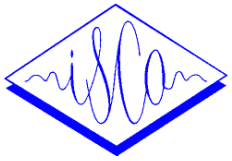
ISCA Archive
http://www.isca-speech.org/archive



7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

Phonetic Normalization Using *z*-Score in Segmental Prosody Estimation for Corpus-based TTS System

Hoeun Song, Jaein Kim*, Kyongrok Lee**, Jinyoung Kim***

* Spoken Language Research Division, Multimedia Technology Laboratory
Research & Development Group, Korea Telecom
**Multimedia DSP Lab., Dept. of Electronic Engineering, Chonnam National University, Gwang-ju, Korea

{*hoeun*, *jaeinkim*}@kt.co.kr, {*kimjin*, *krlee* }@dsp.chonnam.ac.kr

Abstract

Recently, corpus-based text-to-speech (CB-TTS) has been actively studied through the world. Statistical training methods are generally applied for prosodic rules in CB-TTS, and classification and regression tree (CART) is one of the mostly used methods. In this paper, we present an efficient CART training approach of *z*-score based phonetic normalization. The idea of ours comes from the fact that the most important three parameters of CART training for segmental prosody are phone and its right and left phones, especially in Korean language. Our approach reduces the number of CART terminal nodes effectively. The reduction ratios are approximately 14-94% for estimation of segmental duration and 45-70% for intensity estimation. Also, the experimental results show that phonetic normalization slightly lessens the estimation errors.

1. Introduction

The growing popularity of speech synthesizer enabling comfortable man-machine-interfaces demands high quality of the synthesized speech. Corpus-based speech synthesis approach has become one of the most attractive synthesis methods since it guaranteed high perceptual quality with good naturalness [1][2][3]. This high speech quality of the corpus-based synthesizer satisfies the needs of commercial product on the contrary of rule-based approach.

Recently, we have developed our own CB-TTS system[4]. That synthesizer consists mainly of phonetic language processing (PLP), prosody estimation (PE) and unit concatenation (UC) modules. PLP module performs three tasks. After the preprocessing such as deleting special letters and separating sentences, morphological analysis on input sentences is performed in the first step. The next is tagging which generates POS(part of speech) tags, grammatical relations of each word and structures of sentences. At the last, the PLP module performs grapheme-to-phoneme translation of input sentences and gives information of POS, phonetic symbols and structures to prosody estimation module. After the process of PLP, prosodic information on word or phone level is estimated. This paper focuses on the prosody module, especially, estimation of segmental duration and intensity of phone strings. Finally, the synthesizer generates speech by concatenating the desired speech units extracted from our large database containing real speech and related information about the contained segmental units.

For any TTS (text-to-speech) system to have natural and good articulate sound, estimation of prosodic features such as pitch, duration and intensity is required on segmental and supra-segmental level and plays a more important role in the system than other modules. There are several trainable methods of estimating prosodic elements like NN, HMM and CART, *etc*. In our system we used CART to estimate prosodic information. CART is a kind of statistical procedure, which automatically searches for important relationships and uncovers hidden structure even in highly complex data. According to previous research of CART-based prosody training, no preprocessing is done to target values of segmental prosodic values. On the other hand, from our experimental results of prosody estimation it is observed that the most important three parameters are phone, left and right phones in phone sequences. From this observation, we conclude that if we can phonetically normalize prosodic values, the complexity of CART tree will be reduced. In this paper we adopt *z*-score normalization as a method of phonetic normalization. In the rest of this paper we will show that our proposed method achieves the CART tree with lower complexity and also reduces estimation errors slightly.

2. CART-based estimation of segmental prosody information

As described above, we apply CART-based training method to obtain prosodic rules, especially for segmental duration and intensity estimation. Basically Korean syllable is composed of 3-phonemes of initial consonant, vowel and final consonant. We also considered two kinds of break index, which are accentual and intonational phrases. So we classify segmental units into 5 classes as follows.

(1) NB-IC : initial consonant without break
(2) NB-VO : vowel without break
(3) NB-FC : final consonant without break
(4) AP-BP : boundary phone of accentual phrase
(5) IP-BP : boundary phone of intonational phrase.

In the following two sections we explain the preliminary results of CART training for segmental duration and intensity. In our experiments the used prosody DB consisted of 1,000 sentences (11,028 words called eojeol in Korean, 72,447 phonemes), which were selected by the phonetician for having various prosody in the sentences if possible.

2.1 Segmental duration estimation

Table 1. Input parameters for CART training in segmental duration estimation.

