OPTIMIZING SUB-COST FUNCTIONS FOR SEGMENT SELECTION BASED ON PERCEPTUAL EVALUATIONS IN CONCATENATIVE SPEECH SYNTHESIS

Tomoki Toda[†‡]*, Hisashi Kawai*[‡]*, and Minoru Tsuzaki*[‡]

[†]Graduate School of Engineering, Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya-shi, Aichi, 466-8555 Japan

[‡]ATR Spoken Language Translation Research Laboratories 2-2-2 Hikaridai, "Keihanna Science City" Kyoto, 619-0288 Japan

ABSTRACT

In concatenative speech synthesis, various factors affect the naturalness of synthetic speech. A cost for segment selection is calculated by integrating some sub-costs capturing the degradation of naturalness caused by such factors. In this paper, we optimize each sub-cost function for converting a linguistic feature or an acoustic parameter into a sub-cost based on perceptual evaluations. Two types of perceptual experiments are performed with test sets constructed by controlling the variations of sub-costs to evaluate the independent effect of each sub-cost and the interactions between them. We clarify the effectiveness of perceptually optimizing sub-cost functions from a result of a preference test comparing synthetic speech before and after the optimization.

1. INTRODUCTION

Corpus-based approaches to Text-to-Speech (TTS) dramatically improves the naturalness of synthetic speech [1]. As a result, the corpus-based TTS can be used for practical purposes under limited conditions [2]. However, it is necessary to improve the quality of TTS since it cannot synthesize sufficient natural speech consistently for any input text. As one of the ways of achieving that, we focus on improving a measure for selecting segments based on perceptual characteristics.

In concatenative speech synthesis, the measure is calculated with a cost function. In general, a cost consists of target and concatenation costs calculated from several linguistic features and acoustic parameters [3]. Various studies on the relationship between such features or parameters and the naturalness of synthetic speech have been reported, e.g., discontinuities captured by spectral parameters [4][5] and phonetic information [7][8], and the degradation of naturalness caused by modifying phonetic duration [9]. Since these costs are calculated in each segment, such as a phoneme, they can be considered as a local cost that captures the local degradation of naturalness. In order to evaluate the naturalness over an utterance, local costs in a segment sequence need to be integrated into one cost by a certain function.

We have applied a norm function to an integrated cost function and have optimized the function based on perceptual experiments [10][11]. In order to achieve a higher quality of synthetic speech, it is also necessary to optimize each sub-cost function for the features and parameters. In this paper, we perform perceptual evaluations of all the sub-costs used in our segment selection to optimize them.

In order to clarify the relationships between individual sub-costs and the degradation of naturalness, several test sets constructed by controlling the variations of sub-costs are used. Each sub-cost function is estimated so that the sub-cost corresponds to perceptual scores as accurately as possible. Furthermore, we perform a preference test to clarify the effectiveness of the optimization.

Table 1. Sub-costs

	Source information	Vocal tract information
Target cost	SC_{F_0}: F_0 SC_{dur}: Duration	SC_{cep}: Spectrum
Concatenation cost	SC_{F_0c}: F_0	SC_{env}: Phonetic category SC_{spg}: Spectrum

The paper is organized as follows. In **Section 2**, sub-cost functions for segment selection are described. In **Section 3**, a procedure for the perceptual optimization is described. In **Section 4**, perceptual evaluations of sub-costs are described, and optimizing sub-cost functions is described in **Section 5**. The effectiveness of the optimization is described in **Section 6**. Finally, we summarize this paper in **Section 7**.

2. SUB-COST FUNCTIONS FOR SEGMENT SELECTION

We use the six sub-costs shown in **Table 1** for segment selection. Target and concatenation costs are calculated from three different sub-costs, respectively. From these costs, the integrated cost is calculated. A cost on power is not used because we control segmental power in waveform concatenation.

SC_{F_0}, as given by Eq. (1), captures the difference in F_0 contour between a candidate segment u_i and a target t_i.

$$SC_{F_0}(u_i, t_i) = \frac{1}{P}\sum_{p=1}^{P} SF_{F_0}(LF_0(u_i, p), LF_0(t_i, p)), \qquad (1)$$

where $LF_0(u_i, p)$ denotes the averaged F_0 in log-scale for the p-th portion of an equally divided phoneme segment u_i. SF_{F_0} denotes a function to convert the difference of $\log F_0$ into the sub-cost.

