Sums-of-Products Models for Korean Segment Duration Prediction[*]

Hyunsong Chung[**]

ABSTRACT

Sums-of-Products models were built for segment duration prediction of spoken Korean. An experiment for the modelling was carried out to apply the results to Korean text-to-speech synthesis systems. 670 read sentences were analyzed, trained and tested for the construction of the duration models. Traditional sequential rule systems were extended to simple additive, multiplicative and additive-multiplicative models based on Sums-of-Products modelling. The parameters used in the modelling include the properties of the target segment and its neighbors and the target segment's position in the prosodic structure. Two optimisation strategies were used: the downhill simplex method and the simulated annealing method. The performance of the models was measured by the correlation coefficient and the root mean squared prediction error (RMSE) between actual and predicted duration in the test data. The best performance was obtained when the data was trained and tested by "additive-multiplicative models." The correlation for the vowel duration prediction was 0.69 and the RMSE, 31.80 ms, while the correlation for the consonant duration prediction was 0.54 and the RMSE, 29.02 ms. The results were not good enough to be applied to the real-time text-to-speech systems. Further investigation of feature interactions is required for the better performance of the Sums-of-Products models.

Keywords: duration, prosody, Sums-of-Products (SoP) model, CART, speech synthesis

1. Introduction

One of the challenges for research in experimental phonetics in the new millennium is to contribute to both better synthesis systems and a better understanding of human

---

[*] This research was supported in part by the Daegu University Research Grant, 2002.

[**] Department of English Education, Daegu University

speech production. We see this in the research on prosodic phrasing, intonation modelling, and duration modelling that are actively under way in many languages. One particular area of concern is in the analysis and modelling of the prosody of Korean, particularly in the area of segmental durations. Extending the result of the CART duration modelling (Chung, 2001), this paper tries to investigate the possibility of using Sums-of-Products (SoP) models in the duration prediction for speech synthesis systems. This study is based on previous work where possible, but extended to take into account the demands of contemporary approaches to duration modelling as used in English and Japanese synthesis. However, this paper does not just try to build the best predictive model of segment duration in context. It also seeks to learn more about which factors and which structures are most important in Korean prosody. Besides the result from the CART model, the outcome of this work is expected to be both a better model of Korean timing for use in synthesis, and a better understanding of the Korean language. In the following sections, the SoP models are applied to the same data as used in the CART model and the performance results of the two models are compared. The objective quality of the modelling is evaluated by root mean squared prediction error (RMSE) and the correlation coefficient between actual and predicted duration in reserved test data.

2. Overview of Sums-of-Products Model

2.1 Modelling Segment Duration

The underlying linguistic representation in synthesis is symbolic, consisting of entities such as phoneme sequences, in combination with morphological, syntactic and prosodic information. The prosody prediction component computes the timing and pitch contour for the phrase. Prosody modelling refers to the equations involved in these computations, that is using the phonological structure to predict pitch and timing values in Hertz and milliseconds. Prosody modelling is one of the most important factors in determining the naturalness of synthesised speech (Horne, 2000).

Following van Santen (1992), Campbell (2000) categorises current duration prediction systems into three classes: sequential rule systems, equation systems, and binary prediction trees. Such rule systems as Klatt's (1987) are considered sequential rule systems which could be easily converted to equation systems such as van

Santen's SoP models (1992). CART models (Classification and Regression Tree; Breiman et al., 1984; Riley, 1992) are considered binary prediction tree systems which have been criticised by van Santen (1992) as just a collapsed form of lookup table. What these systems share is that they map symbolic input vectors provided by linguistic analysis routines onto acoustic quantities (duration), which may then used by the synthesis component to generate speech with the desired acoustic-prosodic characteristics. SoP models were chosen in this paper because they are believed to be widely used and representative of current duration prediction systems and because linguistic interpretation of their operation is possible.

