Duration Modeling Techniques for Indian Language TTS System: A Review

Chandrika M.
M Tech student, Dept. of Electronics and Communication Engineering, SJCE, Mysore, Karnataka, India.

Shreekanth T.
Assistant Professor, Dept. of Electronics and Communication Engineering, SJCE, Mysore, Karnataka, India.

Abstract- Duration is one of the prosodic features of speech the other two being stress and intonation. The primary goal in duration modeling is to model the duration pattern of natural speech, considering various features that affect the pattern. Accurate estimation of segmental durations is crucial for natural sounding text-to-speech (TTS) synthesis. Variation in segmental duration serves as a cue to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases thereby increasing the naturalness and intelligibility. This review traces the earlier works carried out by the researchers on the duration modeling with a number of features considered and their usefulness and relative contribution for segmental duration prediction.

Key words: syllable, CART, SOP.

I. INTRODUCTION

The text-to-speech (TTS) system will convert ordinary orthographic text into acoustic signal which is indistinguishable from human speech. Text processing and speech generation are two main components of a text to speech system. The objective of text processing component is to produce appropriate sequence of phonemic units [1]. 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. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text[2]. The majority of issues are associated in building a TTS for a new language is associated with handling of real-world text[3]. Current state-of-art TTS system in english and other well-researched languages use such rich set of linguistic resources such as word- sense disambiguation, morphological analyser, part-of-speech tagging, letter-to-sound rules, syllabification, stress-pratterns in one form or the other to build a text processing component of TTS system.

In recent years, researches have come up TTS systems that provides high quality synthetic speech with no compromising in terms of naturalness and intelligibility[4]. Variation in segmental duration serves as a cue to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases thereby increasing the naturalness and intelligibility. In natural speech, segmental durations are highly context dependent. For example, instances of vowel /e/ that are as short as 35 ms in the word "*pehlA*" and as long as 150 ms in the word "*rahegA*"[6]. So the primary intension in duration modeling is to model the duration pattern of natural speech, considering various features that affect the pattern. An important restriction being that, due to the nature of the Text-to-Speech synthesis problem, only those features that can be automatically derived from text can be considered. The approaches to segmental duration modeling can be divided into two categories: rule-based and corpus based.

In the rule based approach, linguistic experts derive a complicated set of rules to model prosodic variations by observing natural speech. In the corpus based approach, speech corpus specially designed and annotated with various levels of prosodic information is used[4]. The corpus is analyzed automatically to create prosodic models which are then evaluated on test data. Modeling the syllable durations by analyzing large databases manually is a tedious process.. Duration models help to improve the quality of Text-to-Speech (TTS) systems. In most of the TTS systems the durations of syllables are estimated using a set of rules derived manually from a limited database[5]. Mapping a string of phonemes or syllables and the linguistic structures (positional, contextual and phonological

information) to the continuous prosodic parameters is a complex nonlinear task [7]. This mapping has traditionally been done by a set of sequentially ordered rules derived based on introspective capabilities and expertise of the individual research workers. Moreover, a set of rules cannot describe the nonlinear relations beyond certain point. The rules are usually as general as possible, and exceptions to them tend to complicate the rule-set.

The work in [5] presents the duration analysis of Indian languages (Hindi, Telugu and Tamil) using syllables as basic units. The syllable is a natural and convenient unit for speech in Indian languages. In Indian scripts syllables are generally used as characters. The syllable-like units are thus more relevant from both speech production and perception point of view. The syllable also captures some coarticulation effects. A character in an Indian language scripts is close to a syllable, and is typically of the following form: V, CV, CCV, CCVC and CVCC, where C is a consonant and V is a vowel. All the Indian languages have a common phonetic base, and the phone set consists of about 35 consonants and 18 vowels.

This paper provides the information about the factors affecting the syllable duration. These factors are phonological, positional and contextual factors. The following sections provide the information about the related works carried out by the various researchers on modeling the duration. The performance analysis based on their final results and a comparison of results of different models proposed.