SC_{dur}, as given by Eq. (2), captures the difference in phonetic duration between a candidate segment and a target.

$$SC_{dur}(u_i, t_i) = SF_{dur}(Dur(u_i), Dur(t_i)), \qquad (2)$$

where $Dur(u_i)$ denotes the duration of a phoneme segment u_i. SF_{dur} denotes a function to convert the difference of duration into the sub-cost.

SC_{cep}, as given by Eq. (3), captures the difference in phonetic mean spectrum between a candidate segment and a target.

$$SC_{cep}(u_i, t_i) = SF_{cep}(Cep(u_i), Cep(t_i)), \qquad (3)$$

where $Cen(u_i)$ denotes the mean spectrum of a phoneme segment u_i. SF_{cen} denotes a function to convert the difference of mean spectra into the sub-cost.

SC_{env}, as given by Eq. (4), captures the discontinuity caused by a mismatch of phonetic environments in segment concatenation.

$$SC_{env}(u_i, u_{i-1}) = SF^{s}_{env}((Ph(u_{i-1}), Ph_s(u_{i-1})), Ph(u_i)) + SF^{p}_{env}((Ph(u_i), Ph_p(u_i)), Ph(u_{i-1})), \quad (4)$$

where $Ph(u_i)$ denotes a phoneme for u_i, and $Ph_s(u_{i-1})$ and $Ph_p(u_i)$ denote a succeeding phoneme of u_{i-1} in the corpus and a preceding phoneme of u_i in the corpus, respectively. SF_{env} denotes a function to convert the substitution of phonetic environments into the sub-cost.

SC_{spg}, as given by Eq. (5), captures the spectral discontinuity around a concatenation boundary.

$$SC_{spg}(u_i, u_{i-1}) = \sum_{f=-l/2}^{l/2-1} w(f) \cdot SF_{spg}(Spg^{(h)}(u_i, f), Spg^{(t)}(u_{i-1}, f)), \quad (5)$$

where $w(f)$ denotes the triangular weighting function. $Spg^{(h)}(u_i, f)$ denotes the spectrum in the f-th frame from the first frame of u_i in the corpus, and $Spg^{(t)}(u_{i-1}, f)$ denotes that from the last frame of u_{i-1} in the corpus. SF_{spg} denotes a function to convert the spectral difference into the sub-cost.

SC_{F_0c}, as given by Eq. (6), captures the F_0 discontinuity at a concatenation boundary.

$$SC_{F_0c}(u_i, u_{i-1}) = SF_{F_0c}(LF_0^{(h)}(u_i), LF_0^{(t)}(u_{i-1})), \quad (6)$$

where $LF_0^{(h)}(u_i)$ denotes the log-scaled F_0 in the first frame of u_i and $LF_0^{(t)}(u_{i-1})$ denotes that in the last frame of u_{i-1}. SF_{F_0c} denotes a function to convert the F_0 difference into the sub-cost.

3. PROCEDURE FOR OPTIMIZING SUB-COST FUNCTIONS

In this paper, we assume that the local cost in each segment is calculated as the weighted sum of all sub-costs. Moreover, we define the sub-cost function described above as follows,

$$SF(x, y) = F(D(x, y)), \quad (7)$$

where F denotes a mapping function and D denotes a distance measure between x and y. We consider various function forms, e.g., linear and non-linear functions, and distance measures, e.g., a mahalanobis distance, which are denoted as $F_1, F_2, \cdots, F_n$ and $D_1, D_2, \cdots, D_m$, respectively.

A large amount of perceptual test data is needed to optimize F, D, and the weights for sub-costs at once. Moreover, since perceptual scores greatly depend on some sub-costs causing a large degradation of naturalness, it seems impossible to clarify the correspondence of the other sub-costs to perceptual scores. To address these problems, we separately perform evaluations for optimizing each sub-cost function and for optimizing the weights.