2.2 Sums-of-Products Model

Van Santen (1992) argued that previous studies on contextual effects on segmental duration have focussed more on theoretical issues and putative underlying processes rather than completeness of empirical description. He said that the first step to construct a duration rule system for a text-to-speech (TTS) system is to make a list of factors which describe the contexts of a segment. The second step is to produce a duration model to explain complex interactions. In his TTS duration model, van Santen tries to show the durational behaviour of a single speaker and produce a simple equation to predict the durations based on contextual factors. Van Santen (1992) also said that duration databases for statistical analysis commonly confound factors in that not all combinations of factors and levels occur with equal frequency. According to him, the factor confounding results in mean durations that correspond to the levels of the factor of interest (the critical factor) being affected by other factors (confounding factors). One such example is word-final lengthening of unstressed syllables. Because, in word-final position, vowels in English are more likely to be unstressed and stressed vowels are more likely to be longer than unstressed vowels, statistics show that word-final vowels are shorter than non-word final vowels when all vowels are analysed altogether. However, when stressed and non-stressed vowels are analysed separately, word-final vowels are longer than non-word final vowels. When a pair of interacting factors such as the vowel and stress factors need to be described, the quasi-minimal pairs technique can be used. Segment durations occurring with a combination of levels on confounding factors and with several levels on the critical factor are divided into "quasi-minimal" sets. If there are not enough duration events for all sets, a piecewise multiplicative correction method can be introduced which assumes that the effect of the critical factor and the joint effects of the remaining factors combine multiplicatively. Van Santen gives the example of the interaction between the syllabic position factor and the stress factor. He argued that these interactions are better described by a multiplicative rule than an additive rule. However, such interactions are not necessarily completely multiplicative, so he uses the term 'piecewise'. Where the quasi-minimal sets and multiplicative correction methods have difficulties with factors that have many levels, he introduces SoP models, which he calls "a special case of an additive-multiplicative models, consisting of the sum of a single product term and any number of single-factor terms."

According to SoP models, the duration for a unit in the context combination described by the feature vector d is given by:

$$\mathrm{DUR}(\mathbf{d}) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j)$$

Here, K is a set of indices, each corresponding to a *product term*. I_i is the set of indices of factors occurring in the i-th product term. Van Santen suggested that major interactions between factors could be described as a complete multiplicative rule (a single product term) or a piecewise multiplicative rule (more than two terms) in a "SoP" model. Otherwise, other interactions are described as additive in the model. The multiplicative interactions predict constancy when effect size is measured as a percentage, the additive interactions do it when it is measured in milliseconds.

In one of his SoP models of English (1994), he provided the following performance result.

Table 1. Performance result of duration modelling in van Santen's (1994) SoP model.

Category	Correlation
Vowels	0.908
Intervocalic consonants	0.903
Consonants in clusters	0.874

In their modelling of segmental durations for Japanese TTS synthesis, Venditti and van Santen (1998) used the following factors:

(1) current phone identity (id), 1 level each segment

(2) length of current vowel (leng): phonological vowel length

(3) preceding phone identity (prev): voiceless stop, voiceless fricative/affricate, voiced stop, voiced fricative/affricate, flap, nasal, glide, vowel

(4) following phone identity (foll): voiceless stop, voiceless fricative/affricate, voiced stop, voiced fricative/affricate, flap, nasal, glide, vowel

(5) left prosodic context (left_pos): major phrase−initial, minor phrase−initial, intonation phrase−initial, accentual phrase−initial, non−initial

(6) right prosodic context (right_pos): major phrase−final, minor phrase−final, intonation phrase−final, accentual phrase−final, non−final

(7) accent status (acc): accented, downstep accented, preceding an accent in an accented AP, following an accent in an accented AP, in an unaccented AP

(8) syllable structure (syll): open or closed by a geminate or moraic consonant

(9) special morpheme status (spec)

The performance results of their models are as follows.

Table 2. Results of SoP model for Japanese vowels in Venditti and van Santen (1998).

Vowel category	Correlation	RMSE
Non−init/non−final	0.87	10.7 ms
Final	0.88	9.3 ms
Initial	0.85	14.9 ms
All vowels	0.88	16.8 ms

Table 3. Results of SoP model for Japanese consonants in Venditti and van Santen (1998).