Variable	Meaning
DlPhon	Left phone of observation phone
Dphon	Observation phone
DrPhon	Right phone of observation phone
DlPOS	Left word's POS
DPOS	Word's POS
DrPOS	Right word's POS
LocPhon	Location in word (initial, middle and final)
LocAccPhr	Location in accentual phrase (initial, middle and final)
LocIntPhr	Location in intonational phrase (initial, middle and final)
CnumSylE	Number of syllables in word
CnumSylP	Number of syllables in accentual phrase
LocWord	Location of word in accentual phrase (initial, middle and final)

In CART training for segmental duration we used input parameters as shown in table 1. In the table the parameters except CnumSylE and CnumSylP are categorical variables. We used Gini index and root mean square error (RMSE) as a measure of node impurity, and Chou algorithm as node splitting method. Also we adopted 10-fold cross validation. The table 2 shows the optimal terminal node number and RMSE values. From the table we can observe that the size of the optimal tree is not small for NB-IC, NB-VO and NB-FC classes. Then the program code becomes complex and it needs high calculation amount. So, we need to reduce the complexity of the generated CART trees with keeping RMSE values. How can we achieve that? We can get an idea from the fact that CART reveals the structure of parameter space of training data. That is, CART tool reports the important values of each parameter. The table 3 shows the important values of input parameters for NB-IC, NB-VO and NB-FC classes.

According to experimental results of the table 3, the parameters of DPhon, DlPhon and DrPhon are very important in estimating

Table 2. CART tree result in segmental duration estimation.

Phone class	Number of nodes	RMSE value
NB-IC	147	13.3msec
NB-VO	159	19.2msec
NB-FC	42	15.6msec
AP-BP	19	22.8msec
IP-BP	7	27.9msec

Table 3. The important values for input parameters in duration estimation.

NB-IC	NB-VO	NB-FC
Dphon : 100.0 DlPhon : 57.1 LocPhon : 17.18 DPOS : 12.1 LocWord : 11.48 etc.	DPhon : 100.0 DrRphon : 81.64 DlPhon : 61.4 LocAccPhr : 16.38 …….. etc.	DrPhon :100.0 Dphon : 59.5 LocWord : 10.5 DPOS : 9.54 …….. etc.

Table 4. CART training result in segmental duration estimation.

Phone class	Number of nodes	RMSE value
IC	120	28.76%
VO	131	17.47%
FC	55	21.60%

Table 5. The importance values for the input parameters.

IC	VO	FC
DPhon : 100.0 DlPhon : 12.9 LocSP : 17.18 DPOS : 12.1 …….. …….. *etc.*	F0 : 100.0 LocSyll : 81.44 LocSP : 72.52 …… DlPhon 55.72 Drphon 44. *etc.*	DPhon : 100.0 DrPhon : 61.5 LocSen :16.75 DRPOS 16.09 …….. *etc.*

segmental duration. In other words, the phonetic conditions have key roles in the determination of segmental duration. So, if we can devise phonetic normalization method, it is possible to reduce the CART tree complexity.

2.2 Segmental intensity estimation

We applied similar approach as discussed above for segmental intensity estimation. We changed some input parameters in intensity estimation. The parameters related pitch and relative position were added. Pitch related parameters are lF0, F0, rF0 that mean pitch values of phone, left and right phones. Of course, the pitch values are obtained by the pitch estimation rules. And the parameters related position were LocSP and LocSen. LocSP is relative position in intonational phrase and LocSen is relative location in a given sentence. The table 4 shows the estimation results. Like the case of duration estimation, the CART tree has many terminal nodes. The important analysis results are shown in the table 5. The important values of DPhon, DlPhon and DrPhon are different from the case of the segmental duration estimation. However, these parameters are still important in the intensity estimation. Thus we have to devise a kind of phonetic normalization to get a downsized CART tree.

From the above two experiments it was shown that the phonetic environment of DlPhon and DrPhon are important as well as segmental phone itself in prosody estimation problem. This tells us that the phonetic characteristics can be described as tri-phone. That is,

Tri-phone = (Left Phone)-(Phone)+(right Phone).

From this observation it is known that the phonetic normalization can be possible if we make CART trees for each tri-phone unit. But, this method is not appropriate. Although each CART tree for each trip-phone has low complexity, the total number of CART trees is very large. It is equal to the number of tri-phone units.

3. Phonetic normalization and its application

3.1 Phonetic normalization

In the above sections we discussed our preliminary experimental results on segmental prosody estimation. From the preliminary results we can deduce that the phonetic environments are very important parameters in segmental prosody estimation. Thus, if we can normalize segmental prosody phonetically, we can reduce CART tree complexity. In this problem the normalization means the elimination of phonetic influences from segmental prosody values.

