A Review of Concatenative text To Speech Synthesis

Smita S. Hande[1]

[1]*EXTC Department, Fr. C R I T,Sector 9A Vashi, Navi Mumbai, Maharashtra State, India*

***Abstract*—Speech is used to convey information, emotions, and feelings. Speech synthesis is the technique of converting given input text to synthetic speech. Speechsynthesis 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 andintelligibility 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 Concatenative synthesismethod 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. Concatenative 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 allpossible 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 individualusers.**

***Index Terms*—Speech synthesis, articulatory synthesizer, formant synthesizer, concatenative synthesizer.**

I. INTRODUCTION

A speech synthesizer takes input as a sequence of words (strings) and converts it into speech [1]-[3] that resembles as close as speaker reading that text. A TTS generally contains two modules: Text Analysis/Linguistic Analysis and Digital Signal Processing. The Linguistic Analysis module takes set of strings (words) as an input and gives a normalized phonetic sentence. These phonetic sentences are the input for DSP module [10] which is responsible for generating the corresponding possible natural speech. Speech Synthesizer can be used for various purposes like: i) can be used by visually impaired ii) can be used by vocally impaired iii) in language pedagogy iv) talking books and talking toys etc. Current area of research is speech prosody for all languages, which is essential in both speech synthesis and speech recognition. Synthesized speech should contain prosodic cues for clear perception of words and the construction of meaning of the utterance for listeners.

This paper covers types of synthesis, details of concatenative synthesis, unit selection in concatenation synthesis and some problems associated with concatenation synthesis.

A. Speech Synthesis

This section covers types of synthesis.

There are two main types of speech synthesis methods: 1. Rule Based 2. Corpus Based

In Rule Based method no pre-recorded speech sound is required because each sound is evaluated by specific set of parameters.

There exist two main Rule based techniques: Formant Synthesis and Articulatory Synthesis.

1.1 Formant Synthesis

Formant Synthesis gives a set of rules which describes how to modify pitch, formant frequencies and other parameters from one sound to other [2]-[4]. These rules are based on source-filter model of speech production [1]. Klatt 1980, described formant synthesis model very clearly. By modifying the filter parameters i.e. formants, one can bring prosody in formant synthesis. "The observed irregularities in the spectrum between the formant peaks are of little perceptual importance; only the strong harmonics near a formant peak and below F1 must be synthesized with the correct amplitude in order to mimic an utterance with a high degree of perceptual fidelity".

1.2 Articulatory Synthesis

Speech synthesis is based on mechanical and acoustic models of speech production. It used to model the physiological effects, such as the movement of the lips, tongue, jaw, and the dynamics of the vocal tract and glottis [2]-[4]. This method of synthesis is still in its infantry stage, hence no need to take care of prosody. It is one of the most complex methods because it very difficult [5][6] to model the dynamics involved in the physiological speech production.

1.3 Concatenative Synthesis

Concatenative Synthesis is a type of Corpus based synthesis technique. Commercially, concatenative synthesis [3] is most

popular and commonly used. It got reason for being popular: storage devices became so cheap that storing pre-recorded wave file for processing is no more any costly affair. In this method, utterance is synthesized by concatenating several natural speech segments. Speech database is createdby storing the speech samples in form of sentences, intonational phrases, phonological words, syllables, diphone or phoneme. If all segments are of same length then it is called fixed inventory otherwise unit selection (variable length segments are stored and system makes decision to the best match). Naturalness in concatenative synthesis is then increased by using PSOLA (Pitch Synchronous Overlap Add) algorithm.

B. Concatenative Synthesis Details

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. The most important aspects in concatenative synthesis is to find correct unit length. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems 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. 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 continuos sentence which makes the continuous speech to sound very unnatural (Allen et al. 1987). Because there are hundreds of thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system.

Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points. Another advantage with diphones is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed. For example, in Finnish the combinations, such as /hs/, /sj/, /mt/, /nk/, and /h p/ within a word are not possible. The number of units is usually from 1500 to 2000, which increases the memory requirements and makes the data collection more difficult compared to phonemes. However, the number of data is still tolerable and with other advantages, diphone is a very suitable unit for sample-based text-to-speech synthesis. The number of diphones may be reduced by inverting symmetric transitions, like for example /as/ from /sa/.

Longer segmental units, such as triphones or tetraphones, are quite rarely used. Triphones are like diphones, but contains one phoneme between steady-state points (half phoneme - phoneme - half phoneme). In other words, a triphone is a phoneme with a specific left and right context. For English, more than 10,000 units are required (Huang et al. 1997).

Building the unit inventory consists of three main phases (Hon et al. 1998). First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled or segmented from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is usually very time-consuming. However, some is this work may be done automatically by choosing the input text for analysis phase properly. The implementation of rules to select correct samples for concatenation must also be done very carefully.

C. Unit selection synthesis

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 [11], 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 morenaturalness [12], 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 [13][14].