Consonant category	Correlation	RMSE
CV non−init (bursts)	0.79	5.0 ms
CV non−init (others)	0.93	7.9 ms
CV initial (all)	0.96	8.1 ms
CyV non−init (all)	0.93	8.8 ms
CyV initial (all)	0.89	10.8 ms
CyV /y/	0.90	3.6 ms
Moraic N non−final	0.54	13.2 ms
Moraic N final	0.90	7.5 ms
All consonants	0.94	12.5 ms

The impressive performance result of this experiment is probably because vowels and consonants were sub−divided into so many groups rather than being modelled together.

3. Design of Corpus

3.1 Material of Main Corpus

News scripts from two main Korean broadcasting stations were recorded by one male speaker of modern standard Korean. The 670 sentences were chosen after the removal of speech errors by the speaker and those utterances which were incomplete or which seemed less grammatical. The sentences were divided into two data sets: 80% went into the training data set (42,103 segments in 535 sentences), while 20% went into the test data set (10,737 segments in 135 sentences).

3.2 Generation of Training and Test Data for Modelling

For the modelling process, a feature string for each segment was automatically generated from the phonological structure using the ProXML scripting language (Huckvale, 1999). The script looked at each segment in turn and constructed an n−ary feature string from the properties of the target segment, the properties of its neighbours and its position in the prosodic structure. Each segment was annotated with the following features together with the actual duration:

(1) Phonemic identity of the target segment, e.g. segment name, or phonemic features of the target segment, i.e. major class features of the segment
(2) Phonemic features of the preceding and the following segments
(3) Syllable structure: position and structure of containing syllable
(4) Position of syllables in UTT (utterance), IP (intonational phrase), AP (accent phrase) and PW (phonological word)

The detailed descriptions of the feature sets can be found in the following tables.

Table 4. Feature set for vowels.

Features	Sub-Levels	Description
id: inherent property of the target segment	one for each segment	
man: manner of the target segment	mono	monophthong
	di	diphthong
prev: the property of the preceding segment	vow	vowel
	nas	nasal
	lat	lateral
	fla	flap
	aspstp	aspirated stop
	aspaff	aspirated affricate
	tnsstp	tense stop
	tnsaff	tense affricate
	tnsfri	tense fricative
	laxstp	lax stop
	laxaff	lax affricate
	laxfri	lax fricative
	pause	pause
foll: the property of the following segment	vow	vowel
	nas	nasal
	lat	lateral
	fla	flap
	aspstp	aspirated stop
	aspaff	aspirated affricate
	tnsstp	tense stop
	tnsaff	tense affricate
	tnsfri	tense fricative
	laxstp	lax stop
	laxaff	lax affricate
	laxfri	lax fricative
	pause	pause
syll: syllable structure	cv	
	cvc	
	v	
	vc	
left_pos: the distance from the syllable to the left edge of the phrase	utt−init	UTT initial
	ip−init	IP initial
	ap−init	AP initial
	pw−init	PW initial
	non−init	Non−initial
right_pos: the distance from the syllable to the right edge of the phrase	utt−fi	UTT initial
	ip−fi	IP initial
	ap−fi	AP initial
	pw−fi	PW initial
	non−fi	Non−initial

Table 5. Additional feature set for consonants.

Features	Sub-levels	Description
id: inherent property of the target segment	one for each segment	
man: the property of the target segment segment	stop aff fri nas lat fla	stop affricate fricative nasal lateral flap
syllpo: segment position in the syllable	on co	onset coda

Seven features for vowels were used in this paper, each of which has a number of sub-levels. For consonants, features "id" and "man" were modified accordingly and a new syllable position feature "syllpo" was added.

4. Analysis of Corpus

Simple additive models, multiplicative models, and additive-multiplicative models are built and evaluated from the Korean corpus. The additive models, multiplicative models, and additive-multiplicative models are described under the headings of SoP models. Using the results of CART modelling (Chung, 2001), SoP models are built and explored.

Separate SoP models were built for vowels and consonants because of differences in the feature sets used to describe them. Firstly, the model formula was specified in terms of which factors were to be incorporated and how the parameters associated with each factor were to be combined in the model. Each parameter was then initialised to a value specified by the experimenter. Limits on the allowed range of values were established. A function optimisation strategy was then employed whereby perturbations in the values of the parameters were investigated in terms of their effects on the model performance. An objective measure of sum squared error of prediction was used for this. Two optimisation strategies were used: the downhill simplex method (Press et al., 1992) which seeks to find a single value for the parameters which minimised the objective measure; and the simulated annealing method (Press et al., 1992) which has additional benefits in its ability to avoid sub-optimal solutions. Models were trained on 19,071 vowels and 23,032 consonants in the training data and tested on 4,829 vowels and 5,908 consonants in the test data. The data were exactly the same as used in the CART modelling.

