Efficient Methodology for Segmentation of Speech Signals in Text to Speech

N.S.Raut
P.G Department of Computer science & Engg SGBAU, Amravati,India

V.M. Thakare, PhD
HOD Of Computer Science & Engg Dept,SGBAU, Amravati,India

S.S. Sherekar, PhD
Associate Professor in Comp Sci & Engg Dept,SGBAU, Amravati.India

ABSTRACT

This paper proposes a method for tuning the weights of unit selection cost functions in syllable based text-to-speech (TTS) synthesis system, two-stage feedforward neural network (FFNN) based approach for modeling fundamental frequency (*F*0) values of a sequence of syllables. Unrestricted Text To Speech System (TTS) is capable of synthesize different domain speech with improved quality. 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. Unit selection cost functions, namely target cost and concatenation cost, are designed appropriate to syllables. The method tunes the weights in such a way that perceptual preference patterns are appropriately considered while selecting the units. The method uses genetic algorithm to derive the optimal weights. From the evaluation, it is observed that prediction accuracy is better for two stage FFNN models, compared to the other different models.

Keywords: Text to Speech Synthesis (TTS), Concatenative synthesis approach, Intonation models, Feed forward neural networks, Unit selection, Target cost, Tuning of weights.

1. INTRODUCTION

A text-to-speech (TTS) synthesis system converts a given input text to corresponding output speech. Unit selection based concatenative approach is one of the corpus based synthesis techniques which can synthesize speech close to natural quality . initially unit selection cost functions namely target cost and concatenation cost are proposed as appropriate to syllables. Then optimal weight tuning method is proposed to tune the weights associated with the subcost of cost functions based on perceptual preference patterns. Tuning of weights is formulated as an optimization problem. Genetic algorithm is used to find the optimal set of weights. Fitness function is designed in such a way that the perceptual preference patterns are mapped into the weights of cost function. Prosody plays an important role in an improving the quality of text-to-speech synthesis (TTS) system both in terms of naturalness and intelligibility. [8]Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. In speech synthesis, intonation directly affects the overall quality of the synthetic speech. From the speaker's view point, intonation can be used to convey pragmatic and emotional information. Linguistic and production constraints are used to predict the *F*0 values of the sequence of syllables. 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. Concatenative-syllable based synthesis approach is used to produce the desired speech through pre-recorded speech waveforms. The Prosody model (duration and intonation) based on the auto associative neural network provides 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 also 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.

2. BACKGROUND

The design and development of an Auto Associative Neural Network (AANN) based unrestricted prosodic information synthesizer[3]. Unrestricted Text To Speech System (TTS) is capable of synthesize different domain speech with improved quality. Auto associative neural networks are employed to model the prosodic parameters of the syllables from their features .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. The implicit knowledge of intonation is usually captured by using modeling techniques. Festival framework is used for developing TTS system [2]. Festival offers general tools for building unit selection synthesizer[4]. Festival provides different rule-based and trained intonation modules for predicting the target *F*0 values. In developing baseline TTS system, trained models like Classification and Regression Trees (CART) are used for predicting the *F*0 values of syllables. Several methods are developed for tuning the weights of subcosts by including subjective criterion into the unit selection cost functions. Tuning of weights is formulated as an optimization problem. The Genetic algorithm is used to find the optimal set of weights. Fitness function is designed in such a way that perceptual preference patterns are mapped into the weights of cost function. Target cost should prompt unit selection process to select appropriate units from the database having the required target features[1].In [1], active interactive genetic algorithms (aiGAs) are used for subjective weight adjustments. Application of aiGAs is to weight adjustment process reduces user fatigue and frustration, and improves user consistency in subjective evaluations.

3. RELATED WORK DONE

Concatenative unit selection speech synthesis systems have been found to be as intelligible as human speech[5].[3]The main problem with concatenation process is that there will be glitches in the joint. These discontinuities present at the unit boundaries are lowered by using the mel-LPC(linear prediction)smoothing technique[10].The baseline TTS system [2] is developed using recorded speech corpus .Festival framework is used for developing TTS system model, first the tilt parameters are derived from the linguistic and production constraints using FFNN.Later, the derived tilt parameters are used for predicting the intonation contours. Target cost should prompt unit selection process to select appropriate units from the database having the required target features. [1]Target cost is formulated in three stages.first two stages are carried out in training phase & third stage is implemented at time of synthesis. Weight space search (WSS) is one such method of tuning the appropriate units from the database having the required target features. [1]Target cost is formulated in three stages.first two stages arecarried out in training phase & third stage is implemented at time of synthesis. Weight space search (WSS) is one such method of tuning the weights through exhaustive search of finite set of possible weights through analysis-by-synthesis exploration.The main limitation of this method is that computational requirement grows exponentially with the number of weights beingtuned. Here work is done on method based on genetic algorithm that can relate weights of subcosts with human perception [9]The genetic algorithm evaluates the fitness function in order to find the inputs which can produce minimum value of the fitness function.

