TR－SLT－0033

Optimizing Segment Selection for High-Quality Text-to-Speech

高品質なテキスト音声合成のための素片選択の最適化

Tomoki Toda　　Hisashi Kawai
戸田　智基　　河井　恒
Minoru Tsuzaki　　Kiyohiro Shikano
津崎　実　　鹿野　清宏 *

March 05, 2003

This report addresses the problem of how to improve the naturalness of synthetic speech in corpus-based Text-to-Speech. To deal with this problem, we focus on two factors: (1) an algorithm for selecting the most appropriate synthesis units from a speech corpus, and (2) an evaluation measure for selecting the synthesis units. We confirm that the proposed segment selection algorithm and the proposed cost function based on perceptual evaluation are effective for improving the naturalness of synthetic speech.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

* 奈良先端科学技術大学院大学 情報科学研究科

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

Contents

Chapter 1

Introduction

1.1 Background and Problem Definition

Speech is the ordinary way for most people to communicate. Moreover, speech can convey other information such as emotion, attitude, and speaker individuality. Therefore, it is said that speech is the most natural, convenient, and useful means of communication.

In recent years, computers have come into common use as computer technology advances. Therefore, it is important to realize a man-machine interface to facilitate communication between people and computers. Naturally, speech is focused on as a medium for such communication. In general, two technologies for processing speech are needed. One is speech recognition, and the other is speech synthesis. Speech recognition is a technique for information input. Necessary information, e.g. message information, is extracted from input speech that includes diverse information. Thus, it is important to find a method to extract only useful information. On the other hand, speech synthesis is a technique for information output. This procedure is the reverse of speech recognition. Output speech includes various types of information, e.g. sound information and prosodic information, and is generated from input information. Moreover, other information such as speaker individuality and emotion is needed in order to realize smoother communication. Thus, it is important to find a method to generate the various types of paralinguistic information that are not processed in speech recognition.

Text-to-Speech (TTS) is one of the speech synthesis technologies. TTS is a technique to convert any text into a speech signal [60], and it is very useful in many practical applications, e.g. car navigation, announcements in railway stations, response services in telecommunications, e-mail reading. Therefore, it is desirable to realize TTS that can synthesize natural and intelligible speech, and research and development on TTS has been progressing.

The current trend in TTS is based on a large amount of speech data and statistical processing. This type of TTS is generally called corpus-based TTS. This approach makes it possible to dramatically improve the naturalness of synthetic speech compared with the early TTS. Corpus-based TTS can be used for practical purposes under limited conditions [12]. However, no general-purpose TTS has been developed that can synthesize sufficient natural speech consistently for any input text. Therefore, it is necessary to improve the performance of corpus-based TTS.

1.2 Report Scope

In this report, we improve the naturalness of synthetic speech in corpus-based TTS.

In corpus-based TTS, three main factors determine the naturalness of synthetic speech: (1) a speech corpus, (2) an algorithm for selecting the most appropriate synthesis units from the speech corpus, and (3) an evaluation measure to select the synthesis units. We focus on the latter two factors.

In a speech synthesis procedure, the optimum set of waveform segments, i.e. portions of speech utterances included in the corpus, are selected, and the synthetic speech is generated by concatenating the selected waveform segments. This selection is performed based on synthesis units. Various units, phonemes, diphones, syllables, and so on have been proposed. In Japanese speech synthesis, syllable units are often used since the number of Japanese syllables is small and transition in the syllables is important for intelligibility. However, syllable units cannot avoid vowel-to-vowel concatenation, which often produces auditory discontinuity, because various vowel sequences appear frequently in Japanese. In order to alleviate this discontinuity, we propose a novel selection algorithm based on two synthesis unit definitions.

Moreover, in order to realize high and consistent quality of synthetic speech, it is important to use an evaluation measure that corresponds to perceptual characteristics in the selection of the most suitable waveform segments. Although a measure based on acoustic measures is often used, the correspondence of such a measure to the perceptual characteristics is indistinct. Therefore, we clarify the correspondence of the measure utilized in our TTS by performing perceptual experiments on the naturalness of synthetic speech. Moreover, we improve this measure based on the results of these experiments.

1.3 Report Overview

The report is organized as follows.

In **Chapter 2**, a corpus-based TTS system is described. We describe the basic structure of the corpus-based TTS system. Then some techniques in each module are reviewed, and we briefly introduce the techniques applied to the TTS system under development in ATR Spoken Language Translation Research Laboratories.

In **Chapter 3**, we propose a novel segment selection algorithm for Japanese speech synthesis. Not only the segment selection algorithms but also our measure for selection of optimum segments are described. Results of perceptual experiments show that the proposed algorithm can synthesize speech more naturally than conventional algorithms.

In **Chapter 4**, the measure is evaluated based on perceptual characteristics. We clarify correspondence of the measure to the perceptual scores determined from the results of perceptual experiments. Moreover, we find a novel measure having better correspondence and investigate the effect of using this measure for segment selection. We also show the effectiveness of increasing the size of a speech corpus.

In **Chapter 5**, we summarize the contributions of this report and offer suggestions for future work.

Chapter 2

Corpus-Based Text-to-Speech

Corpus-based TTS is the main current direction in work on TTS. The naturalness of synthetic speech has been improved dramatically by the transition from the early rule-based TTS to corpus-based TTS. In this section, we describe the basic structure of corpus-based TTS and the various techniques used in each module.

2.1 Introduction

The early TTS was constructed based on rules that researchers determined from their objective decisions and experience [60]. In general, this type of TTS is called rule-based TTS. The researcher extracts the rules for speech production by the Analysis-by-Synthesis (A-b-S) method [3]. In the A-b-S method, parameters characterizing a speech production model are adjusted by performing iterative feedback control so that the error between the observed value and that produced by the model is minimized. Such rule determination needs professional expertise since it is difficult to extract consistent and reasonable rules. Therefore, the rule-based TTS systems developed by researchers usually have different performances. Moreover, synthetic speech by rule-based TTS has an unnatural quality because a speech waveform is generated by a speech production model, e.g. terminal analog speech synthesizer, which generally needs some approximations in order to model the complex human vocal mechanism [60].

