Improved prosody module in a Text-to-Speech system

Wen-Wei Liao and Jia-Lin Shen
Research Center, Delta Electronics, Inc.
wei.liau@delta.com.tw, lynn.shen@delta.com.tw

Abstract

The newly-developed prosody module of our text-to-speech (TTS) system is described in the paper. We present two main works on it's establishment and improvement. On the basis of potential factors influencing prosody parameters, inclusive of duration, pitch and intensity, the prosody model is built as groundwork of this module which is superior to the former rule-based one in generation of natural prosody. In addition, due to the current model's flaw in prediction of the pitch contour, we further employ an technique named "Soft Template Mark-up Language"(STEM-ML) to improve the smoothness of intonation which has the crucial influence on the naturalness of synthetic speech.

Results of the evaluation indicate that the new prosody model is precise enough to predict reliable prosody parameters' values and with the STEM-ML technique, the prosody module can further yield 14.75% reduction in the root mean square (RMS) error of the predicted pitch contour.

1. Introduction

In consideration of severe limitation in the resource afforded by some applications in need of speech response, we choose to develop one storage-saving TTS system which has functioned successfully in our spoken dialogue system. Accordingly, the acoustic inventory used in our system is simply composed of about four hundred base syllable units whose duration and pitch contour will be modified with the algorithm called Pitch-Synchronous Overlap-Add (PSOLA) [1][12] in the synthesizing phrase.

In order to produce natural-sounding synthetic speech, the generation of prosody plays a key role and is a difficult issue yet. Outperforming rule-based method [13][14] which was employed in our system previously, the newly-built statistical model based on sum-of-products approach with key factors affecting prosody [7][8][9][10][11] can predict more accurate values of prosody parameters. And in general, the intonation which is characterized by the pitch contour seems more crucial to the naturalness and intelligibility of synthesized speech in comparison with other prosody elements such as duration, intensity etc [6]. Nevertheless, the pitch contour generated by our current prosody model is still short of smoothness. As a result, we further concentrate our work on this problem. Based on the F0 (fundamental frequency) mean value predicted by the current prosody model, an technique named STEM-ML [2][3][4][5] is adopted to overcome this shortcoming. In the evaluation phrase, we prove that this technique can help to reduce the difference between the predicted and observed pitch contours, which means that a more natural intonation is achieved.

The paper is organized as follows. In the chapter 2, we present the prosody modeling in our system, The chapter 3 reports STEM-ML technique and the result of implementation. The conclusion is described in the chapter 4.

2. Prosody modeling

In general, prosody mainly consists of duration, pitch, intensity of the spoken unit which is one syllable in terms of Mandarin. Besides, the break between units is one of it's important elements as well. Therefore, one utterance's prosody can be regarded as the elaborate composition of these four perceivable characteristics. And the variation in prosody stem from a lot of factors in different dimensions which can be observed in the real speech corpus such as the syllable's position in the sentence, lexical tone even the speaker's emotion and so on. Furthermore the complex interactions between factors further lead to another difficulty in designing the prosody model. As a result, in addition to inferring the reliable factors influencing the prosody, to model the interactions between factors intelligently is also a challenge in this work. .

2.1 Modeling

2.1.1 Base model and sub-models

The potential factors affect one characteristic simultaneously and have additive, multiplicative or repulsive interactions . Thus, it's troublesome to derive their eventual combined effect on the characteristic. However, for the purpose of assuring that the basically reasonable value for the characteristic can be preserved, one major factor in possession of dominant influence are elected to build the base model while the remaining minor factors take charge to constitute sub-models. In other words, under this framework, the base model provides fundamental value for the characteristic and sub-models act on this base value (BV for short) through the mechanism modeling their interaction to obtain the ultimate characteristic value (CV for short).