4. TEXT TO SPEECH CONVERSION TECHNIQUES

Two major methodology have emerged for modeling intonation: (i) the tone sequence approach which follows the traditional phonological description of intonation and (ii) thesuperposition approach. Phonological (tone sequence) models interpret F0 contour as a linear sequence of phonologically distinctive units (tones or pitch accents), which are local in nature.Acoustic-phonetic (superposition or overlay) models interpret F0 contour as a result of superposition of several components of different temporal scopes. A classical superpositional intonational model for Japanese has been proposed. In the Indian context, a rule-based intonation model was proposed for Hindi TTS system. The baseline TTS system [2] is developed using recorded speech corpus.[6]Earlier works in Indian languages have also suggested that the choice of syllables as the basic units for synthesis lead to good quality speech.Optimal weights obtained from the proposed weight tuning system are used in unit selection process. To observe the effectiveness of optimal weights, improvement in the quality of synthesis is evaluated by comparing with the manually tuned weights obtained from informal listening tests[1]. 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[3]. The intonation model is evaluated with the syllables in the test set. The three average *F*0 values located at start, middle and end positions of each syllable in the test set are predicted using FFNN by presenting the feature vector of each syllable as input to the network[2].

In this paper Proposed Methodology using two-stage feedforward neural networks.

4.1) Principle of two-stage intonation Model

Two-stage intonation model consists of two FFNNs. In the first stage, one of the FFNNs is used to model the tilt parameters from PCPA features (Fig).[2]In the second stage, the other FFNN is used to model the *F*0 values of the syllables from the combination of PCPA and the tilt parameters (Fig).

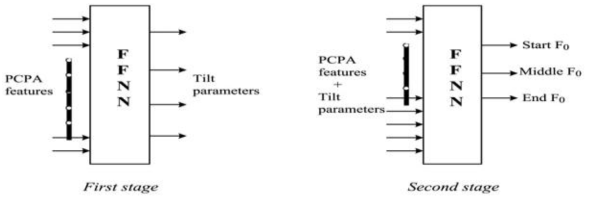

Fig.1

4.2) Evaluation of two-stage Intonation model:

Performance of the two-stage intonation model is evaluated by using objective and subjective measures. The prediction performance of the two-stage intonation model using FFNNs is compared with that of the two-stage LR and CART models. Prediction performance of the intonation models developed using PCPA and PCPA+tilt features. shownin(fig).

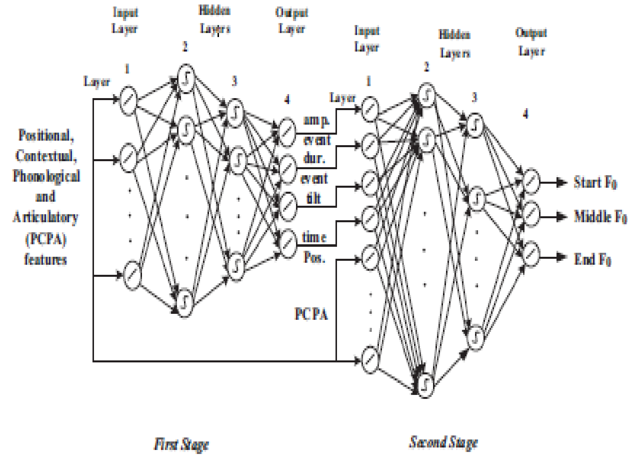

Fig. Architecture of two-stage intonation model using FFNNs.

Fig.2

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 durationand intonation knowledge corresponding to the sequence of syllables is incorporated into the sequence of concatenated syllables[7].In this paper also focused on **Genetic algorithm, Fitness function.**[1]Genetic algorithm is population based evolutionary algorithm,which use derivative free optimization strategy and always yields global optimum values .The algorithm uses the principles derived from natural genetics. GA starts its search from a randomly generated population.Genetic algorithm is based on three basic operators namely selection or reproduction, crossover and Mutation. Fitness function is designed in such a way that perceptual preference patterns are mapped into the weights of cost function.

