Int J Speech Technol (2014) 17:91–98
DOI 10.1007/s10772-013-9210-8

Syllable based text to speech synthesis system using auto associative neural network prosody prediction

Sudhakar Sangeetha · Sekar Jothilakshmi

Received: 9 May 2013 / Accepted: 13 September 2013 / Published online: 24 September 2013

Abstract This paper presents the design and development of an Auto Associative Neural Network (AANN) based unrestricted prosodic information synthesizer. Unrestricted Text To Speech System (TTS) is capable of synthesize different domain speech with improved quality. This paper deals with a corpus-driven text-to speech system based on the concatenative synthesis approach. Concatenative speech synthesis involves the concatenation of the basic units to synthesize an intelligent, natural sounding speech. A corpus-based method (unit selection) uses a large inventory to select the units and concatenate. The prosody prediction is done with the help of five layer auto associative neural network which helps us to improve the quality of speech synthesis. Here syllables are used as basic unit of speech synthesis database. The database consisting of the units along with their annotated information is called annotated speech corpus. A clustering technique is used in annotated speech corpus that provides way to select the appropriate unit for concatenation, based on the lowest total join cost of the speech unit. Discontinuities present at the unit boundaries are lowered by using the mel-LPC smoothing technique. The experiment has been made for the Dravidian language Tamil and the results reveal to demonstrate the improved intelligibility and naturalness of the proposed method. The proposed system is applicable to all the languages if the syllabification rules has been changed.

Keywords Text to Speech Synthesis (TTS) · Tamil TTS · AANN based prosody prediction · Concatenative synthesis approach · Unrestricted Tamil TTS

S. Sangeetha (✉) · S. Jothilakshmi
Annamlalai University, Chidambaram, India
e-mail: sangita.sudhakar@gmail.com

1 Introduction

The essence of text-to-speech synthesis is to convert symbols into signals. Thus, a speech synthesis system occupies a distinctive place in the realm of information technologies. As the signal generation systems, i.e., the speech synthesizers themselves have moved into the domain of sampled speech and stored forms, the main problem of adding naturalness and intelligibility to the systems can largely be solved by incorporating better prosody models. There are several applications based on TTS system. For developing a natural human machine interface, TTS is required for the machine to convey the message. For physically challenged people, TTS systems will be helpful to communicate with others. TTS systems are heavily used in call centers with the replacement of human operators for answering the customer queries. Limited domain TTS systems are already deployed in several commercial systems such as railway and flight schedule queries (Black and Lenzo 2000; Raghavendra and Prahallad 2010).

In state-of-the-art TTS methods, such as unit selection (Black and Cambpbell 1995; Hunt and Black 1996; Donovan and Woodland 1999; Syrdal et al. 2000; Clark et al. 2007) or statistical parametric speech synthesis (Yoshimura et al. 1999, 2000; Zen et al. 2007, 2009; Yamagishi et al. 2009) reasonably high quality synthetic speech can be produced (Karaiskos et al. 2008), especially for normal neutral reading styles. For example, Concatenative unit selection speech synthesis systems have been found to be as intelligible as human speech (Karaiskos et al. 2008). Concatenative speech synthesis (Dutoit 1997) systems combine sound units which are stored in a database, in order to generate the desired speech. The advantage of using unit selection based concatenative synthesis is that there may not be a need for separate prosody modeling, because of the availability of many units under varied contexts. These sound units could be a phoneme, diphone, syllable or word etc. For building Indian language speech synthesis systems, its more appropriate to use syllables as the basic unit. A syllable could be defined as taking the form C^*VC^* where 'C' denotes a consonant and 'V' denotes a vowel. The work (Kishore and Black 2003) suggests the usage of syllables as the basic unit for Indian languages. The earlier efforts—(Rao et al. 2005; Thomas et al. 2006; Venugopalakrishna et al. 2008)—reiterates this fact. Some of the advantages of using syllables as basic units is that they have fairly long duration when compared to phonemes or diphones. Hence, the task of segmentation becomes relatively easier. Also, since the boundaries of most of the syllables are low energy regions (due to consonants), the concatenations would result in reduced perceivable distortions.