4.1 Additive Models

This is the simplest SoP model where there is one parameter per factor level and all parameters are added together based on the assumption that features operate in an additive manner to predict the duration. The additive model for vowels is shown below.

Model 1 for vowels: "pure additive model"

$$DUR(id, man, prev, foll, syll, left\_pos, right\_pos) = S_{1,1}(id)+S_{2,1}(man)+S_{3,1}(prev)+S_{4,1}(foll)+S_{5,1}(syll)+S_{6,1}(left\_pos)+S_{7,1}(right\_pos)$$

The correlation of the model trained by the downhill simplex method was 0.58 and the RMSE was 39.69 ms. The correlation using the simulated annealing method was 0.61 and the RMSE was 36.89 ms. The additive model for consonants was formulated as:

Model 1 for consonants: "pure additive model"

$$DUR(id, man, prev, foll, syll, syllpo, left\_pos, right\_pos) = S_{1,1}(id)+S_{2,1}(man)+S_{3,1}(prev)+S_{4,1}(foll)+S_{5,1}(syll)+S_{6,1}(syllpo)+S_{7,1}(left\_pos)+S_{8,1}(right\_pos)$$

All features described in the data were added together to predict the duration. The correlation by the downhill simplex method was 0.54 and the RMSE was 29.29 ms. The simulated annealing method produced a correlation of 0.51 and an RMSE of 30.08 ms.

The results from "pure additive models" reflect that they were not good enough in predicting durations. They failed to predict that longer durations could be found in some situations.

4.2 Multiplicative Models

The next simplest SoP model is one in which all factors combine multiplicatively. This is equivalent to an additive model working in the log domain. The model for vowels is:

Model 2 for vowels: "pure multiplicative model"

$$DUR(id, man, prev, foll, syll, left\_pos, right\_pos) =$$
$$S_{1,1}(id)\times S_{1,2}(man)\times S_{1,3}(prev)\times S_{1,4}(foll)\times S_{1,5}(syll)\times S_{1,6}(left\_pos)\times S_{1,7}(right\_pos)$$

The multiplicative model for vowels trained by the downhill simplex method gave a correlation of 0.49 and an RMSE of 48.81 ms. The simulated annealing method gave a correlation of 0.51 and an RMSE of 44.09 ms. These results are significantly worse than the additive model for vowels.

The full "purely multiplicative model" for consonants is:

Model 2 for consonants: "pure multiplicative model"

$$DUR(id, man, prev, foll, syll, syllpo, left\_pos, right\_pos) =$$
$$S_{1,1}(id)\times S_{1,2}(man)\times S_{1,3}(prev)\times S_{1,4}(foll)\times S_{1,5}(syll)\times S_{1,6}(syllpo)\times S_{1,7}(left\_pos)\times$$
$$S_{1,8}(right\_pos)$$

The correlation with fitting by the downhill simplex method was 0.14 and the RMSE was 51.42 ms, which was the poorest result of all models. The simulated annealing method gave a correlation of 0.49 and an RMSE of 31.32 ms. The results from "pure multiplicative models" show that they were slightly better at predicting some of the longer durations found in the data, but still they are not very good either in this respect or overall.

