Prosody Modeling Techniques for Text-to-Speech Synthesis Systems - A Survey

Rajeswari K C
Assistant Professor (Sr.G),
Sona College of Technology,Salem
TamilNadu, India

Uma Maheswari P
Professor,
Info Institute of Engineering, Coimbatore,
TamilNadu, India

ABSTRACT

This paper presents a study on prosody modeling for speech synthesis. Any Text to Speech system comprises of two phases. One is text analysis and second is speech synthesis. The task of text analysis is to find the words and the task of speech synthesis is to generate the speech. To attain this, different models are available such as text as language models, grapheme to phoneme models, full linguistic analysis model and complete prosody generation model. In complete prosody generation model, the quantities like phrasing, stress and the like are determined to generate naturalness bearing synthetic voice. Towards generating such a speech, an explicit prosodic model is required. This makes the speech more understandable. Many researches have been done in this stream, but still better solution is required. In this paper, the strength and weaknesses of different approaches of prosody models are discussed.

Keywords

Prosody model, speech synthesis, Text to Speech systems

1. INTRODUCTION

In recent years, researches have been done to provide high quality synthetic speech with no compromising in terms of naturalness and intelligibility in Text to Speech systems. The prosodic information significantly contributes for the quality of synthetic speech in TTS. So the modeling of prosody plays a vital role when Text to Speech systems of high quality is expected. The necessity for striving on prosody modeling is quite interesting. The process of conversion of Text-to-Speech can be achieved easily. But, inclusion of prosodic information makes it sensible by providing naturalness as human voice and thereby makes human computer interaction more likely.

Prosody modeling is devised to model the following constituents of prosody: duration, intonation and phrasing. The rule based approach and the corpus based approach are the two major approaches for prosody modeling. 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. The corpus is analyzed automatically to create prosodic models which are then evaluated on test data. Based on the performance of the test data, the models are improved. In [6], the work by N.Sridhar Krishna *et al.*, mentioned that the challenge in prosody modeling is the consideration of various parameters like syllables for duration model as it contributes for sufficient duration information and from [7] phonemes for intonation model contributes for pitch pattern. Thus, prosody modeling significantly accounts for producing natural sounding synthetic speech in speech processing applications.

2. PROSODY IN TTS

Prosody is an important aspect of speech that helps to maintain expressiveness and intelligibility in speech synthesis systems Prosodic components are Pitch, duration, accent and phrasing. In [2], G.L. Jayavardhana Rama *et al*. developed a complete Text-to-Speech system. The system was designed for Tamil language. Attempts have been made to make the speech natural and provided good synthesis for alien words. One of the prosodic components, pitch marking was studied and this information was used to concatenate speech at synthesis. A preliminary TTS for Tamil language was developed by (G.L.Jayavardhana Rama *et al*., 2001) insist the need for prosody. In [1] syllable was used as basic unit in concatenative speech synthesis process that helps to provide better prosody.

3. PROSODY MODELING

Prosody modeling is the process of building computational models to produce prosodic variations in synthesized speech automatically. Many methods that contribute to prosody in speech exist. Examples are:

Rule based methods, which involve manual analysis of segment durations. But it does not work for large amount of data. This method depends on linguistic and phonetic literature about the factors that affect duration of the units (segments, syllables or phones). In general, rule based methods are difficult to study, due to complex interaction among the linguistic features at various levels.

Statistical data-driven methods are attractive when compared to rule based methods. This method works when large phonetically rich sentences are present in the corpora. This method is based on either parametric or non-parametric model that uses probability or likelihood functions.

Hybrid approach uses a combination of both rule based and statistical method and offers better prosodic model.

3.1 Rule based prosody model

Rule based models are prescriptive in nature and based on implicit or explicit knowledge base. This approach has very distinct nature. In [10], 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.

Problems with rule based approach.

Even though, explicit rule sets are defined, coverage of syllables is very limited. The rule sets are not complete, so it lacks 100% correct syllable detection and offers only up to 98% of correct syllable detection.

Comments

Rule based approach has to be designed adhering to linguistic features of the language concerned. This approach does not ensure feasibility for multilingual speech systems.

3.2 Likelihood based prosody model.

In general, the task of prosody model in concatenative TTS is to predict pitch, duration and energy value either in explicit form or in implicit form. In probability based prosody model, models are built to give prosody predictions. In [4], six models were built namely prosody target model, prosody transition model, duration target model, duration transition model, energy target model and energy transition model. All the models are trained corresponding to a context dependent decision tree and data that are clustered in each leaf are taken as Gaussian mixture. In [5], a decision tree T_i was used and are traversed in regard to the context of the node to get the corresponding Gaussian mixture M ^{l} where l indicates the lth leaf under decision tree $T_{i.}$ The minus log of the likelihood is defined as the cost of the candidate x to Ti, given the context.

$$C\left(\mathrm{x_i}\ |\mathrm{M_i^1}\right) = -logP(\mathrm{x_i}\ |\mathrm{M_i^1})$$

The calculated costs with their weights are called target cost and transition cost.