II. FACTORS AFFECTING THE SYLLABLE DURATION

Acoustic analysis and synthesis experiments have shown that duration and intonation patterns are the two most important prosodic features responsible for the quality of synthesized speech [8]. A good prosodic model should capture the durational and intonational properties of natural speech. In continuous speech large number of factors affect the durations of the basic units. They are broadly classified into phonological, positional and contextual factors. The vowel is considered as a nucleus of a syllable, and consonants may present on either side of the vowel. The syllable duration may be influenced by the vowel position, category of the vowel present in the syllable and the type of the consonants associated with the vowel. Positional factors affect the durations of the basic units according to the position of the unit in the text. Different positions that affect the duration of the basic unit are: Word final position, phrase boundary, sentence ending position and word initial position. The other factors that affect the durations of the basic units depend on the contexts in which the units occur. Contextual factors include the influence of the preceding and following units on the present unit. Different manners and places of articulation of the units in the preceding and following positions also affect the duration of present unit to different extents. Apart from the factors mentioned above, gender of the speaker, psychological state of the speaker (happy, anger, fear etc.), age, relative novelty in the words and words with relatively higher number of syllables also affect the duration. These effects are difficult to describe. Also most of these effects occur relatively less frequently[5].

III. LITERATURE REVIEW ON DURATION MODELING

Rule based models are prescriptive in nature and based on implicit or explicit knowledge base. This approach has very distinct nature. In [20], Ovidiu Buza *et al*. quoted that the rules are needed to be concerned at various stages like text processing stage, speech signal processing stage and rules that adhere to languages. In text processing stage, explicit phonetic rules must be developed for syllable detection, prosodic information retrieval and text processing. In Speech signal processing stage, speech segmentation is an important task that must be carried out. (Ovidiu Buza *et al*., 2010) used Speech segmentation that includes SUV segmentation, regions detection and phonetic segmentation. The phonetic segmentation uses special association rules to realize a coincidence between phonetic groups and regions detected. The most prevalent rule-based duration model is a sequential rule based system proposed by Klatt [9], which is implemented in the MITalk system [10]. In this system, starting from some intrinsic rule, the duration of a segment is modified by rules that are applied sequentially. Models of this type have been developed for several languages [11, 12, 13, 14]. However, rule-based models often over-generalize and cannot handle exceptions well without getting exceedingly complicated. When large speech corpora and the computational means for analyzing these corpora became available, new data-driven approaches based on Classification and Regression Trees (CART) [15, 16], linear statistical models [17] and Artificial Neural Networks [18] have been increasingly used for duration modeling.

Classification and Regression Trees are models based on self learning procedures that sort the instances in the learning data by binary questions about the attributes that the instances have. It starts at the root node and continues to ask questions about the attributes of the instance down the tree until a leaf node is reached [16]. For each node, the decision tree algorithm selects the best attribute, and also the question to be asked about that attribute. The

selection is based on what attribute and question about it divide the learning data so that it gives the best predictive value for right classification. CART modeling is particularly useful in the case of less researched languages like Indian languages, for which the most relevant features that affect the duration pattern and the way they are inter-related have not been studied in detail[6].

In this approach, each single feature is taken in turn and a tree consisting of nodes containing only the conditions imposed by that feature is built. The single best tree is then kept and each remaining feature is taken in turn and added to the tree to find the best tree possible with just two features[6]. The procedure is then repeated for the third, fourth, fifth feature and so on. This process continues until no significant gain in accuracy is obtained by adding more features. The segmental durations are predicted by traversing the decision tree starting from the root node, taking various paths satisfying the conditions at intermediate nodes, till the leaf node is reached. The path taken depends on various features like, the segment identity, preceding and following segment identities, position of the segment in parent syllable and position of the syllable in parent word. The leaf node contains the predicted value of segmental duration.

In the Classification and Regression Tree(CART) model, for some test data speech rate tends to be slower around a topic shift than at other transitions, sentence-final lengthening is similar in topic shifts and continuations. However, at a topic continuation, speech rate increases significantly between the end of the present sentence and the beginning of the following sentence, topic elaborations have significantly less sentence-final lengthening than other transitions. Pauses are of similar durations for elaborations and continuations. Speech rate is faster in elaborations, but does not change at the transition. This study indicates the variability in durations due to the structure present in the spoken material, and the variability in turn reflects in the fluency[6].

The sum of products (SOP) model represents the duration for phonemes/context combination described by the feature vector f as:

$$DUR(f) = \sum_{i \in k} \prod_{j \in i} s_{i,j}(f_i) \qquad (1)$$

Here k represents the set of indices, each of which corresponds to a product term. The sum of products models capture the phenomenon of directional invariance, according to which the effects of a factor, like stress or prepausal position have always effects on the same direction[21]. In an experimental procedure that the mean duration of non-prepausal /O/ is longer than /o/. That is holding all else constant, the same vowels in prepausal position had longer duration values. The effects of sentence position do not effect in the same percentage to all vowels, so neither the factorial model and nor the additive model can capture these interactions fully. A combination of sums and products more beautifully captures and reflects the properties of duration, as directional invariance and interactions. In Several sum-of-products models proposed which model either the duration or logarithm of the duration, and all these models give near about the same results[21].