Test sets for individual sub-costs are prepared assuming independence between sub-costs (A sets). Each test set is constructed under the condition that a distance measure for one sub-cost is varied and those for the other sub-costs are kept as constant as possible. Moreover, another test set in which distance measures for all sub-costs are varied independently is prepared (B set). From the results of A sets, we can determine F and D in each sub-cost function so that the error between the sub-cost and perceptual scores is minimized, because the difference of perceptual scores in one test set greatly depends on only one sub-cost. However, because some sub-costs are not actually independent, e.g., SC_{F_0} and SC_{F_0c}, the determined F and D are not always the best in the case of integrating all sub-costs. Therefore, we estimate only the parameters of a mapping function in each pair of F_n and D_m. As a result, several candidates ($n \times m$) for a sub-cost function are estimated in each sub-cost. Then, the optimum set of sub-cost functions is determined from the results of B set taking into account the interaction of all sub-costs. The weights for sub-costs are also determined. Optimizing sub-cost functions is performed as follows:

1. Constructing test sets for evaluating the correspondences of individual sub-costs to perceptual scores (A sets) and for evaluating the correspondence in the case of integrating all sub-costs (B set).
2. Performing perceptual evaluations to determine perceptual scores in each test set.
3. Estimating several candidates for a sub-cost function using some function forms and distance measures in each sub-cost from the results of A sets.
4. Selecting the optimum set of the sub-cost functions having the best correspondence in the result of B set.

4. PERCEPTUAL EVALUATIONS OF SUB-COSTS

4.1. Designing test sets

We did not use word utterances but phrase utterances extracted from sentence utterances as test stimuli to match experimental conditions to the conditions under which TTS is actually used. In order to easily control the variations of sub-costs, the selection of candidate segments was performed at only one syllable in a target phrase. Target information for the selection was extracted from natural speech. The size of the corpus from which candidate segments were selected was 35 hours. A test set of 503 sentences that were not included in the corpus was used as targets.

A stimulus was synthesized by substituting a syllable segment in the carrier phrase with a candidate segment in the corpus. In each target syllable, 100 candidate segments remaining after preselection with some sub-costs were actually used.

A set of stimuli for each sub-cost was constructed by selecting stimuli so that a distant measure for one sub-cost was varied and distant measures for the other sub-costs were kept as constant as possible. Six test sets (A_{F_0}–A_{F_0c} sets) were constructed for Exp. A. The number of stimuli was 30 each. Moreover, another test set (B set) in which distance measures for all sub-costs were varied independently was constructed for Exp. B by extracting 105 stimuli from A sets.

4.2. Perceptual evaluations

Seven perceptual evaluations with the constructed test sets were performed independently. Pairs of natural speech and synthetic speech were presented to listeners. Parts of substitution were shown to listeners when presenting each stimulus-pair. The degradation of naturalness was evaluated with a 5-point scale. Listeners were instructed to use 5 points widely in each test set. The number of

Table 2. Sub-cost functions before optimization. $F_l(\cdot)$ denotes linear function for scaling. Mel-CD denotes mel-cepstral distance

Sub-cost function	Function form	Distance measure
$SF_{F_0}(\cdot)$	$F_l(\cdot)$	Distance of $\log F_0$
$SF_{dur}(\cdot)$	$F_l(\cdot)$	Distance of duration
$SF_{cen}(\cdot)$	$F_l(\cdot)$	Mel-CD between mean spectra
$SF_{env}(\cdot)$	$F_l(\cdot)$	Mel-CD predicted from phonetic information with linear regression
$SF_{spg}(\cdot)$	$F_l(\cdot)$	Mel-CD between frames
$SF_{F_0c}(\cdot)$	$F_l(\cdot)$	Distance of $\log F_0$

Table 3. Correlation coefficients between individual sub-costs and perceptual scores in Exp. A.

Sub-cost	Before optimization	After optimization
SC_{F_0}	0.748	0.830
SC_{dur}	0.396	0.624
SC_{cen}	0.613	0.773
SC_{env}	0.450	0.666
SC_{spg}	0.456	0.456
SC_{F_0c}	0.748	0.700

listeners who participated in the experiments were 15 to 17 for Exp. A and 22 for Exp. B, respectively. The perceptual score for each stimulus was calculated as an average of the normalized score calculated as a Z-score (mean = 0 and variance = 1 in each test set).

4.3. Experimental results

We use the simple sub-cost functions shown in **Table 2** before the optimization. These sub-costs are basically calculated as the Euclidean distance of acoustic parameters. Moreover, each sub-cost is normalized by linear conversion so that the average of the sub-cost is equal to 1.5.