On the other hand, the current TTS is constructed based on a large amount of data and a statistical process [38][70]. In general, this type of TTS is called corpus-based TTS in contrast with rule-based TTS. This approach has been developed through the dramatic improvements in computer performance. In corpus-based TTS, a large amount of speech data are stored as a speech corpus. In synthesis, optimum speech units are selected from the speech corpus. An output speech waveform is synthesized by concatenating the selected units and then modifying their prosody. Corpus-based TTS can synthesize speech more naturally than rule-based TTS because the degradation of naturalness in synthetic speech can be alleviated by selecting units satisfying certain factors, e.g. a mismatch of phonetic environments, difference in prosodic information, and discontinuity produced by concatenating units. If the selected units need little modification, natural speech can be synthesized by concatenating speech waveform segments directly. Furthermore, since the corpus-based approach has hardly any dependency on the type of language, we can apply the approach to other languages more easily than the rule-based approach.

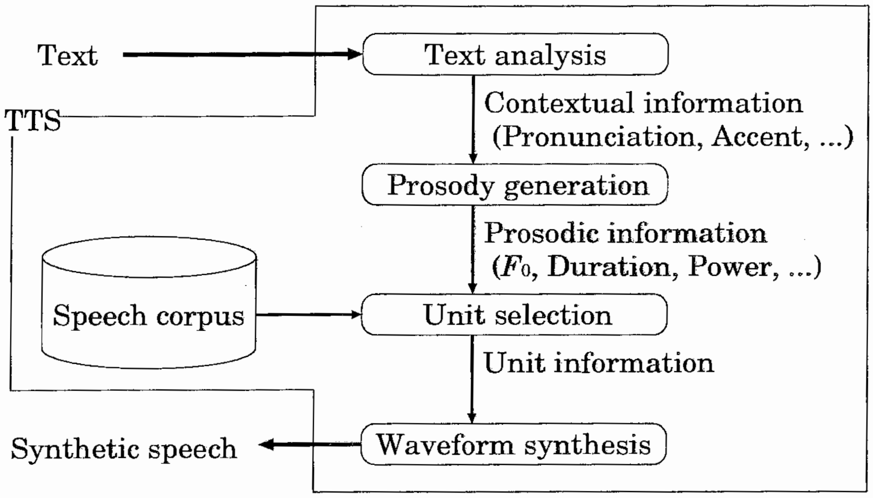

Figure 2.1: Structure of corpus-based TTS.

2.2 Structure of Corpus-Based TTS

In general, corpus-based TTS is comprised of five modules: text analysis, prosody generation, unit selection, waveform synthesis, and speech corpus. The structure of corpus-based TTS is shown in **Figure 2.1**.

2.2.1 Text analysis

In the text analysis, an input text is converted into contextual information, i.e. pronunciation, accent type, part-of-speech, and so on, by natural language processing [72][76]. The contextual information plays an important role in the quality and intelligibility of synthetic speech because prediction accuracy on this information affects all of the subsequent procedures.

First, various obstacles, such as unreadable marks like HTML tags and e-mail headings, are removed if the input text includes these obstacles. This processing is called text normalization.

The normalized text is then divided into morphemes, which are minimum units of letter strings having linguistic meaning. These morphemes are tagged with their parts of speech, and a syntactic analysis is performed. Then, the module determines phoneme and prosodic symbols, e.g. accent nucleus, accentual phrases, boundaries of prosodic clauses, and syntactic structure. Reading rules and accentual rules for word concatenation are often applied to the determination of this information [62][69]. Especially in Japanese, the accent information is crucial to achieving high-quality synthetic speech. In some TTS systems, especially English TTS systems, ToBI (Tone and Break Indices) labels [77] or Tilt parameters [84] are predicted [14][33].

A schematic diagram of text analysis is shown in **Figure 2.2**.

2.2.2 Prosody generation

In the prosody generation, prosodic features such as F_0 contour, power contour, and phoneme duration are predicted from the contextual information output from the text analysis. This

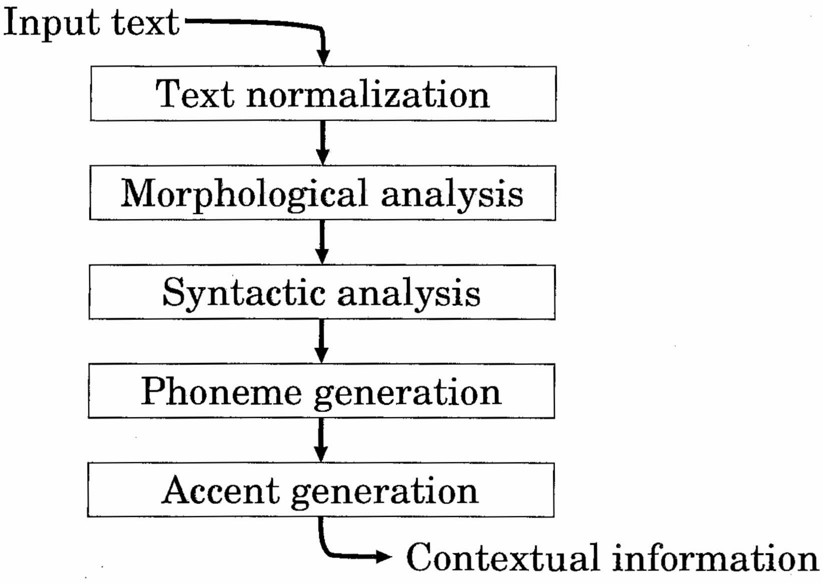

Figure 2.2: Schematic diagram of text analysis.

prosodic information is important for the intelligibility and naturalness of synthetic speech.