2.1.2 Ratio of characteristic value to base value (RCB)

In order that this concept of modeling can be put into practice concretely, the training sample for sub-models, namely the CV of each syllable has to be normalized by it's corresponding BV beforehand. Thus, pre-processed CV is computed as follows.

$$RCB = \frac{CV}{BV} \tag{1}$$

2.1.3 Mechanism

In brief, the ultimate objective of the mechanism devised here is to make combined effect of minor factors quantized to one RCB value used as the multiplier of the BV. The interactions of minor factors are modeled by the approach of sum-of-products and the predicted CV is computed as follows.

$$\begin{aligned} \hat{CV} &= RCB_{comb} \times B_i \\ RCB_{comb} &= \sum_{i}^{SMN} \sum_{j}^{i} C_{ij} S_i^{m_{ij}} S_j^{n_{ij}} \end{aligned} \tag{2}$$

where

Bi is the parameter of the base model for the characteristic i and

SMN is the numbers of sub-models for the characteristic i and

Si is the parameter of the sub-model i and

Cij is a coefficient associating the sub-model i and sub-model j and

mij and nij represent the stress of sub-model i and sub-model j respectively.

2.1.4 Factors

We infer seven potential factors crucial to the characteristics in prosody.
Those are listed and described briefly as below.

- **Base syllable (BS)**

  408 identities

- **Lexical tone (LT)**

  4 lexical tones and one neutral tone

- **Left and right context tones (LRCT)**

  175 levels: 25(bi-tone) + 125(tri-tone)

- **The syllable's position in the word and the syllable number of one word (SInW)**

  15 levels: 1+2+3+4+5 (longest word length)

- **The word's position in the phrase (WInP)**

  4 levels: $WInP = \frac{WordIndex \times 4}{WordNumber \quad OfPhrase}$

- **Right context break (RCBk)**

  4 levels: inter-syllable pause, inter-word pause, comma, period

- **Right context initial (RCIt)**

  32 identities

Accordingly., four kinds of base models and seven kinds of sub-models will be established in light of these factors.

2.2 Estimation

2.2.1 Corpus

Recorded by a single female speaker, the speech corpus contains 3657 sentences (70000 syllables;about 7 hours) with moderate intonation and constant speaking rate. In terms of linguistics ,the properly-designed one has enough coverage to tackle diverse variability of prosody. Among these sentences, around 3200 ones are used as training data and the rest of them are reversed for the purpose of evaluation. The syllable boundaries in the waveform are further calibrated manually after aligned by the automatic speech recognizer.

2.2.2 Objective function

The distortion rate (DR) is defined to measure the precision of predicted value.

$$DR = \left| \frac{O - P}{O} \right| \qquad (3)$$

where

O is the occurrence's CV and

P is the predicted CV.

Accordingly, the objective function is defined as average DRs of all occurrences in the training data.

$$O = \frac{1}{N}\sum_{i} DR_{i} \tag{4}$$

where N is the number of training samples.

2.2.3 Approach

- **Model**

Both base models and sub-models have only one parameter. The parameters of base models and sub-models are calculated as the average of observed occurrences's CVs and RCBs which correspond to them in the training corpus respectively.

$$\mu = \frac{1}{oN}\sum_{i}^{oN} o_{i} \tag{5}$$

where

μ is the parameter of the model and

oi is observed occurrence whose value is either RCB or CV depending on whether the model is a sub-model or base model and

oN is the number of occurrences.

- **Coefficients and Stress**

Firstly, the initial values of coefficients and stress are calculated by means of linear least square error and given value 1 respectively. And furthermore beginning with the initial values, Levenberg-Marquardt algorithm [15][16] with numerical differentiation is employed to find the optimal values of these parameters with the goal of minimizing the objective function O defined in (4).

2.3 Characteristic model

In this section ,the characteristic models, inclusive of duration, pitch and intensity are discussed in terms of the related factors and precision. And as for the break characteristic, we straightforwardly give each type of break an empirical length instead of building the model.

2.3.1 Duration

This characteristic means the time for which one syllable endures in the utterance. Since the boundaries between syllables are demarcated precisely by hand in our speech corpus, it is straightforward to calculate the syllable's duration.

