shorter units like phones and diphones acquire less memory but sample recording, labeling procedure become more tedious job (Kishore and Black, 2003). Hence in present systems units used are usually monosyllables, bisyllable and polysyllable to increase quality of synthesized speech signal.

RELATED WORK

The Pitch Synchronous Overlap and Add (PSOLA) method was originally developed at France Telecom (CNET). It allows prerecorded speech samples smoothly concatenated and provides good controlling for pitch and duration (Lemmety, 1999).

There are several versions of the PSOLA algorithm and all of them work in essence the same way. Time-domain version, TD-PSOLA, is the most commonly used due to its Computational efficiency. The basic algorithm consists of three steps.

a. The analysis step, original speech signal is first divided into separate but often overlapping short-term analysis signals.

b. Modify each analysis signal to synthesis signal.

c. All segmented samples are recombined by means of overlap and addition technique.

Short-term signals $x_m(n)$ are obtained from digital speech waveform $x(n)$ by multiplying the signal by a sequence of pitch-synchronous analysis window $h_m(n)$.

$$x_m(n) = h_m(t_m - n)x_n$$

where,

m is index for short-time signal.

$x_m(n)$ is short term signal.

x_n is digital speech waveform.

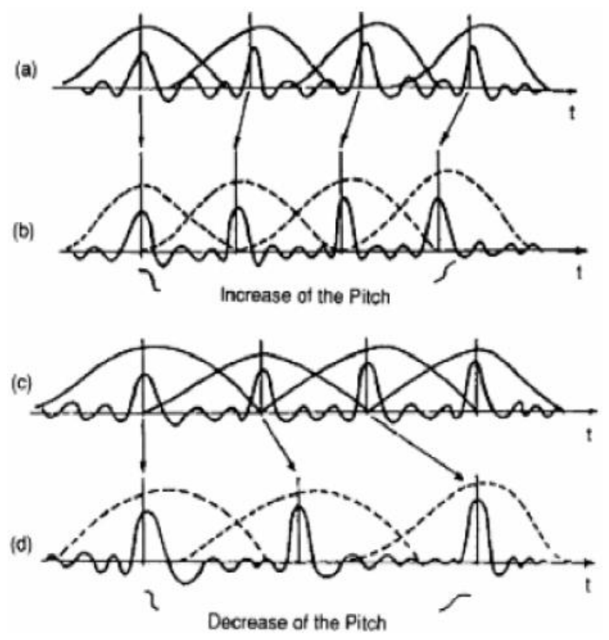
h_m is pitch synchronous analysis window.

t_m is pitch-marks.

Pitch marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts. The used window length is proportional to local pitch period and the window factor is usually from 2 to 4. The pitch markers are determined either by manually inspection of speech signal or automatically by some pitch estimation methods. The segment recombination in synthesis step is performed after defining a new pitch-mark sequence (Lemmety, 1999).

Manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers as shown in Figure 1. The modification of duration is achieved by either repeating or omitting speech segments. In principle, modification of fundamental frequency also implies a modification of duration.

Figure 1: Pitch Modification of a Voiced Speech Segments



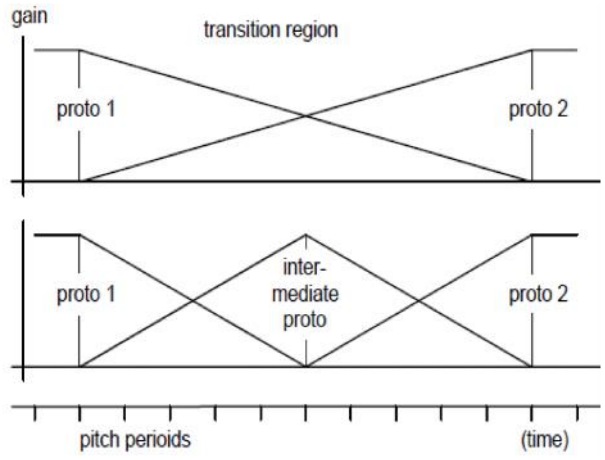
Source: Lemmety (1999)

Frequency Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA) are theoretically more appropriate approaches for pitch-scale modifications because they provide independent control over the spectral envelope of the synthesis signal. FD-PSOLA is used only for pitch-scale modifications and LP-PSOLA is used with residual excited vocoders (Lemmety, 1999; and Yannis Stylianou, 2001).