Fujisaki's model has been proposed as one of the models that can represent the F_0 contour effectively [34]. This model decomposes the F_0 contour into two components, i.e. a phrase component that decreases gradually toward the end of a sentence and an accent component that increases and decreases rapidly at each accentual phrase. Fujisaki's model is often used to generate the F_0 contour from the contextual information in rule-based TTS, particularly in Japanese TTS [40][54]. Then the rules arranged by experts are applied. In recent years, automatic extraction algorithms of control parameters and rules from a large amount of data with statistical methods have been proposed [36][41].

Many data-driven algorithms for prosody generation have been proposed. In the F_0 contour control model proposed by Kagoshima et al. [50], an F_0 contour of a whole sentence is produced by concatenating segmental F_0 contours, which are generated by modifying vectors that are representative of typical F_0 contours. The representative vectors are selected from an F_0 contour codebook with contextual information. The codebook is designed so that the approximation error between F_0 contours generated by this model and real F_0 contours extracted from a speech corpus is minimized. Isogai et al. proposed using not the representative vectors but natural F_0 contours selected from a speech corpus in order to generate an F_0 contour of a sentence [45]. In this algorithm, if there is an F_0 contour having equal contextual information to the predicted contextual information in the speech corpus, the F_0 contour is selected and used without modification. In all other cases, the F_0 contour that most suits the predicted contextual information is selected and used with modification. Moreover, algorithms for predicting the F_0 contour from the ToBI labels or Tilt parameters have been proposed [10][32].

As a powerful data-driven algorithm, HMM-based (Hidden Markov model) speech synthesis has been proposed by Tokuda et al. [85][86][90]. In this method, the F_0 contour, the mel-cepstrum sequence including the power contour, and phoneme duration are generated directly from HMMs trained by a decision-tree based on a context clustering technique. The F_0

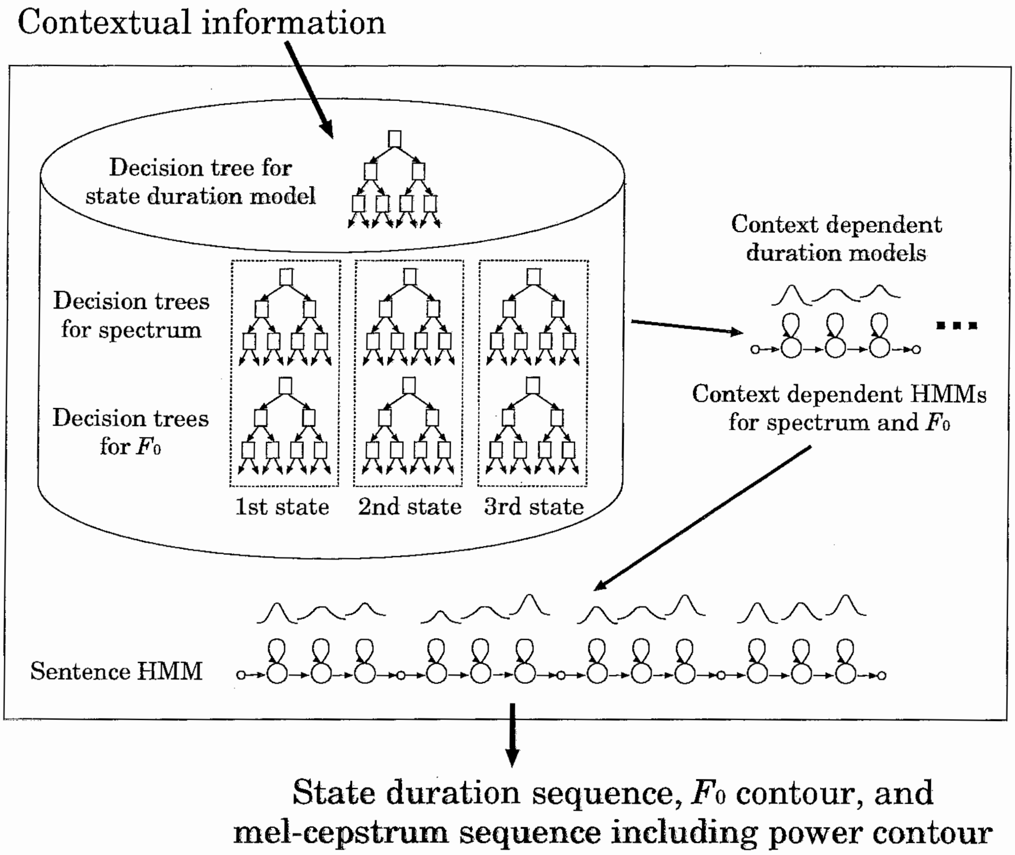
Contextual information
Decision tree for state duration model
Context dependent duration models
Decision trees for spectrum
Decision trees for F_0
Context dependent HMMs for spectrum and F_0
1st state
2nd state
3rd state
Sentence HMM
State duration sequence, F_0 contour, and mel-cepstrum sequence including power contour

Figure 2.3: Schematic diagram of HMM-based prosody generation.

is modeled by multi-space probability distribution HMMs [85], and the duration is modeled by multi-dimensional Gaussian distribution HMMs in which each dimension shows the duration in each state of the HMM. The mel-cepstrum is modeled by either multi-dimensional Gaussian distribution HMMs or multi-dimensional Gaussian mixture distribution HMMs. Decision-trees are constructed for each feature. The decision-tree for the F_0 and that for the mel-cepstrum are constructed in each state of the HMM. As for the duration, one decision-tree is constructed. All training procedures are performed automatically. In synthesis, the smooth parameter contours, which are static features, are generated from the HMMs by maximizing the likelihood criterion while considering the dynamic features of speech [86].