- **Factors**

Major *BS*

Minor *1. LRCT 2. SInW 3. WInP 4. RCBk 5. RCIt*

- **Speaking rate**

Each syllable's duration in the corpus needs to be normalized by the utterance's speaking rate (SR) which is estimated as:

$$SR = \frac{1}{SylN}\sum_{i}^{SylN}\frac{D_i}{\overline{D}_{BSi}} \tag{6}$$

where

D_i is duration of one syllable (named Si), $\overline{D}_{BSi}$ is average duration of base syllable corresponding to Si in the corpus and $SylN$ is the number of syllables in one utterance.

2.3.2 Pitch

Pitch here means the one syllable's pitch contour which is depicted with F0 (fundamental frequency) computed at a constant frame rate. In our task, this characteristic is discussed in two separate aspects, namely the pitch contour 's F0 mean (FM for short) and F0 shape. The former can leave the each syllable's pitch contour in a proper level and the later considerably concerns it's smoothness.

In this chapter, we only concentrate discussion on the F0 mean. In the other hand, one technique named STEM-ML is adopted to deal with F0 shape. This work will be reported in next chapter.

- **Factors**

  Major *LT*

  Minor *1. BS 2. LRCT 3. SinW 4. WinP 5. RCBk*

- **FM rate**

Each syllable's FM in the corpus needs to be normalized by the utterance's FM rate (FMR) which is estimated as:

$$FMR = \frac{1}{SylN}\sum_{i}^{SylN}\frac{F_i}{\overline{F}_{Tonei}} \tag{7}$$

where

F_i is FM of one syllable, $\overline{F}_{Tonei}$ is average FM of Tonei in corpus and $SylN$ is syllable number in one utterance.

2.3.3 Intensity

This characteristic means one syllable's volume in one utterance. We measure one syllable's intensity with it's power. The power can be estimated as below.

$$Power = \log 10\left(\frac{\sum_i X_i^2}{N}\right) \tag{8}$$

where

Xi and N are the sample value and number of samples respectively.

- **Factors**

  Major *LT*

  Minor *1. BS 2. LRCT 3. SInW 4. WInP 5. RCBk*

- **Power rate**

Each syllable's power in the corpus needs to be normalized by the utterance's power rate (PR) which is estimated as:

$$PR = \frac{1}{SylN} \sum_{i}^{SylN} \frac{P_i}{\overline{P}_{Tonei}} \tag{9}$$

where

P_i *is power of one syllable (named Si),* $\overline{P}_{Tonei}$ *is average power of Tonei in the corpus and* $SylN$ *is syllable number in one utterance.*

2.4 Evaluation

The evaluation set consists of 300 stentences, exclusive of the sentence in the training set and the precision of the characteristic models are evaluated with DR defined in (3). The results are shown in the Table 1.

Model	Precision
Duration	11.35%
Pitch	5.6%
Intensity	1.98%

Table 1. The preciosion of characteristic models.

3. Soft Template Mark-up Language (STEM-ML)

The prosody model developed in the previous chapter establishes the groundwork for the prosody module of our TTS system. However, since it merely aims at assuring the accuracy of F0 mean without putting emphasis on the F0 shape, the predicted pitch contour lacks smoothness. For the sake of this drawback , we proceed to employ an model devised by Kochanski, G. P. et al. and called STEM-ML that is abbreviated from "Soft Template Mark-up Language".

It is a tagging system which computes the pitch contour in light of a set of tags serving to interpret the variation in the pitch contour more humanly. In order to make the artificial pitch contour closer to the real one, the mechanism of model has to comply with the constraints actually existing in the human uttering process. Thus, each tag concretely takes effect by imposing constraints on prediction of the pitch curve.