The number of possible models increases rapidly with the number of factors: it is roughly propotional to $[2^{2n-1}\ 1]$, where n is the number of factors. Many factors are ordered such that their effects are never reversed by other factors but their magnitudes may be modulated by other factors. This property is known as single factor independence and a generalized form of it is known as join factor independence, which means that the order of joint effects of two or more factors stay the same where durations of combinations of vowel identities and syllabic stress values have the same order in phrase-final and phrase-medial position. But this is not true in all cases in which no such property is exhibited for the combination of vowel identities and postvocalic consonants[21].

There is a direct mathematical link between the ordinal properties of the data set and the sum-of products model: In the data set when the factors are ordered they exhibit regular patterns of joint independence and amplificatory interactions. The sum-of-products model captures these properties of the segmental duration data. The key assumption made here is that the ordinal structure discovered in the training database can be found in the language in general (restricted to the same speaker and speaking mode). These properties exhibited are the resultant of the stable properties of the speech production apparatus. For example the non-reversal of the syllabic stress factor is linked to the superposition that stressed syllables are pronounced with more sub glottal pressure, increased tension of the vocal chords and larger articulatory excursions than unstressed syllables. The change in timing is a resultant of change in these factors[21].

The difference in the nature and properties of different consonants motivates the division of the model for consonants in a set of subsystems based on the manner of articulation. The capability of extrapolating and generalizing provided by the sum-of-products model proves more effective when the set of descriptor vectors is restricted to similar contexts. Possible subsystems to be analyzed could be the nasals, voiceless plosives, voiced plosives, fricatives and liquids. A sum-of-products model for consonants and the results for the Catalan language are given in (Febrer, A. et. al.). Along with the property of interpolation and covering up for the missing data, the duration models also have the noise suppression property, i.e. even when the data are noisy the durations estimated by the model are close to true durations.(Santen .J; 1994). In the SOP model the number of different sums-of-products grows exponentially with the number of factors. Thus it is difficult to find an SOP model that best describes the data. In addition, the SOP model requires significant preprocessing of data to correct the interaction among factors and data imbalance[21].

Neural network models are known for their ability to capture the functional relation between input-output pattern pairs (Haykin, 1999; Yegnanarayana, 1999). Campbell used a feedforward neural network trained with feature vectors, each representing six features of a syllable (Campbell, 1992). The six features are: Number of phonemes in a syllable, the nature of syllable, position in the tone group, type of foot, stress, and word class. Syllable durations are predicted with these feature vectors as input to the neural network[5]. The durations of the phonemes are estimated from the predicted durations of the syllables using the elasticity principle (Campbell and Isard, 1991). Neural network models were developed using two different databases. (1) Spoken English Corpus (SEC) from broadcast news was used for examples of fluent speech with natural prosody, and (2) Spoken Corpus Readings in British English (SCRIBE) database of phonetically rich sentence readings was used for balanced segmental information. Barbosa and Bailly used a neural network model to capture the perception of rhythm in speech (Barbosa and Bailly, 1994). The model predicts the duration of a unit, known as Inter-Perceptual Center Group (IPCG). The objective of there study is to determine whether the nonlinear neural network models can capture the implicit knowledge of the syllable duration in a language. One way to infer this is to examine the error for the training data. If the error is reducing for successive training cycles, then one may infer that the network indeed captures the implicit relations in the input–output pairs. They propose to examine the ability of neural network models to capture the duration knowledge for speech in different Indian languages using syllable as the basic sound unit. The reason for choosing the syllable as the basic unit is that, it is a natural and convenient unit for production and perception of speech in Indian languages[5].

The prediction performance of the neural network model depends on the nature of the training data used. Distributions of the durations of syllables indicate that majority of the durations are concentrated around mean of the distribution. The training data forces the model to be biased towards mean of the distribution. To avoid this problem, some post processing and preprocessing methods are proposed. Post processing methods modify the predicted values further using some durational constraints. Preprocessing methods involve use of multiple models, one for each limited range of duration. This requires a two-stage duration model, where the first stage is used to segregate the input into groups according to the number of models, and the second stage is used for prediction[5].

IV. PERFORMANCE ANALYSIS

Performance analysis carried out by the researchers for all the models is compared with their strengths and weaknesses along with the results obtained in terms of duration. The comparative analysis for each model can be inferred from the table 1.