Neural networks are known for their ability to generalize according to the similarity of their inputs but also to distinguish different outputs from input patterns that are similar only on the surface. As a consequence, network have the power to predict, after an appropriate learning phase, even patterns they have never seen before. This provides the researcher with a potential solution to the problem of constructing models from imperfect data. In this study auto associative neural network was used to accomplish the prediction of pitch and intonation models. This is done for the same reason that most researchers use decision trees (see for instance Clark and Dusterho 1999; Hirschberg 1993; Black and Cambpbell 1995). That is, neural networks should in principle enjoy the same advantages the decision tree methodology does; they can be automatically trained to learn different speaking styles and they can accommodate a range of inputs from simple text analysis for the problem of synthesis from unrestricted text to more detailed discourse information (Clark and Dusterho 1999) that may be available as a by-product of text generation.

In this paper, we propose a corpus driven text-to-speech system for Tamil language. Concatenative-syllable based synthesis approach is used to produce the desired speech through pre-recorded speech waveforms. The given complex agglunative Tamil input text is morphologically analyzed and based on the linguistics rules, syllabication is performed. The syllables are called basic speech units. The repository of these units is created with its prosodic information. Prosody model (duration and intonation) based on the auto associative neural network provide the duration and intonation information associated with the sequence of syllable units present in the given input text. During the synthesis, appropriate syllable units are selected and concatenated according to the sequence present in the input text and then derived intonation and duration knowledge for the sequence of concatenated syllables is incorporates using pitch modification method. After performing concatenation, the waveform is smoothened at concatenation joints using mel-LPC. This proposed text to speech synthesis system can be used to all languages if the syllabification rules has been changed according to that languages. Since the syllabification rules will vary form one language to other language.

The rest of this paper is organized as follows. In Sect. 2, discuss about the concatenative speech synthesis approach and linguistics rules for syllabification of Tamil language text. In Sect. 3 provides the details of the speech corpus used for developing the unrestricted TTS. In Sect. 4, describe the prosody modeling using auto associative neural network. Section 5 describe the proposed text to speech synthesis system. In Sect. 6, discuss about the quality test. Finally, Sect. 7 provides the conclusion of this paper.

2 Concatenative speech synthesis

Over the past decades, the concatenative synthesis approach was very difficult to implement because of limitation of computer memory. With the advancements in computer hardware and memory, a large amount of speech corpus can be stored and used to produce high quality speech waveforms for a given text. Thus, the synthesized speech preserves the naturalness and intelligibility. Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis yields the supreme natural-sounding synthesized speech. Concatenative speech synthesis uses phones, diphones, syllables, words and sentences as basic units. Speech is synthesized based on selecting these units from the database, called as a speech corpus.

Many researches have been made, selecting each separate unit as the basic unit. When phones are selected as basic units, the size of the database will be less than 50 units for Indian languages. The database may be small, but phones provide very less co-articulation information across adjacent units, thus falling to model the dynamics of speech sounds. Diphones and triphones as basic units, it will minimize the discontinuities at the concatenation points and captures the co-articulation effects. But a single example of each diphone is not enough to produce good quality speech. So we selected syllable as a basic unit. Unit selection provides the greatest naturalness, because it applies only small amounts of digital signal processing (DSP) to the recorded speech (Kominek and Black 2003). DSP often makes recorded speech sound less natural, although some systems use a small amount of signal processing at the point of concatenation to smooth the waveform. The output from the best unit-selection systems is often indistinguishable from real human voices, especially in contexts for which the TTS system has been tuned. However, maximum naturalness typically requires unit-selection speech databases to be very large.

2.1 Linguistics rules for syllabification of Tamil language

Indian languages are syllable centered, where pronunciations are based on syllables. A syllable can be the best unit for Indian languages intelligible speech synthesis. The general form of the 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. This proposed system is tested using the Tamil language. There are 18 consonants and 12 vowels in Tamil language. There are defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. Some of the rules used to perform grapheme to syllable conversion (Saraswathi and Geetha 2010) is:

- 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 of C.
- If the character 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 a CV pattern is followed by CCCV then the syllables are split as CVCC and CV.
- If the VC pattern is followed the 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.