As a result, the pitch curve is eventually generated by the model on condition that those constraints come to a compromise. In fact, such compromise can be considered to be the result of tradeoff between two events with reversal interaction, namely effort and error. The effort term stands for physiological energy consumed in the uttering processing and the error one means the communication error rate caused under the current effort. Obviously, they behave contrary to each other. With more effort, the uttering can achieve more accurate expression on words while the error results from little effort spent on uttering. In conclusion, the model can be also thought to predict the pitch curve with the goal of minimizing the sum of effort and error caused in the uttering process.

3.1 Model

3.1.1 Soft templates

Soft templates consists of pitch contours of four lexical tones (tone1,tone2,tone3 tone4) and the neutral tone (graphed in Figure 1).Since the syllable's tone shape varies considerably due to the affection from syllables nearby, five templates aren't apparently equal to express such variability . However, the adjective, "Soft" significantly implies that their shapes are allowed to change properly (see Figure 2). Consequently, these templates with the elastic property can form smoother pitch contour.

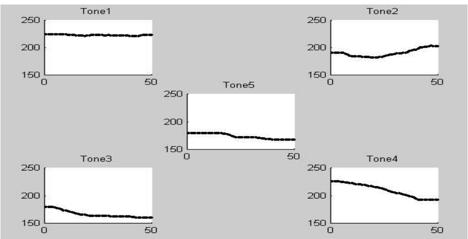

Fig 1. 5 tone templates.

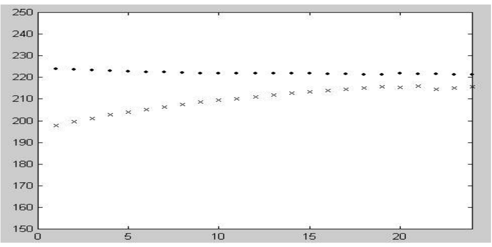

Fig2. A example of how one syllable is effected by it's neighbor. Succeeding to Tone3, the original shape of Tone1 template (dot line) is bended under control of the model and turns out to be the one (cross line) with tilt in the front part.

3.1.2 Tags

The tags function as adjustable parameters of the model. Each kind of tag governs the pitch curve's variability in one certain dimension. For instance, the tag ***smooth*** determines the permissible velocity of change in pitch values and the priority over one pitch curve's shape and F0 mean is dependent on the tag ***syllable-type*** **.** Thus, the tags have the critical influence on the generated pitch curve's look and should be given proper values so that the one can has good quality. The estimation of tags will be reported in the section 3.3. 10 kinds of tags in total are used in our work as listed below.

max, min, base, range, add, slope, smooth, pdroop, adroop syllable-type, syllable-strength

Moreover, to account for the more detailed pitch curve's variation inside one word, the tag ***syllable-strength*** is specially given a distinct value depending on the syllable's position inside the word. As the case for the sub-model **SInW**, this actually leads to 15 kinds of ***syllable-strength*** tags considered in the model.

3.2 Calculation of pitch contour

Based on the templates and tags, the process of calculating the pitch curve mainly includes two steps.

Step1

The first step purposes to prepare the plain templates assembling a prototype of the pitch curve.

1. Select the templates according to each syllable's tone among five basic templates as mentioned above.

2. The templates have to be modified to conform to the desired duration and F0 mean predicted by the prosody model.

Step2

In this step, the tags start to be applied in the calculation along with ready templates. The constraints on generation of the pitch curve are realized by translating the tags to a number of conditional equations with pitch instants (F0) as unknown variables to be solved. One tag can brings in one equation or one group of equations. For example, the ***slope*** tag which controls the pitch's increasing or decreasing rate in the phrase level yields the equation $Pt+1 - Pt = S$ where P and S are the pitch variable and the ***slope*** tag's value respectively. These joint conditional equations can be written as the form $Ax = b$ where A is matrix with rows composed of the coefficients in the left-hand side of all equations and x is a vector containing the unknown variables and the b is a vector with elements consisting of the right-hand side of all equations .Consequently, the pitch values of the curve are the solution of the algebraic problem $Ax = b.$

Furthermore, the calculation proceeds in the order of phrase level and the syllable level. Riding on the phrase's pitch curve solved firstly, the syllable's one is calculated . The process in the phrase level aims at deciding the trend of the whole resultant pitch curve which is finally obtained in the syllable level. Step2 is illustrated in Figure 3.