Micro phonemic method is used for variable length units derived from natural speech. The units may be words, syllables (monosyllable, bisyllable and polysyllable), phonemes (allophones, diphones and halfphones), Pitch periods, Transients or noise segments. Based on the unit selected for concatenation a segments of particular units (Prototype) are collected (Lemmety, 1999).

These collected units are concatenated in time axis with PSOLA technique. If the formant distances between consecutive sound segments are less than two critical bandwidths (Barks), the concatenation is made by simple linear amplitude-based interpolation between the prototypes. If the difference is more than two Barks, an extra intermediate prototype must be used because the simple amplitude-based interpolation is not sufficient for perceptually acceptable formant movements (Lemmety, 1999). The overlap-add process of prototypes are shown in Figure 2.

Figure 2: Linear Amplitude Based Interpolation



Source: Lemmety (1999)

Special attention should be provided for some consonants, i.e., stop consonants can be stored as direct waveform segments as several variants in the different vowel context. Considering fricatives as prototype of about 50 ms of total length and 10 ms units from that prototype is randomly selected for concatenation with this above technique (Lemmety, 1999).

Ravi and Sudarshan Patilkulkarni (2011) have developed the speech database using PRAAT utility software. This system developed is basically for phoneme level for Kannada language. The entire phone set of Kannada language are recorded as mono sound using PRAAT utility software and saved as .wav files in database based on their UNICODES and decimal equivalent. After that direct waveform concatenation technique is used to develop TTS system (Ravi and Sudarshan Patilkulkarni, 2009).

Concatenating and modifying the prosody of speech units without introducing audible artifacts by appropriate linguistic design of the text corpus and careful preparation of the speech database. Manually editing the WAV files and direct waveform concatenation to identify the right constituent segments for any word in the vocabulary, the appropriate segments are concatenated programmatically to yield the synthesized speech. Sentences could also be synthesized with the prosody corresponding to those embedded in the segments (Ravi and Sudarshan Patilkulkarni, 2011).

Duration is one of the prosodic features of speech, the other two being stress (intensity) and intonation (pitch). Generating prosodic features from text is one of the most difficult problems faced by current TTS systems. Most TTS systems generate prosodic features by rules, based on the linguistic information. However, making such rules is an extremely complex and human-dependent task (Ravi and Sudarshan Patilkulkarni, 2009; and Paul Taylor, 2009).

A variety of methods for the optimum selection of units have been proposed. The target cost and a concatenation cost are attributed in each candidate unit. The target cost is calculated as the weighted sum of the differences between elements such as prosody and phonetic context of the target and candidate units. The concatenation cost is also determined by the weighted sum of cepstral distance at the point of concatenation and the absolute differences in log power and pitch. The total cost for a sequence of units is the sum of the target and concatenation costs. Then, optimum unit selection is performed with a Viterbi search. Even though a large speech database is used, it is still possible that a unit with a large target and/or concatenation cost has to be selected because a better unit (prosody) is lacking. This results in a degradation of the output synthetic speech. Moreover, searching large speech databases can slow down the speech synthesis process (Hunt and Black, 1996; and Paul Taylor, 2009).

In the context of Harmonic plus Noise Model (HNM), speech signals are represented as a time-varying harmonic component plus a modulated noise component. The decomposition of a speech signal into these two components allows for more natural-sounding modifications of the signal. The parametric representation of speech using HNM provides a straightforward way of smoothing discontinuities of acoustic units around concatenation points. Formal listening tests have shown that HNM provides high-quality speech synthesis while outperforming other models for synthesis in intelligibility, naturalness, and pleasantness (Yannis Stylianou, 2001).

Concatenation based speech synthesis system using approximate matching of syllable provides better sounding speech than diphone and phone, the coverage of all syllables is a non-trivial issue (Veera Raghavendra *et al.*, 2008).

All Indian language scripts have a common phonetic base, and a universal phoneset consists of about 35 consonants and about 15 vowels. Theoretically possible syllable combinations in an Indian language are V, CV, CCV, CVC, and CCVC. The reason for using syllable as a basic unit is, larger the unit length lesser will be concatenations and reduces the co-articulation effects during the synthesis. Text to- speech synthesis based on syllables seems to be a good possibility to enhance the quality of synthesized speech with comparison to diphone based synthesizers (Hunt and Black, 1996; and Kishore and Black, 2003).