And some of the new rules that have been added in this work to perform grapheme to syllable conversion is:

- If the character after CV pattern are of type CV then the syllables are split as CVCV.
- Similarly If the character after CV pattern are of type CVCV then the syllables are split as CVCVCV.
- If the CV pattern if followed by CVC then syllables are split as CVCVC.
- If the CV pattern if followed by CCV then syllables are split as CVCCV.

Researches were conducted to find which order of the syllables is best accepted for synthesis. The following are the recommended combinations:

- Monosyllables at the beginning of a word and bisyllables at the end.
- Bisyllables at the beginning of a word and monosyllables at the end.
- Monosyllables at the beginning and trisyllables at the end of a word.
- Trisyllables at the beginning and monosyllables at the end of a word.

3 Speech corpus

Building a speech corpus is a difficult task for the text to speech synthesis system. Text to speech synthesis system based on the concatenative syllable based approach needs labeled speech corpora. The quality of the synthesized speech depends on the number of variants of various units present in the database. Prosodic information such as pitch, duration and intonation prediction has to be done in the corpus development stage itself, some more information has to be specified with the basic speech units after storing them in the corpus. The problem such as mispronunciation, untranscribed speech units, phrase boundary detection, pronunciation variants are to be identified and addressed.

For developing the corpus for TTS system, we have considered 500 sentences and they are split into syllables based on linguistic rules and the syllables are recorded using two female and three male persons, whose age is in the range of 25–35 years. The reason for recording the speech syllables with multiple speakers is to select the preeminent speaker for recording the final speech to develop the unrestricted and full-fledged TTS. In this TTS corpus creation, best speaker means the speech produced by that speaker should have uniform characteristics with respect to speaking rate, intonation, energy profile and pronunciation. At concatenative TTS, the quality of speech corpus is very important, because the characteristics of synthesized speech are directly related to nature of speech corpus.

Along with the above mentioned characteristics, the speech of the selected speaker should give least distortion when speech segments are manipulated as per requirements. First, the syllables are recorded by each of the five speakers in an echo free audio recording studio. The recorded speech signal is sampled at 48 kHz and stored in 16-bit PCM-data format. After recording the speech from each speaker each of the speech file is down sampled to 16 kHz. The speech wave files are saved according to the requirement. The speech wave files corresponding to the words are named according to their corresponding Romanized names. Each syllable and word file contains text transcriptions and timing information in a number of samples. The fundamental frequencies of the syllables are computed using the autocorrelation method. The words collected comprises dictionary words, commonly used words, newspapers and story books, also different domain such as sports, news, literature and education for building unrestricted TTS (Clark and Dusterho 1999).

4 Prosody modeling with auto associative neural networks

Prosodic parameters of speech at the syllable level depend on positional, contextual or background and phonological

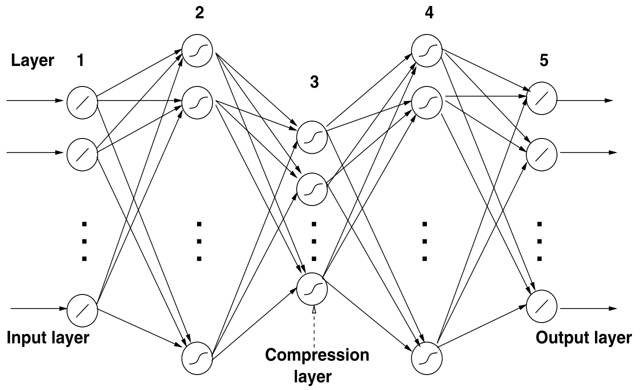

Fig. 1 A five layer AANN model

features of the syllables (Yegnanarayana 1999). In this paper, auto associative neural networks are employed to model the prosodic parameters of the syllables from their features. The prosodic parameters considered in this work are duration and sequence of pitch (F0) values of the syllables. This technique is used in text to speech synthesis process. Neural network models in voice conversion system are explored or employed for capturing the mapping functions between source and target speakers at source, system and prosodic levels, also it is used to characterize the emotions present in speech. Auto-associative neural networks are a particular class of neural networks in which the target output pattern is identical to the input pattern. The aim is to approximate as closely as possible the input data themselves. The input and output feature vectors of the given training data are expected to capture the functional relationship of AANN model. A five-layer auto associative neural network shown in the Fig. 1 is used for modeling the duration/F0 patterns of syllables. The structure of the AANN is shown in Fig. 1.