Problems with likelihood

- Cost can be smaller than zero.
- Each Gaussian mixture is to optimize the output locally but not globally.
- Hard to tune, the weights for different models.

Comments.

Rich language specific skills are required for the weight tuning. Cost factor cannot hold negative value and it becomes invalid when it is less than zero. This method lacks optimization.

3.3 Posterior based prosody model

In [5], posterior probability is defined as P (M_i^l| x_i). The following formula is used to calculate the posterior probability value.

$$P(\mathrm{M_i^1}|\mathrm{x_i}) = \frac{\mathrm{P(x_i\ |M_i^1)P(M_i^1)}}{\sum_{j=1}^{N}\mathrm{P(x_i\ |M_i^1)P(M_i^1)}}|$$

$$C_{new} = -\log P(M_i^1|\mathrm{x_i}) = -\log\frac{\mathrm{P(x_i\ |M_i^1)P(M_i^1)}}{\sum_{j=1}^{N}\mathrm{P(x_i\ |M_i^1)P(M_i^1)}}$$

The cost function C_{new} can be greater than zero since both the numerator and denominator are positive and numerator is always smaller than denominator and definitely gives the value greater than zero. So, it is assured that the posterior probability can achieve global optimization better than likelihood function.

Problems with posterior based model

Real data sets in limited range only use this approach.

Comments

Posterior probability is less applicable for real data sets but reduces the negative impacts of the over-train issues and weight adjustment issues. The experimental results obtained by (Wei Zhang *et al.*, 2009) shows posterior probability based prosody model has greater advantages on robustness, flexibility and overall quality when compared to likelihood function based model.

3.4 Hybrid model

The hybrid model is a combination of both rule based and statistical model based approach. CART is a hybrid model widely used for prosody modeling. As the earlier work was done using pre-clustering the syllables based on their position in the word. Results of listening tests show that pre-clustering is inadequate. In [8], (Ashwin Bellur *et al.*, 2011), in his work used a cluster unit framework based on clustering the syllables of same type, preferring a set of higher level prosodic and phonetic features. As a part of this, CART is used and acoustic distance measure is defined to distinguish between syllables. First, the syllable set is defined and then the features are selected. The choice of features must be selected in such a way that they enable gross acoustic properties of the syllables are captured.

In [8] the acoustic distance measure is calculated, using this and feature set, a decision tree CART was built. As a result, the decision tree predicts whether the phrase boundary exist after the word. In [8] a new feature say morpheme tag was used to identify the phrase boundary predictions. Two synthesizers, one with manually marked phrase boundaries and other with the facility of automatically predicting the phrase boundaries using decision tree were used and MOS test was conducted and observed that the synthesizer that automatically predicts the phrase boundaries are fairly good when compared to one which annually predicts the phrase boundaries. In [18] , a data-driven modeling of three fundamental features F0, intensity and segmental duration using CART based approach was tried for Czech language speech system and were able to achieve better assessment of intonation. This model relies on separate CART tree for each phoneme and this may not work for large corpus systems.

Problems with hybrid model.

Morpheme tags need to be listed out separately for each language especially for the languages that lacks punctuation.

Comments.

Hybrid model takes the advantage of both rule based and statistical approaches but it has to be further analyzed whether it is applicable for TTS that suits as many languages as possible.

3.5 Prosody model and HMM based synthesis

Hidden Markov Model is one of the best models currently in use for most of the speech synthesis systems. In [17], the drawback of current HMM was stated that it lacks variations in prosodic parameters. To overcome the drawback, an improved HMM based synthesis approach was proposed that uses linguistic oriented approach through which high level linguistic features can be extracted from the text, so that the quality of prosody modeling can be improved. In [20], it has been highlighted, the use of a combination of an explicit and implicit duration model helps to improve the quality of speech in HMM based speech synthesis systems.