Some TTS systems do not perform the prosody generation [20]. In these systems, contextual information is used instead of prosody information for the next procedure, unit selection.

In our corpus-based TTS under development, HMM-based speech synthesis is applied to a prosody generation module. A schematic diagram of HMM-based prosody generation is shown in **Figure 2.3**.

2.2.3 Unit selection

In the unit selection, an optimum set of units is selected from a speech corpus by minimizing the degradation of naturalness caused by various factors, e.g. prosodic difference, spectral difference, and a mismatch of phonetic environments [47][70]. Various types of units have been proposed to alleviate such degradation.

Nakajima et al. proposed an automatic procedure called Context Oriented Clustering (COC) [64]. In this technique, the optimum synthesis units are generated or selected from a speech corpus of a single speaker in advance in order to alleviate degradation caused by spectral difference. All segments of a given phoneme in the speech corpus are clustered in advance into equivalence classes according to their preceding and succeeding phoneme contexts. The decision trees that perform the clustering are constructed automatically by minimizing the acoustic differences within the equivalence classes. The centroid segment of each cluster is saved as a synthesis unit. In the speech synthesis phase, the optimum synthesis units are selected from leaf clusters that most suit given phonetic contexts. As the synthesis units, either spectral parameter sequences [35] or waveform segments [46] are utilized.

Kagoshima and Akamine proposed an automatic generation method of synthesis units with Closed-Loop Training (CLT) [2][49]. In this approach, an optimum set of synthesis units is selected or generated from a speech corpus in advance to minimize the degradation caused by synthesis processing such as prosodic modification. A measure capturing this degradation is defined as the difference between a natural speech segment prepared as training data cut off from the speech corpus and a synthesized speech segment with prosodic modification. The selection or generation of the best synthesis unit is performed on the basis of the evaluation and minimization of the measure in each unit cluster represented by a diphone [68]. Although the number of diphone waveform segments used as synthesis units is very small (only 302 segments), speech with natural and smooth sounding quality can be synthesized.

There are many types of basic synthesis units, e.g. phoneme, diphone, syllable, VCV units [73], and CVC units [74]. The units comprised of more than two phonemes can preserve transitions between phonemes. Therefore, the concatenation between phonemes that often produces perceptual discontinuity can be avoided by utilizing these units. The diphone units have unit boundaries at phoneme centers [27][65][68]. In the VCV units, concatenation points are vowel centers in which formant trajectories are stabler and clearer than those in consonant centers [73]. While, in the CVC units, concatenation is performed at the consonant centers in which waveform power is often smaller than that in the vowel centers [74]. In Japanese, CV (C: Consonant, V: Vowel) units are often used since nearly all Japanese syllables consists of CV or V.

In order to use stored speech data effectively and flexibly, Sagisaka et al. proposed Non-Uniform Units (NUU) [47][70][83]. In this approach, the specific units are not selected or generated from a speech corpus in advance. An optimum set of synthesis units is selected by minimizing a cost capturing the degradation caused by spectral difference, difference in phonetic environment, and concatenation between units in a synthesis procedure. Since it is possible to use all phoneme subsequences as synthesis units, the selected units, i.e. NUU, have variable lengths. The ATR ν-talk speech synthesis system is based on the NUU represented by a spectral parameter sequence [71]. Hirokawa et al. proposed that not only factors related to spectrum and phonetic environment but also prosodic difference are considered in selecting the optimum synthesis units [39]. In this approach, speech is synthesized by concatenating the selected waveform segments and then modifying their prosody. Campbell et al. also proposed utilization of prosodic information in selecting the synthesis units [16][17]. Based on these works, Black et al. proposed a general algorithm for unit selection by using two costs [8][18][42]. One is a target cost, which captures the degradation caused by prosodic difference and difference in phonetic environment, and the other is a concatenation cost, which captures the degradation caused by concatenating units. In this algorithm, the sum of the two costs is minimized using a dynamic programming search based on phoneme

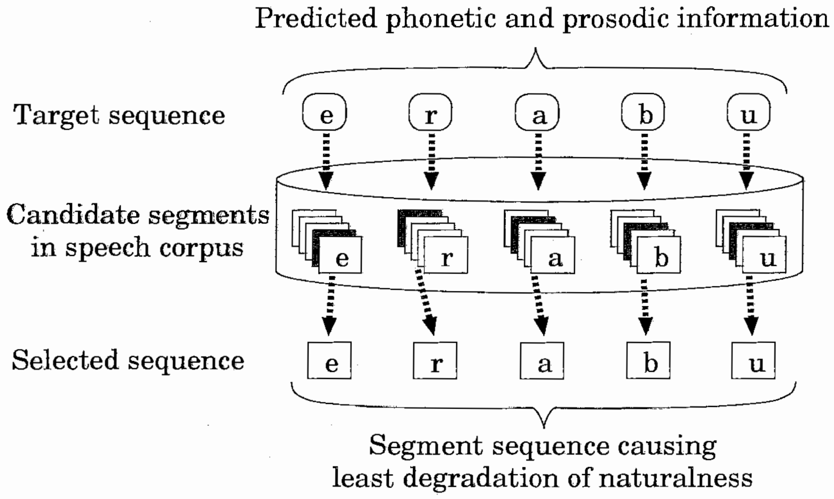
Predicted phonetic and prosodic information
Target sequence
e
r
a
b
u
Candidate segments in speech corpus
e
r
a
b
u
Selected sequence
e
r
a
b
u
Segment sequence causing least degradation of naturalness

Figure 2.4: Schematic diagram of segment selection.

units. By introducing these techniques to ν-talk, CHATR is constructed as a generic speech synthesis system [7][9]. Since the number of considered factors increases, a larger-sized speech corpus is utilized than that of ν-talk. If the size of a corpus is large enough and it's possible to select waveform segments satisfying target prosodic features predicted by the prosody generation, it is not necessary to perform prosody modification [19]. Therefore, natural speech without degradation caused by signal processing can be synthesized by concatenating the waveform segments directly. This waveform segment selection has become the main current of corpus-based TTS systems for any language. In recent years, Conkie proposed a waveform segment selection based on half-phoneme units to improve the robustness of the selection [24].