- The first and fifth layers are called the input and output layers respectively. Data dimensionality is nothing but the number of neurons that they contain.
- The mapping and demapping layers are referred to as the second and fourth layers respectively. Both are not essential to have the same number of neurons. To reduce the number of free parameters in the model architecture, they are fixed to be the same. The transfer (sigmoid) function in the mapping and demapping layers is defined as follows:

  $$\sigma(x) = \frac{2}{1 - e^{-2x}} - 1 \tag{1}$$

  This non-linear sigmoid function provides the capability for modelling arbitrary functions f and g.
- The third layer is called as the bottleneck layer. It should be a smaller dimension than either the input or output layer. The outputs of the bottleneck layer are referred as non-linear principal components which could be viewed as a non-linear generalization of principal components. The bottleneck layer may be linear or nonlinear activation functions. Commonly speaking, linear transfer functions are preferred unless a bounded response is preferred.

In this paper, we consider the mapping function is between the 25-dimensional input vector and the one-dimensional output. Several network structures are explored in this study. The final structure of the network is 19L 38N 5N 38N 19L, where L denotes a linear unit, and N denotes a non-linear unit. The integer value indicates the number of units used in that layer. For modeling the durations of syllables, neural network models are urbanized with respect to the specified language. In case of intonation modeling, a separate model is developed for each of the speakers in that language. For each syllable, a 25-dimension input vector is formed, representing the positional, contextual and phonological features. The input vector features and the number of input nodes to the AANN are given in Table 1.

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100, \tag{2}$$

Where x_i and y_i are the actual and predicted F0/duration values, respectively. To evaluate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation coefficient ($\gamma_{X,Y}$) are computed using actual and predicted F0/duration values. The definitions of average prediction error (μ), standard deviation (σ) and the linear correlation coefficient ($\gamma_{X,Y}$) are given below.

$$\mu = \frac{\sum_i |x_i - y_i|}{N}, \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \tag{4}$$

$$d_i = e_i - \mu, \quad e_i = x_i - y_i \tag{5}$$

where x_i, y_i are the actual and predicted F0 values, respectively, and e_i is the error between the actual and predicted F0 values. The deviation in error is d_i, and N is the number of observed F0 values of the syllables. The correlation coefficient is given by

$$\gamma_{x,y} = \frac{V_{X,Y}}{\sigma_X \cdot \sigma_Y}, \tag{6}$$

where

$$V_{X,Y} = \frac{\sum_i |(x_i - \overline{x)}| \cdot |(y_{i-}\overline{y})|}{N} \tag{7}$$

The quantities σ_X, σ_Y are the standard deviations for the actual and predicted F0 values, respectively, and $V_{X,Y}$ is the correlation between the actual and predicted F0 values.

To estimate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation

Table 1 List of input feature vectors and a number of input nodes to AANN

Factors	Features	No. of nodes
Syllable position in the phrase	Position of syllable from beginning of the phrase Position of syllable from end of the phrase Number of syllables in the phrase	3
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word	3
Word position in the phrase	Position of word from beginning of the phrase Position of word from end of the phrase Number of words in the phrase	3
Syllable identity (represented by four dimensional feature vector)	Segments of the syllable (consonants and vowels)	4
Context of the syllable (represented by four dimensional feature vector)	Previous syllable Following syllable	4 4
Syllable nucleus	Position of the nucleus Number of segments before the nucleus Number of segments after the nucleus	3
Pitch	F0 of the previous syllable	1
Gender	Gender of the speaker	1

Table 2 Performance of the prosody models for predicting the F0 of the syllables, $N = 1276$ syllables

Predicted syllables within deviation				Objective measures		
5 %	10 %	15 %	25 %	μ (ms)	σ (ms)	γ
41	74	92	99	16.02	13.04	0.80

Table 3 Performance of the prosody models for predicting the durations of the syllables

Predicted syllables within deviation			Objective measures		
10 %	25 %	50 %	μ (ms)	σ (ms)	γ
36	78	98	27	24	0.83

coefficient ($\gamma_{X,Y}$) are computed using actual and predicted F0/duration values. These results are given in Tables 2 and 3.