Table 3 ("Before optimization") shows correlation coefficients between the individual sub-costs and perceptual scores in Exp. A. It can be seen that the sub-costs on F_0, i.e., SC_{F_0} and SC_{F_0c}, have better correspondences than the other sub-costs.

A multiple correlation coefficient between an estimated cost, which is calculated as a weighted sum of all sub-costs, and the perceptual scores obtained in Exp. B is shown in **Table 4** ("Before optimization"). This correlation is equal to that between a cost, which is calculated with the optimum weights for sub-costs, and perceptual scores. This result indicates that only optimizing weights for sub-costs is not enough to achieve a good correspondence of the cost to the perceptual scores.

Table 4. Multiple correlation coefficients between a cost calculated as a weighted sum of all sub-costs and perceptual scores in Exp. B

Before optimization	After optimization
0.528	0.696

5. OPTIMIZING SUB-COST FUNCTIONS

5.1. Estimating candidates for each sub-cost function

Although some sub-costs, e.g., SC_{F_0} and SC_{F_0c}, have good correspondences in **Table 3**, it is possible that a non-linear mapping is more effective than a linear mapping. Moreover, it is possible that the degradation of naturalness cannot be perceived in the case where the distance in a measure is less than a certain threshold. Therefore, we consider the following function forms as candidates for a mapping function:

$$F_l(D) = a \cdot D - b, \tag{8}$$

$$F_g(D) = -a \cdot \exp\left(-\left(\frac{D}{c}\right)^2\right) + b, \tag{9}$$

$$F_s(D) = \frac{a}{(1 + \exp(-c \cdot (D - d)))} - b, \tag{10}$$

where a, b, c, and d denote parameters in each function, and D denotes a distance measure.

Parameters are estimated by minimizing the mean square error between the sub-cost and perceptual scores on the result of Exp. A. The bias parameter b is adjusted so that the minimum value of the sub-cost is equal to 0. We consider various distance measures, $D_1, D_2, \cdots, D_m$ for each sub-cost. Therefore, many candidates $(3 \times m)$ for a sub-cost function are estimated.

5.2. Selecting the best sub-cost functions

The optimum set of sub-cost functions is selected from the estimated candidate functions while considering the interaction between sub-costs as follows.

Multiple correlation coefficients between the cost calculated as a weighted sum of all sub-costs and perceptual scores in Exp. B are calculated for all possible combinations of the estimated candidate functions. A set of sub-cost functions having the best multiple correlation coefficient is selected.

5.3. Results after optimization of sub-cost functions

The optimized sub-cost functions are shown in **Table 5**. All sub-cost functions except for SF_{spg} are different from those before the optimization.

Correlation coefficients between individual sub-costs after the optimization and perceptual scores in Exp. A are also shown in **Table 3** ("After optimization"). It can be seen that the correspondences of almost all sub-costs are improved except for SC_{F_0c}.

The partial correlation coefficients between individual sub-costs and perceptual scores in Exp. B are shown in **Table 6**. A partial correlation of SC_{F_0c} before the optimization is almost 0 although the correspondence of it to perceptual scores in Exp. A is good. In other words, this sub-cost is not meaningful in the case of integrating all sub-costs. After the optimization, improvement of the partial correlation of this sub-cost can be seen. This fact shows that it is effective to optimize sub-cost functions taking into account the interaction of all sub-costs.

A multiple correlation coefficient between the estimated cost and perceptual scores is much improved by the optimization as shown in **Table 4** ("After optimization"), which indicates the effectiveness of the procedure for optimizing sub-cost functions as described in this section.

Table 5. Sub-cost functions optimized by perceptual evaluations

Sub-cost function	Function form	Distance measure
$SF_{F_0}(\cdot)$	$F_s(\cdot)$	Distance of $\log F_0$
$SF_{dur}(\cdot)$	$F_l(\cdot)$	Distance of duration normalized by standard deviation calculated in each phoneme
$SF_{cep}(\cdot)$	$F_g(\cdot)$	Mel-CD between mean spectra normalized by determinant of covariance calculated in each phoneme
$SF_{env}(\cdot)$	$F_l(\cdot)$	Perceptual scores in experiments described in [8]
$SF_{spg}(\cdot)$	$F_l(\cdot)$	Mel-CD between frames
$SF_{F_0c}(\cdot)$	$F_s(\cdot)$	Distance of $\log F_0$ at voiced phoneme boundary