4. OTHER CONTRIBUTIONS OF PROSODY MODEL

In [12], it has been developed a method for modeling and generating the prosodic component, pitch using Hidden Markov model. This method uses S-CART for predicting prosodic breaks and U-CART for generating pitch contour. In [13], HMM models combined with ANN models were discussed by (Hung-Yan GU *et al*., 2010) to promote simultaneously both prosodic and acoustic fluencies. In [14], Maximum entropy based automatic prosody labeling framework that exploits both language and speech information was addressed. In [15] , it has been demonstrated a prosodic feature extraction tool to identify regional Tamil dialects by monitoring the parameters duration, F0 and other important values like range, movement and slope. It is important to consider all these methods and parameters while designing prosody model for any language specific speech synthesis system. In [16], a multi level context dependent prosodic model has been defined to estimate the linguistic units that contribute for the variations of prosodic parameters on each level independently. By adopting this method performance is improved in terms of both better duration predictions and relative prediction error. In [17], an improved HMM (Hidden Markov Model) has been proposed in order to overcome the drawback of present HMM based synthesis that lacks variation in prosodic parameters. In [20], an external duration model which is one of the important parts of any prosodic model was used to analyze three different approaches to improve the quality of speech in HMM based speech synthesis systems. The three different approaches are an explicit duration approach, implicit duration approach and hybrid approach. The results were comparable that, in explicit duration approach, better estimation of phone duration was obtained. In implicit approach, phone duration was not better compared to explicit approach. The hybrid approach takes the advantages of previous approaches. A linguistic oriented approach is adopted to extract high level linguistic features from the text to improve prosody modeling. The work [19] carried out by (Yu-Lun Chou *et al*., 2010) shows significant exploration of prosodic information when PLM (prosody labeling and modeling) method was used for spontaneous speech applications. Table. 1 highlights the strength and weaknesses of each prosody modeling technique.

TABLE 1. Comparison and Discussion

Technique	Strength	Weakness
Rule based approach	Requires fewer resources.	Prescriptive in nature. Does not work for large amount of data.
Statistical approach	Large amount of data can be tested.	Data sparsity. Less applicable for real data sets. Lacks optimization.
Hybrid approach	Combines advantages of both rule based and statistical approach.	Language that lacks punctuation need to use other features like morpheme tag.
Context dependent model	prosodic forms can be jointly observed and each prosodic level can be modeled and controlled independently from each other	Degradation in relative error performance and not suitable for laboratory speech
Labeling and modeling approach	Able to obtain rich prosodic information	Most appropriate only in dialog corpus
Any prosodic model in HMM based system	Better prosodic information, best suited for Unit selection paradigm, able to achieve Intelligible speech	Appropriate work to be carried out for bringing naturalness bearing speech.

5. PROSODY MODELING IN VARIOUS SYNTHESIS SYSTEMS

Prosody modeling is applicable to many other synthesis systems like Concept-to-speech, Emotional speech synthesis and Table-to-speech synthesis, etc. The great challenge of browsing complex data tables has been addressed using diction based prosody modeling in [21].Table-to-speech synthesis needs proper navigation manner and semantic structure of data understanding. For this, a set of prosodic parameters was derived and analyzed in concern with phrase accent tones and pauses without violating consistency in cell content and visual structure. Prosody model significantly contributes to emotional speech synthesis systems is evident from [22] and [23]. In [22], the pitch contour is decomposed into hierarchical structure having sentence, prosodic word and syllable/ sub syllable level. It has been observed that, such hierarchical prosodic structure reduces the prediction error in the model. The use of Discrete Legendre polynomial coefficient for pitch contour conversion function, GMM for prosodic word level conversion, Linear regression based clustering for sub syllable level features put together improves the prosodic conversion in Mean square Error measure. In [23], it has been clearly stated the importance of prosody modeling for emotional speech synthesis system by investigating whether prosodic features alone can attain appropriateness (emotional expression in concern with verbal content) and efficiency (emotional expression in concern with speaker’s attitude) in communications. The result obtained shows that the prosody features helps to obtain meaningful results for some emotions but proves that not necessary to use special emotional rich corpus.

6. CONCLUSIONS

In this paper, the various techniques used for prosody modeling for Text to Speech synthesis systems were analyzed. The tools that are used for developing prosody model are also highlighted. The issues, advantages and disadvantages of each method found out that the scope for research in prosody modeling is highly expected to provide high quality Text-to-Speech synthesis systems. The challenges that have been identified may be over aid with the help of linguistic experts and technical expert's effort. Prosody modeling has its significance not only in Text-to-Speech synthesis systems but, also in other systems like Table-to-Speech and emotional speech synthesis systems are also analyzed and highlighted. Hence, the future research is focused in developing the prosodic model to improve the quality of speech synthesis systems.

7. REFERENCES

[1] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, "Text-to-speech synthesis using syllable like units," in National Conference on Communication, Kharagpur, India, Jan 2005, pp 277-280.

[2] G.L.Jayavardhana Rama, A G Ramakrishnan, R. Muralishankar and Vijay Venkatesh" Thirukkural – A text to speech synthesis system". Proc. Tamil Internet 2001, Kuala Lumpur 2001, 92-97.