In this paper we suggest a normalization approach of *z*-score. *Z*-scores are a special application of the transformation rules in statistics. The *z*-score for an item indicates how far and in what direction, that item deviates from its distribution's mean expressed in units of its distribution's standard deviation. The mathematics of the *z*-score transformation is such that if every item in a distribution is converted to its *z*-score, the transformed scores will necessarily have a mean of zero and a standard deviation of one. Thus the application of *z*-score makes the all the prosody values for each segmental unit have standard Gaussian distributions under the assumption that segmental prosodic values have Gaussian pdf. This procedure has the effect of phonetic normalization, for we apply z-score approach for each tri-phone units.

Based on this concept we propose a training model of prosody model as shown in figure 1. In the figure ① means conventional approach and ② means our proposed approach. As discussed above, the *z*-scores, means and standard variations of prosody values are obtained for each tri-phone unit. So the proposed algorithm is as follows.

1) Perform z-score normalization to each triphone using eq. (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Store all the mean and standard variation of each triphone $\{\mu_i, \sigma_i\}$ in this step.

2) Make CART tree for segmental prosody estimation using z-score values.

As we normalize prosody values with z-score approach, the de-normalization process should be performed after CART tree encoding. That is, if the segment belongs to *i*-th tri-phone, the estimated value *x* is transformed using eq. (2).

$$y = \sigma_i x + \mu_i \tag{2}$$

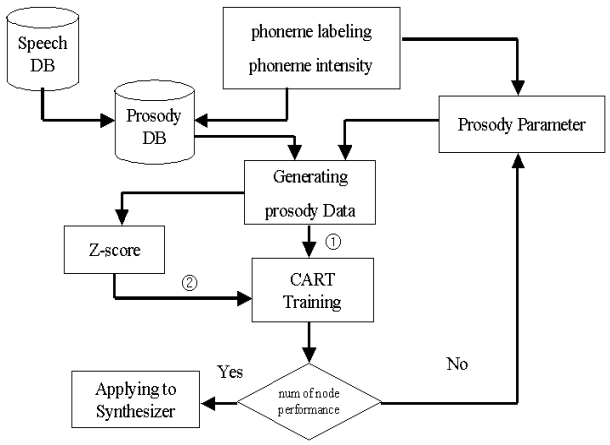


Figure 1. Training Process of Prosody model

Table 6. CART tree result in segmental duration estimation.

Phone class	Number of nodes	RMSE value
NB-IC	55	12.6msec
NB-VO	10	15.6msec
NB-FC	12	14.7msec
AP-BP	9	20.2msec
IP-BP	6	24.5msec

Table 7. CART tree result in segmental intensity estimation.

Phone class	Number of nodes	Relative RMSE value
IC	37	27.5%
VO	41	15.7%
FC	30	20.9%

in the following section of 3.2 we will explain the result of the z-score based segmental prosody estimation.

3.2 Segmental prosody estimation results using the phonetic normalization

In this section, we present the experimental results of segmental prosody estimation using the *z*-score based phonetic normalization. The table 6 shows the results of segmental duration estimation and the table 7 shows the segmental intensity estimation results. From these tables we can observe that the performances of segmental prosody estimation are enhanced. The number of terminal nodes is highly lessened. Also, the estimation errors are slightly reduced.

To directly compare our proposed approach with the conventional method, we show the reduction ratios of terminal nodes in figure 2 and 3. In the figure 2, the reduction ratios are 14-94% for duration estimation. Especially, the 94% reduction of terminal nodes is achieved in the case of non-break vowel. This means that the phonetic normalization is very effective for the vowel class. For the case of IP-BP (boundary phone of intonational phrase) class, the reduction ratio is 14%. But, this is not a problem, for the number of terminal nodes is originally small in that case. The total reduction ratio is 70% and the average reduction ratio is 58.4%. Here the total reduction ratio (TRR) is calculated as follows.

$$TRR = \frac{NC - NP}{NC} \cdot 100\%, \tag{3}$$

Where *NC* is the total terminal nodes of the conventional method and *NP* is the total terminal nodes of the proposed method. And the average reduction ratio is the mean of reduction values of each phone class.

On the other hand, from the figure 3 we observe that the terminal node reduction of 45-70% is achieved in intensity estimation. The total reduction ratio is 64.7% and the average reduction ration is 60.7%. The experimental results prove that our proposed phonetic normalization is very useful in CART-based segmental prosody estimation.