Table 6. Partial correlation coefficients between individual sub-costs and perceptual scores in Exp. B

Sub-cost	Before optimization	After optimization
SC_{F_0}	0.362	0.421
SC_{dur}	0.174	0.324
SC_{cep}	0.157	0.128
SC_{env}	0.102	0.244
SC_{spg}	0.118	0.203
SC_{F_0c}	0.043	0.269

6. EXPERIMENTAL VERIFICATION OF EFFECTIVENESS OF PERCEPTUAL OPTIMIZATION

We performed a preference test to clarify the effectiveness of the above described method for optimizing sub-cost functions to the improvement of the naturalness of synthetic speech. Synthetic speech before and after the optimization were compared in pairs. Weights for sub-costs were set to equal values to focus on the effectiveness of the optimization of sub-cost functions. The corpus size was 35 hours. A set of 53 sentences that were not included in the corpus was used to synthesize the stimuli. Natural prosody was used as the target for segment selection. Ten Japanese listeners participated in the test.

The preference score in the case of optimizing sub-cost functions was $61.70 \pm 2.93\%$ (95% confidence interval). Therefore, perceptually optimizing sub-cost functions is effective for improving the naturalness of synthetic speech.

We also analyzed the preference scores for individual listeners and those for individual sentences. For listeners, the preference scores were over 50% (the minimum score = 51.9%, the maximum score = 68.9%). The preference scores in 7 listeners were significantly larger than 50%. On the other hand, the preference scores for individual sentences were not always over 50%. In fact, the preference scores for 4 sentences were significantly smaller than 50%. These results reveal that it is difficult to improve the naturalness of synthetic speech for any input text by optimizing cost functions.

7. CONCLUSION

We optimized each sub-cost function for converting a linguistic feature or an acoustic parameter into a sub-cost based on perceptual evaluations. Two types of perceptual experiments were designed so that the independent effect of each sub-cost and the interactions between them could be evaluated. We also performed a preference test comparing synthetic speech before and after optimizing sub-cost functions. As a result, it was clarified that the perceptual optimization of sub-cost functions is effective for improving the naturalness of synthetic speech. Since weights for sub-costs were also determined in this optimization, we need to clarify the effectiveness of optimizing the weights.

Acknowledgment: This research was supported in part by the Telecommunications Advancement Organization of Japan and JSPS Research Fellowships for Young Scientists.

8. REFERENCES

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. ICASSP*, pp. 679–682, New York, U.S.A., Apr. 1988.

[2] A.W. Black and K. Lenzo. Limited domain synthesis. *Proc. ICSLP*, Vol. 2, pp. 411–414, Beijing, China, Sep. 2000.

[3] A.W. Black and N. Campbell. Optimizing selection of units from speech database for concatenative synthesis. *Proc. EUROSPEECH*, pp. 581–584, Madrid, Spain, Sep. 1995.

[4] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 39–51, 2001.

[5] Y. Stylianou and A.K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. ICASSP*, pp. 837–840, Salt Lake City, U.S.A., May 2001.

[6] W. Ding, K. Fujisawa, and N. Campbell. Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification. *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 191–194, Jenolan Caves, Australia, Nov. 1998.

[7] A.K. Syrdal. Phonetic effects on listener detection of vowel concatenation. *Proc. EUROSPEECH*, pp. 979–982, Aalborg, Denmark, Sep. 2001.

[8] H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis. *Proc. ICSLP*, pp. 2621–2624, Denver, U.S.A., Sep. 2002.

[9] H. Kato, M. Tsuzaki, and Y. Sagisaka. Acceptability for temporal modification of single vowel segments in isolated words. *J. Acoust. Soc. Am.*, Vol. 104, No. 1, pp. 540–549, 1998.

[10] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Perceptual evaluation of cost for segment selection in concatenative speech synthesis. *IEEE Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[11] T. Toda, H. Kawai, and M. Tsuzaki. Optimizing integrated cost function for segment selection in concatenative speech synthesis based on perceptual evaluations. *Proc. EUROSPEECH*, pp. 297–300, Geneva, Switzerland, Sep. 2003.