In this TTS synthesis, prosody models provide the specific duration and modulation information associated with the sequence of sound units present in the given text. In this study, employ waveform concatenation for synthesizing the speech. The basis of concatenative synthesis is to join short segments of speech, usually taken from a pre-recorded database, and then impose the associated prosody by appropriate signal processing methods.

5 Proposed text to speech synthesis system

5.1 Front end

The block diagram of the proposed unrestricted AANN based TTS is shown in Fig. 2. In the front end, the first step is text analysis which is an automated process. In this stage, removing of punctuations such as double quotes, full stop, comma and all are performed here. After the preprocessing process a pure sentence will be obtained. All the abbreviations present in the input text are expanded and also unwanted punctuation are removed, and then normalizing non-standard words likes abbreviations and numbers are performed. Then the text is syllabified based on the linguistic rules as discussed in Sect. 2.1. Then the positional, contextual and phonological features (linguistics and production constraints) for each of the syllable present in the given text are derived. These features are given to the prosody prediction models (duration and intonation model) which will generate the appropriate duration and intonation information corresponding to the syllables as discussed in Sect. 4.

5.2 Synthesis stage

At the synthesis stage, first, the concatenation is performed based on the pre-recorded syllables according to the sequence in the text. Using prosody modification methods the derived duration and intonation knowledge corresponding to the sequence of syllables is incorporated into the sequence of concatenated syllables. The prosody parameters (duration and pitch) are incorporated by manipulating the instants of significant excitation of the vocal tract system during the production of speech (Rao and Yegnanarayana 2003). Instants of significant excitation are computed from the linear prediction (LP) residual of the speech signals by using the average group-delay of minimum phase signals. The main problem with concatenation process is that there will be glitches in the joint. These discontinuities present at the

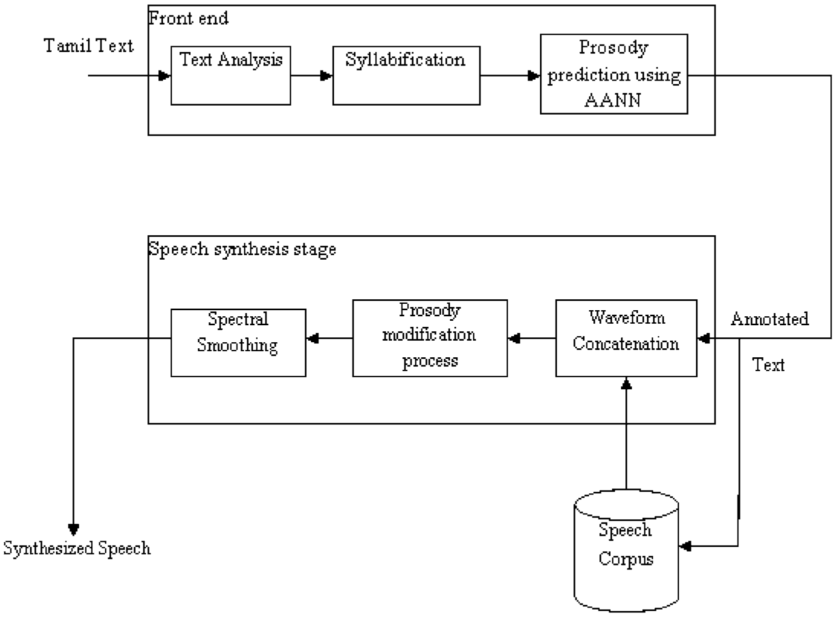

Fig. 2 Block diagram of proposed Tamil text to speech synthesis system

unit boundaries are lowered by using the mel-LPC smoothing technique.

5.2.1 Spectral smoothing

To introduce individual smoothness for syllable in Tamil text-to-speech, a time scale modification is carried out for each syllable. Smoothing at concatenation joints is performed using mel-LPC. In general, mel-LPC is used for representing the spectral envelope of a digital signal of speech using the information of a linear predictive model. The basic idea behind the mel-LPC analysis is that a speech sample can be estimated as mel cepstrum by applying mel scale filter bank on linear combination of speech samples. By decreasing the sum of the squared differences between the actual speech samples and the mel-linearly predicted ones, a unique set of predictor coefficients is determined. Speech is modeled as the output of linear, time varying system excited by periodic pulses.