4. Conclusion

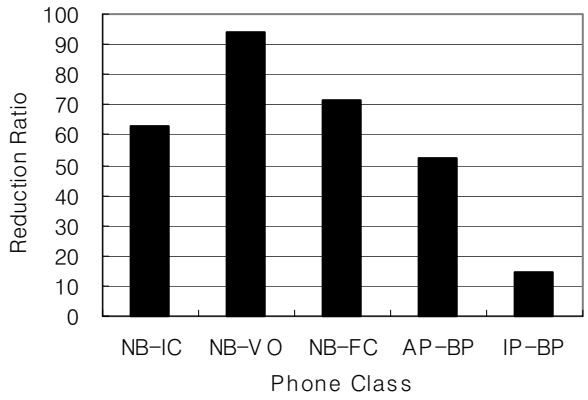


Figure 2. Reduction ratio of terminal nodes in segmental duration estimation.

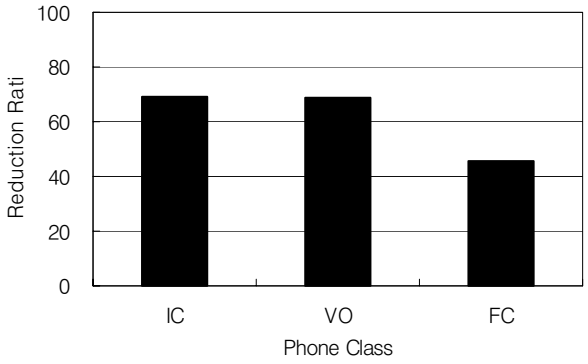


Figure 3. Reduction ratio of terminal nodes in segmental intensity estimation.

For commercial synthesizer, synthesized speech should be closer to human sound and the performance like processing speed or DB size *etc.* have to be also better as much as possible. Also, it is desirable that the implemented TTS has small–sized engine.

In this paper, we introduced phonetic normalization approach by using *z*-score to get a downsized CART tree in prosody estimation. From Figure 2 and Figure 3, we can see the result of estimating segmental duration and intensity. From these figures it is concluded that z-score normalization approach is better than the conventional method.

It means that the synthesizer can obtain better quality of speech and run more rapidly from searching for a smaller tree. Actually, an element affecting speed of searching is the depths of tree rather than the number of terminal nodes. However, because CART creates trees having almost uniform number whether the largest depths or the smallest, we can imply the depths is proportional to the number of terminal node. In addition, we can catch one more fact from the experimental results. In *z*-score based method, the accuracy rate has been slightly improved, although the number of terminal nodes is lessened highly. For example 14-94% reduction is achieved in duration estimation. This implies that we can prune optimal tree as small as possible, if not so much worse. In the future, we need a study about searching minimum number of tree as an error of estimating prosody modules increases not so much by trial and error.

Our future tasks, therefore, should be investigated a relationship with prosody, parameters, phonetic classification methods like locations and styles of intonation of speech for improvement of naturalness and articulation. Finally, we will study whether the method can be combined with logarithmic transformation, which could assure better results from perceptually point of view.

Reference

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," Joint Meeting of ASA, EAA, and DAGA, pp.18-24, 1998.

[2] H. Hong, A. Acero, X. Huang, J. Liu and M. Plumpe, "Automatic generation synthesis units for trainable text-to-speech systems," Proceedings of ICASSP'98, pp.293-296, 1998.

[3] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proceedings of ICASSP'96, vol. 1, pp.373-376, 1996.

[4] A. Ferencz, S. Choi, H. Song, M. Koo, "Hansori 2001 – Corpus-based Implementation of the Korean Hansori Text-to-Speech Synthesizer," Proceedings of EUROSPEECH2001, pp. 841-844, 2001.

[4] L. Breiman, J. H. Friedman, J. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984.

[5] D. Steinberg and P. Colla, *CART manual*, Salford Systems, 1997.

[6] Jan P.H. Van Santen, Richard W. Sproat, Joseph P. Olive, Julia Hirschberg, *Progress in Speech Synthesis*, pages 278 ~ 292, Springer, 1996

[7] Sangho Lee, *Tree-based modeling of Prosody for Korean TTS Sytstems,* doctorate thesis, 1999.

[8] Y. K. Hong et al., *A Korean Morphological Analyzer for Speech Translation System*, Proceedings of SCSP93, 1993 (in Korean).

[9] Ji-Young Shin, *Understanding of Speech Sound*, Hanguk-Munhwa Sa, 2000