CV* units [53] and multiform units [82] have been proposed as synthesis units by Kawai et al. and Takano et al., respectively. These units can preserve the important transitions for Japanese, i.e. V-V transitions, in order to alleviate the perceptual discontinuity caused by concatenation. The units are stored in a speech corpus as sequences of phonemes with phonetic environments. A stored unit can have multiple waveform segments with different F_0 or phoneme duration. Therefore, optimum waveform segments can be selected while considering both the degradation caused by concatenation and that caused by prosodic modification. In general, the number of concatenations becomes smaller by utilizing longer units. However, the longer units cannot always synthesize natural speech, since the number of candidate units becomes small and the flexibility of prosody synthesis is lost.

Shorter units have also been proposed. Donovan et al. proposed HMM state-based units [28][29]. In this approach, decision-tree state-clustered HMMs are trained automatically with a speech corpus in advance. In order to determine the segment sequence to concatenate, a dynamic programming search is performed over all waveform segments aligned to each leaf of the decision-trees in synthesis. In the HMM-based speech synthesis proposed by Yoshimura et al. [90], the optimum HMM sequence is selected from decision-trees by utilizing phonetic and prosodic context information.

In our corpus-based TTS, the waveform segment selection technique is applied to a unit selection module. A schematic diagram of the segment selection is shown in **Figure 2.4**.

2.2.4 Waveform synthesis

An output speech waveform is synthesized from the selected units in the last procedure of TTS. In general, two approaches to waveform synthesis have been used. One is waveform concatenation without speech modification, and the other is speech synthesis with speech modification.

In the waveform concatenation, speech is synthesized by concatenating waveform segments selected from a speech corpus using prosodic information to remove need for signal processing [19]. In this case, instead of performing prosody modification, raw waveform segments are used. Therefore, synthetic speech has no degradation caused by signal processing. However, if the prosody of the selected waveform segments is different from the predicted target prosody, degradation is caused by the prosodic difference [39]. In order to alleviate the degradation, it is necessary to prepare a large-sized speech corpus that contains abundant waveform segments. Although synthetic speech by waveform concatenation sounds very natural, the naturalness is not always consistent.

In the speech synthesis, signal processing techniques are used to generate a speech waveform with the target prosody. The Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) algorithm is often used for prosody modification [63]. TD-PSOLA does not need any analysis algorithm except for determination of pitch marks throughout the segments. Speech analysis-synthesis methods can also modify the prosody. In the HMM-synthesis method, a mel-cepstral analysis-synthesis technique is performed [90]. Speech is synthesized from a mel-cepstrum sequence generated directly from the selected HMMs and the excitation source by utilizing a Mel Log Spectrum Approximation (MLSA) filter [44]. A vocoder type algorithm such as this can modify speech easily by varying speech parameters, i.e. spectral parameter and source parameter [31]. However, the quality of the synthetic speech is often degraded. As a high-quality vocoder type algorithm, Kawahara et al. proposed the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) analysis-synthesis method [51]. STRAIGHT uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region to remove signal periodicity and designs an excitation source based on phase manipulation. Moreover, STRAIGHT can manipulate such speech parameters as pitch, vocal tract length, and speaking rate while maintaining high reproductive quality. Stylianou proposed the Harmonic plus Noise Model (HNM) as a high-quality speech modification technique [78]. In this model, speech signals are represented as a time-varying harmonic component plus a modulated noise component. Speech synthesis with these modification algorithms is very useful in the case of a small-sized speech corpus. Synthetic speech by this speech synthesis sounds very smooth, and the quality is consistent. However, the naturalness of the synthetic speech is often not as good as that of synthetic speech by waveform concatenation.

In our corpus-based TTS, both the waveform concatenation technique and STRAIGHT synthesis method are applied in the waveform synthesis module. In the waveform concatenation, we control waveform power in each phoneme segment by multiplying the segment by a certain value so that average power in a phoneme segment selected from a speech corpus becomes equal to average target power in the phoneme. When the segments modified by this power are concatenated, an overlap-add technique is applied in the frame-pair with the highest correlation around a concatenation boundary between the segments. A schematic diagram of the waveform concatenation is shown in **Figure 2.5**. In the other synthesis method based on STRAIGHT, speech waveforms in voiced phonemes are synthesized with STRAIGHT by using a concatenated spectral sequence, a concatenated aperiodic energy sequence, and target prosodic features. In unvoiced phonemes, we use original waveforms modified only by power. A schematic diagram of the speech synthesis with prosody modification by STRAIGHT is shown in **Figure 2.6**.

2.2.5 Speech corpus

A speech corpus directly influences the quality of synthetic speech in corpus-based TTS. In order to realize a consistently high quality of synthetic speech, it is important to prepare a speech corpus containing abundant speech segments with various phonemes, phonetic environments, and prosodies, which should be recorded while maintaining high quality.

Abe et al. developed a Japanese sentence set in which phonetic coverage is controlled [1]. This sentence set is often used not only in the field of speech synthesis but also in speech recognition. Kawai et al. proposed an effective method for designing a sentence set for utterances by taking into account prosodic coverage as well as phonetic coverage [55]. This method selects the optimum sentence set from a large number of sentences by maximizing the measure of coverage. The size of the sentence set, i.e. the number of sentences, is decided in advance. The coverage measure captures two factors, i.e. (1) the distributions of F_0 and phoneme duration predicted by the prosody generation and (2) perceptual degradation of naturalness due to the prosody modification.