Mel-linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the time-varying system representing vocal tract (Lokesh and Balakrishnan 2012). The difference equation describing the relation between speech samples $s(n)$ and excitation $u(n)$ for proposed method is

$$S_{mel}(n) = \sum_{k=1}^{P} a_k s_{mel}(n-k) + G_{mel} U(n) \tag{8}$$

The system function is of the form

$$h(z) = \frac{s_{mel}\ (Z)}{u_{mel}\ (z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{9}$$

A mel-LPC of order p with prediction coefficients axis defined as a system whose output is

$$S_{mel}(n) = 1 - \sum_{k=1}^{P} a_k z^{-k} \tag{10}$$

where, S is the z transform of the error signal and has the $h(z)$ characteristics of a noise, since a Mel filter separates the uncorrelated. In this case it is enough to calculate the auto-correlation function of the signals (n) which belongs to a speech signal from the database. It is one of the most powerful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters (Varho and Alku 1997).

6 Quality test

Voice quality testing is performed using subjective test. In subjective tests, human listeners hear and rank the quality of processed voice files according to a certain scale. The most common scale is called a Mean Opinion Score (MOS) and is composed of 5 scores of subjective quality, 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent. The MOS score of a certain TTS system is the average of all the ranks voted by different listeners of the different voice file used in the experiment. The tests were conducted in a laboratory environment with 50 students in the age group of 20–28 years by playing the synthesized Tamil speech signals through headphones. In this case, the subjects should possess the adequate speech knowledge for accurate assessment of the speech signals and were examined to evaluate the articulacy and spontaneity of the synthesized speech. They have to assess the quality on

Table 4 Mean opinion score for the quality of synthesized speech in Tamil language

Mean opinion score				Level of confidence (%)	
TTS with prosody		TTS without prosody			
Intelligence	Naturalness	Intelligence	Naturalness	Intelligence	Naturalness
4.5	4.2	3.9	3.1	>99.5	>99.5

a 5-point scale for each of the sentences. The mean opinion scores for assessing the intelligibility and naturalness of the synthesized Tamil speech is given in Table 4.

The MOS scores show that the intelligibility of the synthesized Tamil speech is honestly acceptable, whereas the naturalness appears to be little degree of degradation. Naturalness is mostly attributed to distinct perception. It can be enhanced to some degree by integrating the stress and co articulation information along with duration and pitch. The accuracy of the prediction of prosody models can be also analyzed by conducting the listening tests for judging the intelligibility and naturalness on the synthesized speech without incorporating the prosody. In this case, speech samples are the derivative of concatenating the neutral syllables without integrating the prosody. The MOS of the excellence of the synthesized speech without incorporating the prosody have been observed to be low compared to the speech synthesized by combining the prosody. The consequence of the differences between the pairs of the MOS for intelligibility and naturalness is verified using hypothesis testing and the level of confidence is high (> 99.5 %) for both cases.

7 Results and conclusion

In this proposed work, a prototype text to speech synthesis system for Tamil using syllable as the basic unit was developed using AANN prosody prediction each of the five speakers. Five speaker's speech samples and synthesized speech from prototype TTS systems are analyzed. The best speaker who has uniform characteristics of pitch, energy dynamics and speaking rate is selected. Speech corpus has been created from the best speaker is used to build unrestricted TTS. Text corpus has been created from various domains. An optimal unit selection algorithm is used to reduce redundancy in the text corpus. Linguistic rules are derived from the text to syllable conversion in Tamil. Clustering of units is done based on syllable specific positional, contextual and phonological features. Prosody prediction of the syllables has been done based on the auto associative neural network models to enhance the intelligibility and naturalness. Spectral discontinuities were lowered at unit boundaries based on the mel-LPC method. Thus unrestricted TTS have been developed. Based on the subjective quality test results we can conclude that the proposed TTS system producing the synthesized speech with naturalness and good quality. This proposed TTS system can be generalized to all the languages if we change the syllabication rules according to that language's.