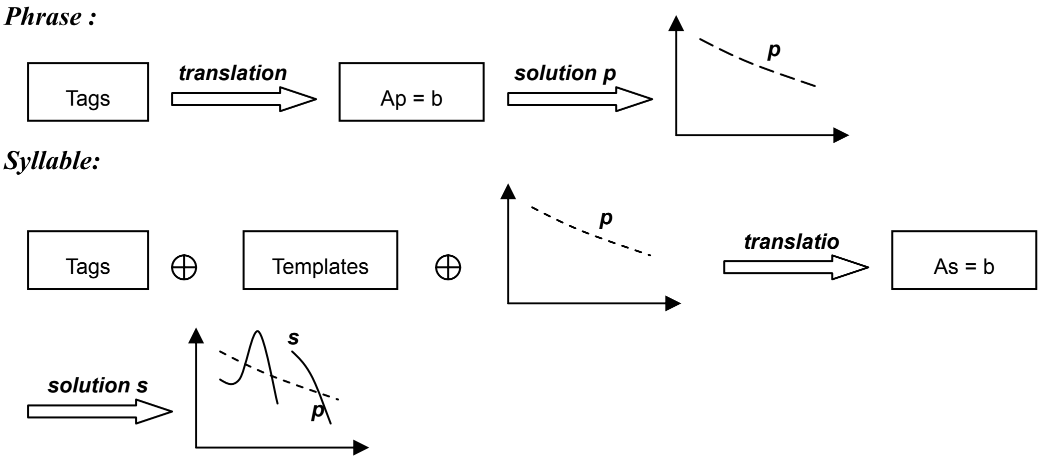

Figure 3. The procedure for calculating pitch contour which is carried out in the order of the phrase and syllable levels .

A real case for the syllable's pitch curve (dot line) and phrase's one (dash line) generated by the model is plotted in Figure 4 .

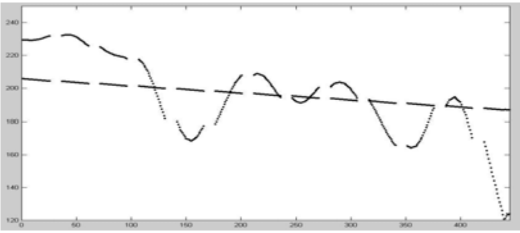

Fig 4. A example of the pitch contour generated by the model.

3.3 Estimation of tags

3.3.1 Approach

We estimate the tags by data fitting with the objective to minimize root mean square (RMS) error of the predicted F0 in comparison with the observed F0 in the data. The development data set composed of 300 sentences is designed to cover enough occurrences for each kind of tag and templates. Similarly, Levenberg-Marquardt algorithm with numerical differentiation is employed in this task. In addition, the number of pitch samples per syllable in the data is normalized to a constant and the syllable's un-voiced position is excluded.

3.3.2 Results

The process of minimization ends in RMS error that is equal to 16.16 (Hz) One example of fitting results is shown in Figure5.

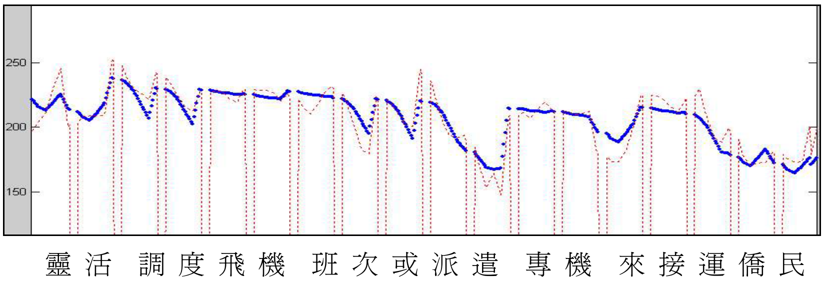

Fig5. A example of one utterance's simulated pitch curve (dot line) along with the real one (dash line) in the data-fitting result.