In general, the degradation of naturalness caused by a mismatch of phonetic environments and prosodic difference can be alleviated by increasing the size of the speech corpus. However, variation in voice quality is caused by recording the speech of a speaker for a long time in order to construct the large-sized corpus [56]. Concatenation between speech segments with different voice qualities produces audible discontinuity. To deal with this problem, previous studies have proposed using a measure capturing the difference in voice quality to avoid concatenation between such segments [56] and normalization of power spectral densities [75].

In our corpus-based TTS, a large-sized corpus of speech spoken by a Japanese male who is a professional narrator is under construction. The maximum size of the corpus used in this report is 32 hours. A sentence set for utterances is extracted from TV news articles, newspaper articles, phrase books for foreign tourists, and so on by taking into account prosodic coverage as well as phonetic coverage.

2.3 Summary

This section described the basic structure of corpus-based Text-to-Speech (TTS) and reviewed the various techniques in each module. We also introduced some techniques applied to the corpus-based TTS under development in ATR Spoken Language Translation Research Laboratories.

Corpus-based TTS improve the naturalness of synthetic speech dramatically compared with rule-based TTS. However, its naturalness is still inadequate.

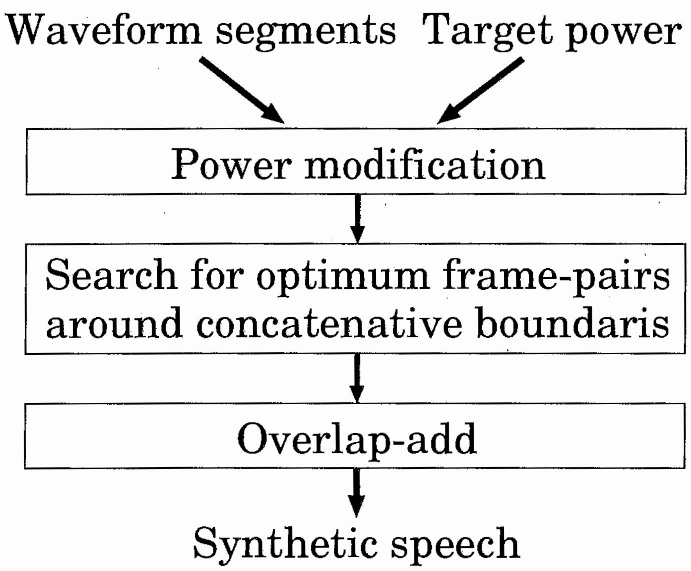

Figure 2.5: Schematic diagram of waveform concatenation.

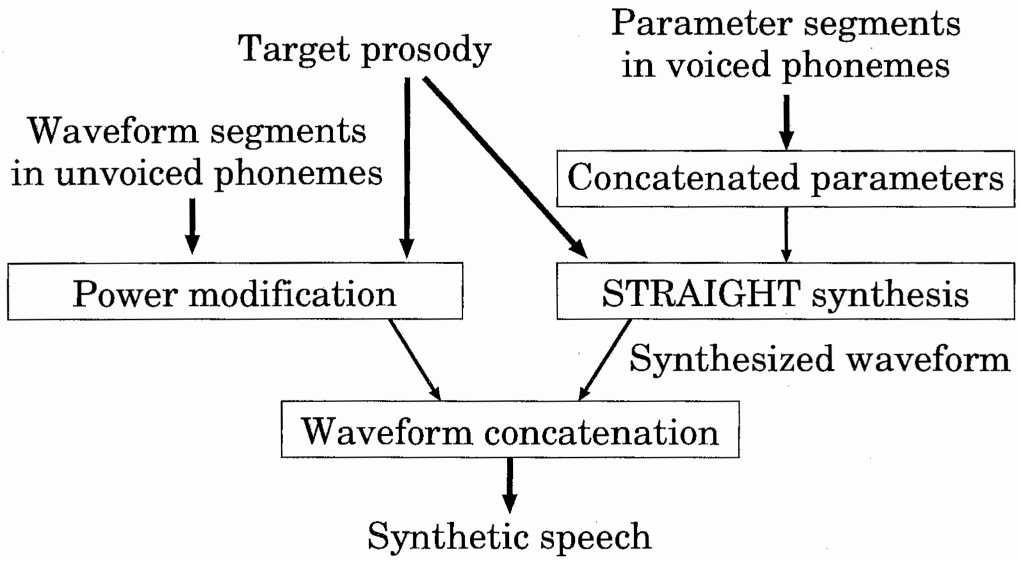

Figure 2.6: Schematic diagram of speech synthesis with prosody modification by STRAIGHT.

Chapter 3

A Segment Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme and Diphone Units

This section describes a novel segment selection algorithm for Japanese TTS systems. Since Japanese syllables consist of CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) or V, except when a vowel is devoiced, and these correspond to symbols in the Japanese 'Kana' syllabary, CV units are often used in concatenative TTS systems for Japanese. However, speech synthesized with CV units sometimes has discontinuities due to V-V or V-semivowel concatenation. In order to alleviate such discontinuities, longer units, e.g. CV units, have been proposed. However, since various vowel sequences appear frequently in Japanese, it is not realistic to prepare long units that include all possible vowel sequences. To address this problem, we propose a novel segment selection algorithm that incorporates not only phoneme units but also diphone units. The concatenation in the proposed algorithm is allowed at the vowel center as well as at the phoneme boundary. The advantage of considering both types of units is examined by experiments on concatenation of vowel sequences. Moreover, the results of perceptual evaluation experiments clarify that the proposed algorithm outperforms the conventional algorithms.*

3.1 Introduction

In Japanese, a speech corpus can be constructed efficiently by considering CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) syllables as synthesis units, since Japanese syllables consist of CV or V except when a vowel is devoiced. CV syllables correspond to symbols in the Japanese 'Kana' syllabary and the number of the syllables is small (about 100). It is also well known that transitions from C to V, or from V to V, are very important in auditory perception [70][82]. Therefore, CV units are often used in concatenative TTS systems for Japanese. On the other hand, other units are often used in TTS systems for English because the number of syllables is enormous (over 10,000) [60]. In recent years, an English TTS system based on CHATR has been adapted for diphone units by AT&T [4]. Furthermore, the NextGen TTS system based on half-phoneme units has been constructed [5][24][80], and this system has proved to be an improvement over the previous system.