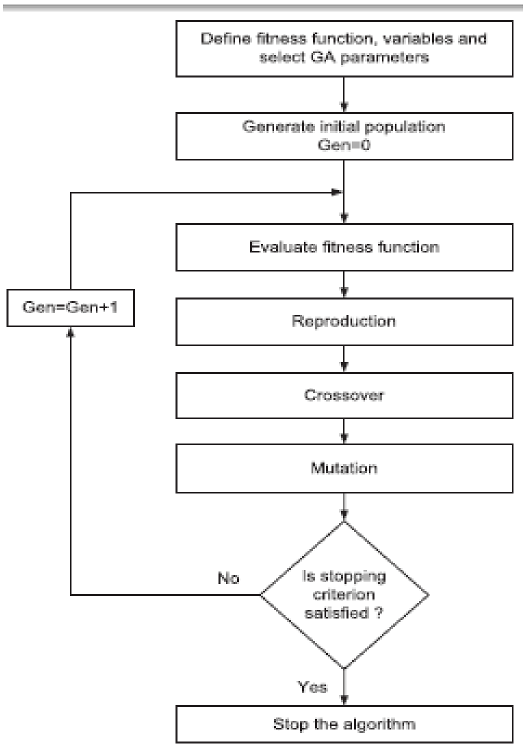
Define fitness function, variables and select GA parameters
Generate initial population Gen=0
Evaluate fitness function
Gen=Gen+1
Reproduction
Crossover
Mutation
No
Is stopping criterion satisfied ?
Yes
Stop the algorithm

Fig.3

4. POSSIBLE OUTCOME AND RESULT

Unit selection cost functions, namely concatenation cost and target cost, are proposed for syllable based synthesis. Concatenation costs are defined based on the type of segments present at the syllable joins. The target cost formulation is proposed in three stages for enhancing the quality of the TTS system. The predicted tilt parameters are concatenated with the PCPA features, and they are given as input to TTS system at perceptual level. 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. The output of fitness function is the average dissimilarity. The second stage of the network to predict the $F0$ values at the output of second stage. The contribution of tilt parameters to (positional, contextual, phonological and articulatory) PCPA features is significant in improving the quality ofmeasure between the reference rankings and test rankings of all test words for a given set of input weights.Thederived optimal weights can synthesize good quality speech compared to manually tuned weights.

5. CONCLUSION

Genetic algorithm is used to adjust the weights of subcosts such that the ranking obtained from the total cost of all instances and the ranking obtained from perceptual preference tests are nearly same. Fitness function is designed such that weights of the cost function can select the units which are perceptual preferred by listeners. Spectral discontinuities were lowered at unit boundaries based on the mel-LPC method.

Thus unrestricted TTShave been developed TTS system producing the synthesized speech with naturalness and goodquality, to improve the performance in all possiable direction.

6. FUTURE WORK

All the methods have been refined for further improvement in TTS system can be generalized to all the languages & also to improve the quality of speech.

7. REFERENCES

[1] N.P. Narendra, K. Sreenivasa Rao,” Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis”, Science Direct, VOL. 13, NO. 2, PP.773-781.Feb-2013.

[2] V. Ramu Reddy, K. Sreenivasa Rao,” Two-stage intonation modeling using feedforward neural networksfor syllable based text-to-speech synthesis”, Science Direct, VOL. 27, NO. 5, PP.1105-1126, Aug-2013.

[3] Sudhakar Sangeetha , Sekar Jothilakshmi,” Syllable based text to speech synthesis system using auto associative neural network prosody prediction”, Springer Science, Vol.17,No.2,PP.91-98, June 2014.

[4] Black, A.W., Taylor, P., Caley, R.,.”The Festival speech synthesis system”, Manual and source code available at www.cstr.ed.ac.uk/ projects/festival.html, 2009.

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

[6] Rao, K.S., Yegnanarayana, B.,” Intonation modeling for Indian languages” Computer Speech and Language 23 (April), 240–256.2009.

[7] Rao, K. S., & Yegnanarayana, B..”Prosodic manipulation using instants of significant excitation” In Proc. IEEE int. conf. multimediaand expo, Baltimore Maryland, USA,pp. 389–392 ,2003.

[8] Rao, K.S.,” Acquisition and incorporation prosody knowledge for speech systems in Indian languages”. PhD Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May 2005.

[9] D.E. Goldberg, “The Design of Innovation”, Lessons From and For CompetentGenetic Algorithms, Kluwer Academic Publisher, Dordrecht, 2002.

[10] Lokesh, S., & Balakrishnan, G.Speech enhancement using mel-LPC cepstrum and vector quantization for ASR. *EuropeanJournal of Scientific Research*, *73*(2), 202–209.2012.