Table 1: Comparitive analysis of all duration models

MODEL	STRENGTH	WEAKNESS
Klatt duration model	**Requires fewer resources**	**Does not work for large amount of data**
SOP duration model	**Predicts for combined context**	**Grows exponentially with increased number of factors**
CART duration model	**Large amount of data can be tested.**	**Language that lacks punctuation need to use other features like morpheme tag.**
Neural networks model	**increasing the naturalness and intelligibility**	**separate model for each syllable**

V. CONCULSION

Duration of the speech is one of the prosodic features which gives the variation in the speech depending on the timing based on the position, context and phonological factors. Different models for duration analysis proposed by various researchers have been studied. Comparison of the models based on their strengths and weaknesses is performed. Duration modeling using neural network is suggested as the better model as compared to other models. In Duration modeling using nueral networks, speech rate is faster in elaborations but does not change at the transition which gives increased variation in naturalness and intelligibility.

REFERENCES

[1] Venditti, Jennifer J. and Jan P. H. van Santen. "*Modeling Segmental Durations for Japanese Text- To-Speech Synthesis"*, In SSW3, pages 31–36, 1998.

[2] Anand Arokia Raj, Tanuja Sarkar, Satish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad, Alan W Black, *"Text Processing For Text- To-Speech Sytems in Indian Languages"* in IIIT, hyderabad, IIT Kanpur, India.

[3] spoart R., black A.W., chen S., kumar S., Ostendorf M., and Richards C., *"Normalization of non-standard words,"* Computer speech and language, pp. 287-333, 2001.

[4] Rajeshwari K S,Uma Maheshwari P, "Prosody Modeling Techniques for Text-to-Speech Synthesis Systems - A Survey" SCT, salem, IIE, Coimbatore, Tamilnadu, India.

[5] K. Sreenivasa Rao and B. Yegnanarayana," *Modeling Syllable Duration In Indian Languages Using Neural networks"* in Indian Institute of Technology Madras, Chennai, 2009.

[6] N. Sridhar Krishna and Hema A Murthy,"*Duration modeling of Indian Languages Hindi and Telugu"* in Indian Institute of Technology Madras, Chennai, 2009.

[7] M. Vainio and T. Altosaar, "Modeling the microprosody of pitch and loudness for speech synthesis with neural networks," in *Proc. Int. Conf. Spoken Language Processing*, (Sidney, Australia), Sept. 1998.

[8] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*. Prentice-Hall, Inc., 2001.

[9] Dennis H. Klatt, "Synthesis by rule of segmental durations in English sentences", In B. Lindblom and S. Ohman, Editors, Frontiers of Speech Communication Research, pages 287–300, Academic Press, New York, 1979.

[10] JonathanAllen, M. SharonHunnicut, and DennisH. Klatt, "From Text to Speech: The MITalk system", Cambridge University Press, Cambridge, 1987.

[11] Carison, R. and B. Granstrom, "A search for durational rules in real speech database", Phonetica, vol. 43, pp. 140-154, 1986.

[12] van Santen, J. P. H., "Contextual effects on vowel durations", Speech Communication, vol. 11, pp. 513-546, 1992.

[13] Bartkova, K. and C. Sorin, "A model of segmental duration for speech synthesis in French", Speech Communication, vol. 6, pp. 245-260, 1987.

[14] Simoes, A.R.M., "Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portugese", In Workshop on Speech Synthesis, ESCA, Autrans, pp. 173-176, 1990.

[15] Riley, M.D., "Tree-based modeling for speech synthesis", In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 265-273, 1992.

[16] Hyunsong Chung andMark A. Huckvale, "Linguistic factors affecting timing in Korean with applicationto speech synthesis", in Eurospeech, Denmark, 2001.

[17] van Santen, J.P.H., "Assignment of segmental duration in text-to-speech synthesis", Computer Speechand Language, vol. 8, pp. 95-128, 1994.

[18] Campbell, W., "Syllable-based Segmental Durations", In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 43-60, 1992.

[19] Mitchell, T.M., Machine Learning, McGraw-Hill, New York, 1997.

[20] Ovidiu Buza, Gavril Toderean, Jozsef Domokos, "A rule based approach to build a Text to speech system for Romanian", in proceedings of international Conference on communications, June 2010, pp. 33-36.

[21] sadaf nawaz, "Duration Model For Urdu Using The Sum Of Products Model" at National University of Computer & Emerging Sciences, in 2005.