In Japanese TTS, speech synthesized with CV units has discontinuities due to V-V or V-semivowel concatenation. In order to alleviate these discontinuities, Kawai et al. extended the CV unit to the CV* unit [53]. Sagisaka proposed non-uniform units to use stored speech data effectively and flexibly [70]. In this algorithm, optimum units are selected from a speech corpus to minimize the total cost calculated as the sum of some sub-costs [47][71][83]. As a result of dynamic programming search based on phoneme units, various sized sequences of phonemes are selected [8][18][42]. However, it is not realistic to construct a corpus that includes all possible vowel sequences, since various vowel sequences appear frequently in Japanese. The frequency of vowel sequences is described in **Appendix A** . If the coverage of prosody is also to be considered, the corpus becomes enormous [55]. Therefore, the concatenation between V and V is unavoidable.

Formant transitions are more stationary at vowel centers than at vowel boundaries. Therefore, concatenation at the vowel centers tends to reduce audible discontinuities compared with that at the vowel boundaries. VCV units are based on this view [73], which has been supported by our informal listening test. As typical Japanese TTS systems that utilize the concatenation at the vowel centers, TOS Drive TTS (Totally Speaker Driven Text-to-Speech) has been constructed by TOSHIBA [49] and Final Fluet has been constructed by NTT [82]. The former TTS is based on diphone units. In the latter TTS, diphone units are used if the desirable CV* units are not stored in the corpus. Thus, both TTS systems take into account only the concatenation at the vowel centers in vowel sequences. However, concatenation at the vowel boundaries is not always inferior to that at the vowel centers. Therefore, both types of concatenation should be considered in vowel sequences. In this section, we propose a novel segment selection algorithm incorporating not only phoneme units but also diphone units. The proposed algorithm permits the concatenation of synthesis units not only at the phoneme boundaries but also at the vowel centers. The results of evaluation experiments clarify that the proposed algorithm outperforms the conventional algorithms.

The section is organized as follows. In **Section 3.2**, cost functions for segment selection are described. In **Section 3.3**, the advantage of performing concatenation at the vowel centers is discussed. In **Section 3.4**, the novel segment selection algorithm is described. In **Section 3.5**, evaluation experiments are described. Finally, we summarize this section in **Section 3.6**.

3.2 Cost Function for Segment Selection

The cost function for segment selection is viewed as a mapping, as shown in **Figure 3.1**, of objective features, e.g. acoustic measures and contextual information, into a perceptual measure. A cost is considered the perceptual measure capturing the degradation of naturalness of synthetic speech. In this report, only phonetic information is used as contextual information, and the other contextual information is converted into acoustic measures by the prosody generation.

The components of the cost function should be determined based on results of perceptual experiments. A mapping of acoustic measures into a perceptual measure is generally not practical except when the acoustic measures have simple structure, as in the case of F_0 or phoneme duration. Acoustic measures with complex structure, such as spectral features that are accurate enough to capture perceptual characteristics, have not been found so far [26][59][79][88].

On the other hand, a mapping of phonetic information into perceptual measures can be determined from the results of perceptual experiments [57]. Therefore, it is possible to capture the perceptual characteristics by utilizing such a mapping. However, acoustic measures that can represent the characteristic of each segment are still necessary, since phonetic information can only evaluate the difference between phonetic categories.

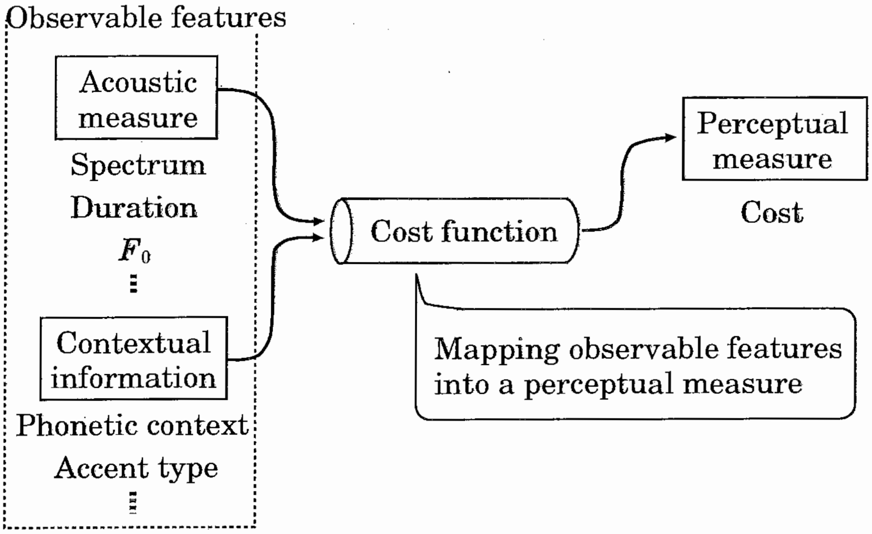

Figure 3.1: Schematic diagram of cost function.

Table 3.1: Sub-cost functions

Source information	Prosody (F_0, duration) F_0 discontinuity	C_{pro} C_{F_0}
Vocal tract information	Phonetic environment Spectral discontinuity Phonetic appropriateness	C_{env} C_{spec} C_{app}

Therefore, we utilize both acoustic measures and perceptual measures determined from the results of perceptual experiments.