3.4 Evaluation

The evaluation data set is the same to one in the chapter 2 and the prosody model is used as the baseline of this task. In the baseline, the templates are unvaried in the shape but shifted to have the F0 mean predicted by the prosody model. The accuracy of the pitch contour generated by the model is measured by the RMS error of predicted F0 .The result is shown in the Table 2.

Prosody model (baseline)	19.46 (Hz)
Prosody model + STEM-ML	16.59 (Hz)

Table 2. The RMS F0 error of the pitch contour generated by the prosody model and prosody model + STEM-ML.

The result indicates that based on the prosody model, this technique can further reduce 14.75% RMS error of F0 in the predicted pitch contour.

4. Conclusions

In this paper, we successively report two works on the development of the prosody module in our TTS system, Firstly, the prosody model based on the framework of base models and sub-models and sum-of-products approach has been proven to have the capability of predicting reliable prosody parameters' values. Furthermore, the employment of the STEM-ML technique further bring in the improvement in the smoothness of the intonation which the prosody model originally lacks

In order to raise the accuracy of the prosody model, the refinement of the mechanism in the modeling should be necessary . Besides, we consider expanding the types of STEM-ML tags defined in our system to generate more natural and lively intonation.

References

[1] Moulines, E. and Charpentier, F. Pitch Synchronous Waveform Processing
Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication* 9, 453-467, 1990.

[2] Kochanski, G. P. and Shih, C., “Prosody modeling with soft templates,” accepted by *Speech Communication.*

[3] Kochanski, G. P. and Shih, C., “Automatic modeling of Chinese intonation in continuous,” in Proceedings of EUROSPEECH 2001, pp.911-914.

[4] Grep P. Kochanski and Chilin Shih, “Stem-ml: Language independent prosody description,” in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[5] Chilin Shih and Greg P. Kochanski, “Chinese tone modeling with stem-ml,” in *ICSLP*, Beijing, China, 2000

[6] Plumpe, M., Meredith, S. Which is more Important in a concatenative Text To Speech System – Pitch, Duration or Spectral Discontinuity ?, *Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan, Autralia, Nov. 25-29, 1998

[7] Van Santen, J. P. H. Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language*, 8, 1994.

[8] *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, Richard Sproat, editor, Kluwer Academic Publishers, 1998.

[9] J. van Santen, “Prosodic modeling in text-to-Speech synthesis”, *Proceedings of EuroSpeech’97*, KN-19,Rhodes 1997.

[10] Febrer, A.; Padrell, J.; & Bonafonte, A. 1998. Modeling phone duration: Application to Catalan TTS. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 43-46.

[12] K.M. Law and Tan Lee, “Cantonese text-to-speech synthesis using sub-syllable units", in *Proceedings of the 7th European Conference on on Speech Communication and Technology*, Vol.2, pp.991 - 994, Aalborg, Denmark, September 2001.

[13] L.S.Lee, C.Y. Tseng, and M. Ouh-Young, “The synthesis rules in a chinese text-to-speech system”, *IEEE trans. Acoust., speech, signal Processing,* Vol. 37, pp. 1309-1320, 1989.

[14] 許文龍,“使用時間比例基週波形內差之國語語音合成器”,國立台灣科技大學電機工程研究所,民國 84 年.

[15] K Levenberg, “A method for the solution of certain problems in least sqrares,” *Quart. Applied Math*., vol. 2, pp. 164-168, 1944.

[16] D. Marquardt, “A algorithm for least-squares estimation of non-linear parameters,” *SIAM J. Applied Math*, vol. 11, pp.431-441, 1963.