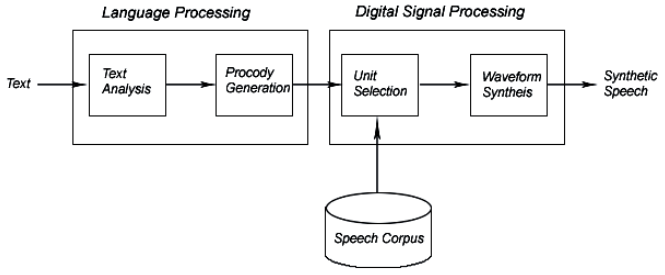

Fig. 1 Unit Selection Synthesis system.

There are several problems in concatenative synthesiscompared to other methods.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.
- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.

Some of the problems may be solved with methods described below and the use of concatenative method is increasing due to better computer capabilities (Donovan 1996).

In order to find the best units in the database, unit selection is based on two costfunctions, target cost and concatenation cost. Concatenation cost refers to how well adjacent units can be joined. The problem of findinga concatenation cost function can be broken into two subproblems; into finding the proper parameterizationsof the signal and into finding the right distance measure. Recent studies attemptedto specify which concatenation distance measures are able to predict audible discontinuities andthus, highly correlates with human perception of discontinuity at concatenation point. However, none of the concatenation costs used so far, can measure the similarity (or, (dis-)continuity) oftwo consecutive units efficiently.

Many features such as line spectral frequencies (LSF) and Mel frequency cepstral coefficients(MFCC) have been used for the detection of discontinuities. In this study, three new sets of features for detecting discontinuities are introduced. The first set of features are obtained bymodeling the speech signal as a sum of harmonics with time varying complex amplitude, whichyield a nonlinear speech model. The second set of features is based on a nonlinear speech analysistechnique which tries to decompose speech signals into AM and FM components. The thirdfeature set exploits the nonlinear nature of the ear.

D. Spectral discontinuity in concatenated speech

When the join between two speech units is clearly audible, it refers to discontinuity. The mismatch in spectra of the speech units on either side of join causes this discontinuity. Audible spectral discontinuities in concatenated signals were researched in. Signal components can change in a number of ways at the join; an abrupt termination of signal components, an abrupt onset of signal components and more subtle changes in signal components sustained across the join. The synthesised speech can sound very natural if the discontinuities at the concatenation points are inaudible. But when these joins are audible, their presence can be very frustrating to the listener and it also reduces the overall perceived quality of synthesized speech.

In systems which use databases containing longer speech units and where the variety of output is limited, the problem of spectral discontinuity is less severe. This is because with longer speech units, there will be lesser concatenation points. However, in systems which create speech by combining large number of smaller speech units, the presence of spectral discontinuity at the concatenation boundaries is a major problem; since there is an increase in the number of joins, therefore, there is an increase in the number of discontinuities. There are a number of reasons for the presence of spectral discontinuities. Audible discontinuity may arise due to inconsistencies in fundamental frequencies, or different levels of loudness (energy of the segments), or due to the contextual differences and variations of speaking style of the speaker.

In order to avoid the problem of spectral discontinuity at concatenation boundaries, an appropriate signal processing technique must be applied. Ideally, a signal processing approach would include algorithms that will examine the synthetic speech waveform at concatenation points and then manipulate the waveform at these points to produce a more natural sounding continuity. In the next section we propose one such signal processing technique to reduce the effect of spectral discontinuities in the original acoustic signal.

II. CONCLUSION

For Text to speech conversion the concatenation speech synthesis is the simplest method where phonemes are concatenated which are called units. The unit plays important role. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. The most important aspects in concatenative synthesis is to find correct unit length. With longer units high naturalness, less concatenation points and good control of coarticulation 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] L. R.Rabiner and R. W.Schafer ,Digital Processing of Speech Signals, Prentice- Hall, Englewood Cliffs, NJ, 1978.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," The Journal of the Acoustical Society of America, vol. 87, pp. 820-857, 1990.

[3] D. H. Klatt, "Review of text-tospeech conversion for English," Journal of the Acoustical Society of America, vol. 82, pp. 737-793, 1987.

[4] Klatt, D. H.: Software for a Cascade/Parallel Formant Synthesiser, The Journal of the Acoustical Society of America, 67(3), Mar. 1980, 971-995, 1980.

[5] Coker, C. H., "A model of articulatory dynamics and control," Proc. IEEE, 64(4),1976, 452-460.

[6] Mermelstein, P., "Articulatory model for the study of speech production," J. Acoust. Soc. Am.,53(4), 1973, 1070-1082.

[7] Sondhi, M. M. and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer, " IEEE Trans. Acoust., Speech, and Signal Processing, 35(7), 1987, 955-967.

[8] A. Black and K. Lenzo, "Limited domain synthesis," in ICSLP2000, Beijing, China., 2000, vol. II, pp. 411–414.

[9] Kain, A. and Macon M.," Spectral voice conversion for text-to-speech synthesis", In: Proc.ICASSP, Seattle,1998.

[10] Atal B. S and Hanauer Suzanne L., "Speech analysis and synthesis by linear prediction of the speech wave", *The journal of acoustic society of America*, 1971, pp 637-655.

[11] 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.
[12] 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.
[13] 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.
[14] MunkhtuyaDavaatsagaan, and Kuldip K. Paliwal, "Diphone-Based Concatenative Speech Synthesis System for Mongolian", IMECS, March, 2008, Hong Kong, pp. 19-21.