[3] Vinodh M Vishwanath, Ashwin Bellur, Badri Narayan K, Deepali M Thakare, Anila Susan, Suthakar N M and Hema A Murthy,"Using Polysyllabic units for Text to Speech Synthesis in Indian languages," Proceedings of National Conference on Communication,pp.1-5, 29-31, Jan. 2010.

[4] X.J. Ma, W. Zhang, W.B. Zhu, Q. Shi and L. Jin, "Probability Based Prosody Model for Unit Selection", ICASSP 2004, Montreal, Canada

[5] Wei Zhang, Liang Gu and Yuqing Gao "Recent improvements of probability based prosody model for unit selection in concatenative Text to Speech", in the proceedings of ICASSP 2009, pp 3777-3780

[6] N. Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, A. G. Ramakrishnan, "Duration Modeling for Hindi Text to Speech Synthesis System", in Proc. ICSLP 2004, South Korea, 2004.

[7] A. S. Madhukumar, S. Rajendran and B. Yegnanarayana, "Intonation component of a Text to Speech system for Hindi", Proceedings of International journal of Computer Speech and Language, 1993, Volume7, pp 283-301

[8] Ashwin Bellur, K Badri Narayan, Raghava Krishnan K, Hema A Murthy, "Prosody modeling for syllable based concatenative speech synthesis of Hindi and Tamil", in National conference on Communications, Jan 2011, pp 28-30.

[9] Samuel Thomas, M. Nageshwara Rao, Hema A.Murthy and C.S. Ramalingam, "Natural sounding TTS based on syllable-like units," in the proceedings of the 14th European Signal Processing Conference, Florence, Italy, Sep 2006.

[10] 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.

[11] G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and R Prathibha, "A Complete Text-To-Speech Synthesis System in Tamil", in 0-7803-7395-2/02, IEEE proceedings of ICASSP,2002.

[12] Chi-Chun Hsia, Chung-Hsien Wu, and Jung-Yun Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM based speech synthesis', in Inernational journal of Audio, Speech and Language processing, Nov 2010,Volume 18, pp,1994-2003.

[13] Hung-Yan GU, Ming-Yen LAI and Sung-Feng TSAI, "Combining HMM spectru models and ANN prosody models for speech synthesis of syllable prominent languages", in Inernational journal of Audio, Speech and Language processing, 2010, pp,451-454.

[14] Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan,"Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework", in Inernational Journal of Audio, Speech and Language processing, May 2008, Volume 16, pp,797-811.

[15] Raja Mohamed S, Raviraj P," Prosodic Feature Extraction for Regional Tamil dialects", in Inernational Conference on emerging Trends in electrical and Computer Technology, March 2011, pp 922-925.

[16] Nicolas Obin, Xavier Rodet and Anne Lacheret Dujour,"A multi-level context-dependent prosodic model applied to duration modeling", in the tenth annual conference,Inerspeech,France,2009.

[17] Nicolas Obin, Pierre Lanchantin, Mathieu Avanzi, Anne Lacheret-Dujour and Xavier Rodet," Towards improved HMM-based speech synthesis using high-level syntactical features", in the fifth International Conference on Speech Prosody, Chicago, 2010.

[18] Jan Romportl and Jiri Kala, "Prosody Modeling in Czech Text-to-Speech Synthesis", in the proceeding of Sixth International workshop on speech synthesis, 2007.

[19] Yu-Lun Chou, Chen-Yu- Chiang, Yih-Ru Wang, Hsui-Min Yu and Sin-Horng Chen, "Prosody labeling and modeling for Mandarin spontaneous Speech",in the International Conference on Speech Prosody, Chicago, 2010.

[20] Javier Latorre, Sabine Buchholz, Masami kamine, "Usages of an external duration model for HMM-based speech synthesis", in fifth International conference on Speech Prosody, Chicago, 2010

[21] Dimitris Spiliotopoulos, Gerasimos Xydas, and Georgios Kouroupetroglou," Diction Based Prosody Modeling in Table-to-Speech Synthesis", in LNAI 3658, pp. 294–301, 2005.

[22] Chung-Hsien Wu, Chi-Chun Hsia, Chung-Han Lee, and Mai-Chun Lin," Hierarchical Prosody Conversion using Regression-Based Clustering for Emotional Speech Synthesis", in IEEE Transactions on Audio, Speech and Language Processing, Vol.18, No.6, August 2010.

[23] Dan-ning Jiang, Wei Zhang, Li-qin Shen and Lian-Hong Cai," Prosody Analysis and Modeling for Emotional Speech Synthesis", in IEEE proceedings of ICASSP,0-7803-8874-7/05,pp 281-284, 2005.