References

Black, A. W., & Cambpbell, N. (1995). Optimising selection of units from speech database for concatenative synthesis. In *Proc. EUROSPEECH* (pp. 581–584).

Black, A. W., & Lenzo, K. A. (2000). Limited domain synthesis. In *Proc. ICSLP*, Beijing, China.

Clark, R. A., Richmond, K., & King, S. (2007). Multisyn: open-domain unit selection for the festival speech synthesis system. *Speech Communication*, *49*(4), 317–330.

Donovan, R., & Woodland, P. (1999). A hidden markov-model-based trainable speech synthesizer. *Computer Speech & Language*, *13*(3), 223–241.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Norwell: Kluwer Academic.

Hirschberg, J. (1993). Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence*, *63*(13), 305–340.

Hunt, A., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP* (pp. 373–376).

Karaiskos, V., King, S., Clark, R. A. J., & Mayo, C. (2008). The Blizzard Challenge 2008. In *Proc. Blizzard Challenge workshop*, Risbane, Australia.

Kishore, S. P., & Black, A. (2003). Unit size in unit selection speech synthesis. In *Proc. of EUROSPEECH* (pp. 1317–1320).

Kominek, J., & Black, A. (2003). CMU ARCTIC databases for speech synthesis. Language Technologies Institute.

Raghavendra, E., & Prahallad, K. (2010). A multilingual screen reader in Indian languages. In *National conference on communications (NCC)*, Chennai, India.

Rao, K. S., & Yegnanarayana, B. (2003). Prosodic manipulation using instants of significant excitation. In *Proc. IEEE int. conf. multimedia and expo*, Baltimore Maryland, USA (pp. 389–392).

Rao, M. N., Thomas, S., Nagarajan, T., & Murthy, H. A. (2005). Text-to-speech synthesis using syllable like units. In *Proc. of national conference on communication (NCC)*, IIT Kharagpur, India (pp. 227–280).

Clark, R. A. J., & Dusterho, K. E. (1999). Objective methods for evaluating synthetic intonation. In *Proc. Eurospeech*, Budapest, Hungary.

Lokesh, S., & Balakrishnan, G. (2012). Speech enhancement using mel-LPC cepstrum and vector quantization for ASR. *European Journal of Scientific Research*, *73*(2), 202–209.

Saraswathi, S., & Geetha, T. V. (2010). Design of language models at various phases of Tamil speech recognition system. *International Journal of Engineering Science and Technology*, *2*(5), 244–257.

Syrdal, A., Wightman, C., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Storm, V., Lee, K., & Makashay, M. (2000). Corpus-based techniques in the at and t nextgen synthesis system. In *Proc. ICSLP* (pp. 411–416).

Thomas, M. S., Rao, N., Murthy, H. A., & Ramalingam, C. S. (2006). Natural sounding tts based on syllable-like units. In *Proc. of 14th European signal processing conference*, Florence, Italy.

Varho, S., & Alku, P. (1997). Linear predictive method using extrapolated samples for modeling of voiced speech. In *Proc. of IEEE workshop on applications of signal processing to audio and acoustics* (pp. 13–16).

Venugopalakrishna, Y. R., Vinodh, M. V., Murthy, H. A., & Ramalingam, C. S. (2008). Methods for improving the quality of syllable based speech synthesis. In *Proc. of spoken language technology (SLT) workshop*, Goa (pp. 29–32).

Yamagishi, J., Nose, T., Zen, H., Ling, Z. H., Toda, T., Tokuda, K., King, S., & Renals, S. (2009). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1208–1230.

Yegnanarayana, B. (1999). *Artificial neural networks*. New Delhi: Prentice-Hall.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum pitch and duration in hmm-based speech synthesis. In *Proc. EUROSPEECH* (pp. 2350–2374).

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. S. (2000). Simultaneous modeling of spectrum pitch and duration in hmm-based speech synthesis. *IEICE Transactions*, *83-D-II*(11), 2099–2107.

Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of nitech hmm based speech synthesis system for the blizzard challenge 2005. *IEICE Transactions on Information and Systems*, *90-D*(1), 325–333.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064.