The preliminary attempt to add emotion to concatenative synthetic speech is successfully done in this work. A new intonation contour (including both pitch and duration changes) was applied to the concatenated segments during production of the final audible utterance, and some of the available synthesis parameters were systematically modified to increase the affective content. The output digital speech samples were then subject to further manipulation with a waveform editing package, to produce the final output utterance. The results of this process were a small number of manually-produced utterances, but which illustrated that affective manipulations were possible on this type of synthesizer (Iain *et al.*, 2000).

Text to Speech system developed by Sangamitra Mohanty describes syllable based Indian language TTS system which tells us how different syllable from the text are chunked and kept in database. Individual polysyllables are also stored in database. All these are .wav files and named after C, V, CV, VC, etc. Concatenative algorithm is developed using Visual C++ and codes written for phone based concatenation is modified with respect to syllable (Hunt and Black, 1996; and Sangamitra Mohanthy, 2011).

Database creation is one of the major processes in TTS system design. For the study and analysis of prosody, a number of sentences that will compose of high frequency components are extracted from phonetically rich text corpus. The corpus was designed such a way that each phoneme resides in various positions in a word initial, medial, and final position of occurrence in that way the extraction of them is possible and can be used as a structural element in a Text-To-Speech system (TTS) inventory. Using their phonetic transcription sentences were segmented using PRAAT utility software (Boersma and Weenink, 2001).

The work in Ravi and Sudarshan Patilkulkarni (2009) explains in detail the phoneme-based concatenative TTS system, with sufficient degree of customization and which uses linguistic analysis to circumvent most of the problems of existing concatenative systems.

PERFORMANCE EVALUATION

Quality of the speech is subjective in nature, speech which appears good to one person may not appear good to other, and hence collecting Mean Opinion Score (MOS) is better option for testing speech quality. The MOS that is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality and 5 is the highest perceived quality. The MOS is generated by averaging the results of a set of standard, subjective tests where a number of listeners rate the perceived audio quality of test sentences read aloud by both male and female speakers over the communications medium being tested. A listener is required to give each sentence a rating using the rating scheme in Table 1. The perceptual score of the method MOS is calculated by taking the mean of the all scores of each sentence (Ravi and Sudarshan Patilkulkarni, 2009).

Table 1: Mean Opinion Score (MOS)

MOS	Quality	Impairment
1	Bad	Very Annoying
2	Poor	Annoying
3	Fair	Slightly Annoying
4	Good	Perceptible
5	Excellent	Imperceptible

Source: Ravi and Sudarshan Patilkulkarni (2009)

Each listener is subjected to MOS, i.e., score between 1 (worst) to 5 (best) and AB-Test, i.e., the same sentence synthesized by two different synthesizers is played in random order and the listener is asked to decide which one sounded better (Kishore and Black, 2003). The MOS obtained from different techniques are shown in Table 2. Intelligibility provides, how well the synthesized speech understands to human ear. This is shown in Table 3.

Table 2: MOS Evaluated for Speech Quality

No.	Algorithm	MOS
1	Direct Wave Concatenation	4.50
2	Symbol Based Concatenation	3.82
3	HNM	3.45
4	TD-PSOLA	3.14
5	Approximate Matching of Syllable	3.08
6	Microphonemic	Not Known

Table 3: MOS Evaluated for Intelligibility

No.	Algorithm	MOS
1	DTW	4.18
2	Vowel Classification	4.32
3	HNM	3.98
4	TD-PSOLA	Not Known
5	Approximate Matching of Syllable	3.10
6	Microphonemic	Not Known

A number of subjective tests are used to measure the success of TTS system. Mean Opinion Score (MOS) test is carried out to examine the naturalness of the synthesized output of TTS system as shown in Table II. From the above tabulation we can clearly view that from informal and formal listening tests, Direct wave concatenation (4.5) (Ravi and Sudarshan Patilkulkarni, 2012), Harmonic plus Noise model (HNM) (3.45) (Yannis Stylianou, 2001), Time Domain – Pitch synchronous overlap and add (TD-PSOLA) (3.14) (Yannis Stylianou, 2001) and approximate matching of syllables (3.08) (Veera Raghavendra *et al.*, 2008) and Microphonemic technique (Not known) MOS are obtained. Hence from evaluated MOS we can clearly say that direct waveform concatenation technique gives better quality speech output compared to HNM, TD-PSOLA, approximate matching, and microphonemic technique, with MOS of 4.5 out of 5.