4.3 Additive-multiplicative Models

As van Santen (1997) suggested, models need to be elaborate enough to capture systematic variability in the data, yet use few parameters. He claims that additive-multiplicative models capture the "directional invariance" of the segment duration. He said "directional invariance is the property that, holding all else constant, the effects of a factor have always the same direction." We have not attempted to derive a new additive-multiplicative model for Korean but instead we have adapted Venditti and van

Santen's (1998) model of Japanese timing. Because feature classes of this model are different to those we have used for Korean, they were adjusted as in Model 3. Venditti and van Santen (1998) suggested the following SoP model for Japanese vowels:

DUR(id, man, prev, foll, left_pos, right_pos, acc, syll, spec) =
$S_{1,1}$(id) + [$S_{2,1}$(leng)×$S_{2,2}$(prev)] + [$S_{3,1}$(leng)×$S_{3,2}$(foll)] +$S_{4,1}$(left_pos) + [$S_{5,1}$(leng)×
$S_{5,2}$(right_pos)] + [$S_{6,1}$(leng)×$S_{6,2}$(acc)] + [$S_{7,1}$(foll)×$S_{7,2}$(syll)] +$S_{8,1}$(spec)]

where 'id' is the identity of the vowel, 'leng' is the length, 'prev' the manner of the preceding vowel, 'foll' the manner of the following vowel, 'left_pos' the syllable distance to the left phrase boundary, "right_pos" the syllable distance to the right phrase boundary, "acc" the accent status of the syllable, "syll" the syllable structure, "spec" the special morphological information. Because the "acc" and "spec" features are not available in the Korean data, these features were ignored. The "leng" feature was substituted by "man" (manner) in the modified model, because the phonemic length is not assumed to exist in the modern Korean. Other than the phonemic length, the manner feature of the target segment is believed to be the best candidate to describe the property of the segment. The modified model for the Korean language is thus:

Model 3 for vowels: "additive–multiplicative model"
DUR(id, man, prev, foll, syll, left_pos, right_pos) =
$S_{1,1}$(id) + [$S_{2,1}$(man)×$S_{2,2}$(prev)] + [$S_{3,1}$(man)×$S_{3,2}$(foll)] + [$S_{4,1}$(foll)×$S_{4,2}$(syll)] +
$S_{5,1}$(left_pos) +$S_{6,1}$(right_pos)

The correlation of this model by the downhill simplex method was 0.69 and the RMSE was 31.80 ms. The correlation by the simulated annealing method was 0.68 and the RMSE was 32.13 ms. These are the best results for SoP model for vowels.

To obtain an additive–multiplicative model for consonants, we adapted the model used for vowels, having no specific information which might guide an alternative design. The model is:

Model 3 for consonants: "additive–multiplicative model"

$$\mathrm{DUR}(\mathrm{id},\ \mathrm{man},\ \mathrm{prev},\ \mathrm{foll},\ \mathrm{syll},\ \mathrm{syllpo},\ \mathrm{left\_pos},\ \mathrm{right\_pos}) =$$

$$S_{1,1}(\mathrm{id}) + [S_{2,1}(\mathrm{man}) \times S_{2,2}(\mathrm{prev})] + [S_{3,1}(\mathrm{man}) \times S_{3,2}(\mathrm{foll})] + S_{4,1}(\mathrm{syll}) + S_{5,1}(\mathrm{syllpo}) +$$

$$S_{6,1}(\mathrm{man}) \times S_{6,2}(\mathrm{left\_pos})] + [S_{7,1}(\mathrm{man}) \times S_{7,2}(\mathrm{right\_pos})]$$

where 'syllpo' is the segment position in the syllable, i.e. onset or coda. The correlation was 0.54 both by the downhill simplex method and simulated annealing method. The RMSE was 29.02 ms by the downhill simplex method and 28.86 ms by the simulated annealing method. These results were the best among the SoP models for consonants.

The additive–multiplicative models still failed to predict longer durations and the performance was not significantly improved

4.4 Summary

A summary of the performance results of all of the above models are illustrated as follows.

Table 6. Performance results summary for vowels using SoP models.

	Downhill simplex		Simulated annealing	
	RMSE	Correlation	RMSE	Correlation
Model 1	39.69 ms	0.58	36.89 ms	0.61
Model 2	48.81 ms	0.49	44.09 ms	0.51
Model 3	31.80 ms	0.69	32.13 ms	0.68

Table 7. Performance results summary for consonants using SoP models.

	Downhill simplex		Simulated annealing	
	RMSE	Correlation	RMSE	Correlation
Model 1	29.29 ms	0.54	30.08 ms	0.51
Model 2	51.42 ms	0.14	31.32 ms	0.49
Model 3	29.02 ms	0.54	28.86 ms	0.54

The results showed that the CART decision tree models had overall better performance than the SoP models. The previous results using CART models (Chung, 2001) are summarized as follows:

Table 8. CART performance results for vowels and consonants.