CONCLUSION

This paper reveals the issue in selecting an appropriate algorithm among the concatenative speech synthesis technique. The various TTS systems developed using concatenative speech synthesis technique are discussed and the results of each method is tabulated and evaluated using Mean Opinion Score (MOS) obtained from the audience. The perceptual test results obtained from each technique is compared with each other. From the perceptual MOS result, it is observed that the direct wave concatenation and Symbol based concatenation technique performs better speech synthesis compared to HNM, TD-PSOLA and microphonemic technique. Hence we can say that, direct waveform concatenation technique is best suitable for TTS synthesis system. Unit selected in building the speech database defines the quality of speech output. This method can be implemented for all Indian language with suitable speech database.

REFERENCES

1. Boersma and Weenink (2001), “PRAAT: A Tool for Phonetic Analysis and Sound Manipulations”, 1992-2001, www.praat.org
2. Huang X, Acero A and Hon H W (2001), “Spoken Language Processing A Guide to Theory, Algorithm and System Development”, Prentice Hall, New Jersy.
3. Hunt A and Black A (1996), “Unit selection in a Concatenative Speech Synthesis System Using Large Speech Database”, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 373-376.
4. Iain R Murray, Mike D Edgington, Diane Campion and Justin Lynn (2000), “Rule-Based Emotion Synthesis Using Concatenated Speech”, Department of Applied Computing, The University, Dundee DD1 4HN.
5. Kishore S P and Black A (2003), “Unit Size in Unit Selection Speech Synthesis”, in Proceedings of Eurospeech, September, pp. 1317-1320.
6. Kishore S P, Rohith Kumar and Rajeev Sangal (2002), “A Data Driven Synthesis Approach for Indian Languages Using Syllable as Basic Unit”, in Proceedings of International Conference on National Language Processing (ICON).
7. Klatt D H (1987), “Review of Text-to-Speech Conversion for English”, *Journal of the Acoustical Society of America (JASA)*, Vol. 82, No. 3, pp. 737-739.
8. Lawrence R Rabiner and Ronald W Schafer (2007), “Digital Processing of Speech Signals”, Pearson Education, Inc. and Dorling Kindersley Publishing Inc.
9. Lemmety S (1999), “Review of Speech Synthesis Technology”, M.S. Thesis, Dept. Elec. and Comm. Engg., Helsinki University of Technology.
10. Marian Macchi (1993), “Issues in Text-to-Speech Synthesis”.
11. Pamela Chaudhari, Madhuri Rao and Vinod Kumar K (2009), “Symbol Based Concatenation Approach for Text to Speech System for Hindi Using Vowel Classification Technique”, *IEEE*.
12. Paul Taylor (2009), “Text-to-Speech Synthesis”, Cambridge University Press.

13. Ravi D J and Sudarshan Patilkulkarni (2011), "A Novel Approach to Develop Speech Database for Kannada Text-to-Speech System", *Int. J. on Recent Trends in Engineering & Technology*, Vol. 05, No. 01.

14. Ravi D J and Sudarshan Patilkulkarni (2012), "Evaluation of Kannada Text-to-Speech [KTTS] System", *International Journal of Advanced Research in Computer Science and Software Engineering*.

15. Ravi D J and Sudarshan Patilkulkarni (2011), "Text-to-Speech Synthesis System for Kannada Language", *International Journal of Advanced Research in Computer Science*, Vol. 2, No. 1.

16. Ravi D J and Sudarshan Patilkulkarni (2009), "Kannada Text-to-Speech Systems: Duration Analysis", Proc. of ISCO, p. 53, Coimbatore.

17. Sangamitra Mohanthy (2011), "Syllable Based Indian Language Text to Speech System", *International Journal of Advances in Engineering & Technology*.

18. Thomas S (2007), "Natural Sounding Text-to-Speech Synthesis Based on Syllable Like Units", M.S. Thesis, Indian Institute of Madras.

19. Veera Raghavendra E, Yegnanarayana B and Kishore Prahallad (2008), "Speech Synthesis Using Approximate Matching of Syllables", Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

20. Yannis Stylianou (2001), "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1.