	Vowels		Consonants	
	RMSE	Correlation	RMSE	Correlation
Name & manner	27.51 ms	0.78	24.20 ms	0.71
Name only	27.68 ms	0.78	24.56 ms	0.70
Manner only	28.50 ms	0.76	26.76 ms	0.63
z-score	26.01 ms	0.77	25.21 ms	0.70

Performance was best when the segment names and manner features were used in the CART model. Among SoP models, the additive-multiplicative models were better than the "pure additive models" or the "pure multiplicative models." This might be evidence for presence of interactions between factors that could not be modelled by uniform addition or multiplication. Further investigation of these interactions is clearly required.

5. Conclusion

This paper showed that SoP models and CART analysis can be used to address linguistic issues such as the boundary effect, the syllable structure effect, and the effect of surrounding segments on the segment duration. In addition, these models also provide duration predictions which can be used in the prosody component of TTS systems. Based on the results of the CART models, this paper tried to show that contextual interactions could be represented in the SoP duration models. However, the results from SoP models for Korean in this paper were worse than those for English or Japanese found in other studies. They also failed to show the systematic interactions among contextual features. The best performance results from CART models were similar to other studies on Korean segment duration (Lee, 1996; Lee & Oh 1999). Future studies might examine the internals of the CART decision trees to find interactions among features, and this might feed into better SoP models. The calculation of the values for the same feature in different product terms should be pursued to improve the performance of SoP models.

References

Breiman, L., J. Friedman, R. Olshen & C. Stone. 1984. *Classification and Regression Trees*, Monterey, Chapman & Hall/CRC.

Campbell, W. Nick. 2000. Timing in speech: a multi-level process, in *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce,* Merle Horne (ed.), Dordrecht, the Netherlands: Kluwer Academic Publisher, 281-334.

Chung, H. 2001. "Analysis of the timing of spoken Korean using a Classification and Regression Tree (CART) model." *Korean Journal of Speech Sciences*, 8(1), 77-91.

Horne, Merle (ed.). 2000. *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce*, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Huckvale, Mark. 1999. "Representation and processing of linguistic structures for an all-prosodic synthesis system using XML." *Proceedings of Eurospeech '99*, 4, 1847-1850.

Klatt, D. H., 1987. Review of text-to-speech conversion for English, *Journal of Acoustical Society of America*, vol. 82 (3), 737-793.

Lee, Sangho & Oh. Young-Hwan. 1999. CART-based modelling of Korean segmental duration, *Proceedings of the Second International Workshop on East-Asian Language Resources and Evaluation*, Oriental COCOSDA '99, 109-112.

Lee, Yang Hee. 1996. Modelling of segment duration in Korean speech synthesis, *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee.* Seoul National University Press, 249-274.

Press, William, H., Saul, A. Teukolsky, William, T., Vetterling, and Flannery. Brian P. 1992. *Numerical Recipes in C*, 2nd edition, Cambridge: Cambridge University Press.

Riley, M. 1992. Tree-based modelling of segmental durations, in *Talking Machines. Theories, Models and Designs*, G. Bailly, C. Benôit (eds.), Amsterdam, the Netherlands: North-Holland, 265-273.

Van Santen, J. P. H. 1992. Contextual effects on vowel duration, *Speech Communication 11*, 513-546.

Van Santen, J. P. H. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language 8*, 95-128.

Van Santen, J. P. H. 1997. Prosodic modelling in Text-to-Speech synthesis, *Proceedings of Eurospeech '97.*

Venditti, Jennifer J. & J. P. H. van Santen. 1998. Modelling segmental durations for Japanese text-to-speech synthesis, *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 31-36.

Received: October 29, 2003.
Accepted: December 15, 2003.

▲ Hyunsong Chung
Department of English Education
Daegu University
15 Naeri, Jinryang, Gyeongsan
Gyeongbuk 712–714
Korea
Tel: +82–53–850–4125
Fax: +82–53–850–4121
E–mail: hchung@daegu.ac.kr