3.2.1 Local cost

The local cost shows the degradation of naturalness caused by utilizing an individual candidate segment. The cost function is comprised of five sub-cost functions shown in **Table 3.1**. Each sub-cost reflects either source information or vocal tract information.

The local cost is calculated as the weighted sum of the five sub-costs. The local cost $LC(u_i, t_i)$ at a candidate segment u_i is given by

$$\begin{aligned} LC(u_i, t_i) &= w_{pro} \cdot C_{pro}(u_i, t_i) \\ &\quad + w_{F_0} \cdot C_{F_0}(u_i, u_{i-1}) \\ &\quad + w_{env} \cdot C_{env}(u_i, u_{i-1}) \\ &\quad + w_{spec} \cdot C_{spec}(u_i, u_{i-1}) \\ &\quad + w_{app} \cdot C_{app}(u_i, t_i), \qquad (3.1) \\ w_{pro} + w_{F_0} &+ w_{env} + w_{spec} + w_{app} = 1, \qquad (3.2) \end{aligned}$$

where t_i denotes a target phoneme. All sub-costs are normalized so that they have positive values with the same mean. These sub-cost functions are described in the following

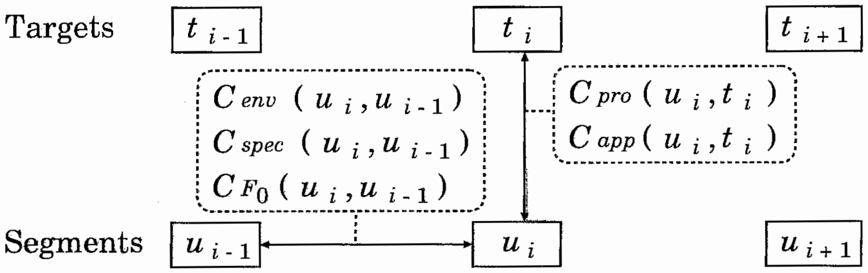

Figure 3.2: Targets and segments used to calculate each sub-cost in calculation of the cost of a candidate segment u_i for a target t_i. t_i and u_i show phonemes considered target and candidate segments, respectively.

subsections. $w_{pro}, w_{F_0}, w_{env}, w_{spec}$, and w_{app} denote the weights for individual sub-costs. In this report, these weights are equal, i.e. 0.2. The preceding segment u_{i-1} shows a candidate segment for the $(i-1)$-th target phoneme t_{i-1}. When the candidate segments u_{i-1} and u_i are connected in the corpus, concatenation between the two segments is not performed. **Figure 3.2** shows targets and segments used to calculate each sub-cost in the calculation of the cost of a candidate segment u_i for a target t_i.

3.2.2 Sub-cost on prosody: C_{pro}

This sub-cost captures the degradation of naturalness caused by the difference in prosody (F_0 contour and duration) between a candidate segment and the target.

In order to calculate the difference in the F_0 contour, a phoneme is divided into several parts, and the difference in an averaged log-scaled F_0 is calculated in each part. In each phoneme, the prosodic cost is represented as an average of the costs calculated in these parts. The sub-cost C_{pro} is given by

$$C_{pro}(u_i, t_i) = \frac{1}{M} \sum_{m=1}^{M} P(D_{F_0}(u_i, t_i, m), D_d(u_i, t_i)), \tag{3.3}$$

where $D_{F_0}(u_i, t_i, m)$ denotes the difference in the averaged log-scaled F_0 in the m-th divided part. In the unvoiced phoneme, D_{F_0} is set to 0. D_d denotes the difference in the duration, which is calculated for each phoneme and used in the calculation of the cost in each part. M denotes the number of divisions. P denotes the nonlinear function and is described in **Appendix B** .

The function P was determined from the results of perceptual experiments on the degradation of naturalness caused by prosody modification, assuming that the output speech was synthesized with prosody modification. When prosody modification is not performed, the function should be determined based on other experiments on the degradation of naturalness caused by using a different prosody from that of the target.

3.2.3 Sub-cost on F_0 discontinuity: C_{F_0}

This sub-cost captures the degradation of naturalness caused by an F_0 discontinuity at a segment boundary. The sub-cost C_{F_0} is given by

$$C_{F_0}(u_i, u_{i-1}) = P(D_{F_0}(u_i, u_{i-1}), 0), \tag{3.4}$$

where D_{F_0} denotes the difference in log-scaled F_0 at the boundary. D_{F_0} is set to 0 at the unvoiced phoneme boundary. In order to normalize a dynamic range of the sub-cost, we utilize the function P in Equation (3.3). When the segments u_{i-1} and u_i are connected in the corpus, the sub-cost becomes 0.

3.2.4 Sub-cost on phonetic environment: C_{env}

This sub-cost captures the degradation of naturalness caused by the mismatch of phonetic environments between a candidate segment and the target. The sub-cost C_{env} is given by

$$\begin{aligned} C_{env}(u_i, u_{i-1}) &= \{S_s(u_{i-1}, E_s(u_{i-1}), u_i) + S_p(u_i, E_p(u_i), u_{i-1})\}/2, & (3.5)\\ &= \{S_s(u_{i-1}, E_s(u_{i-1}), t_i) + S_p(u_i, E_p(u_i), t_{i-1})\}/2, & (3.6)\end{aligned}$$

where we turn Equation (3.5) into Equation (3.6) by considering that a phoneme for u_i is equal to a phoneme for t_i and a phoneme for u_{i-1} is equal to a phoneme for t_{i-1}. S_s denotes the sub-cost function that captures the degradation of naturalness caused by the mismatch with the succeeding environment, and S_p denotes that caused by the mismatch with the preceding environment. E_s denotes the succeeding phoneme in the corpus, while E_p denotes the preceding phoneme in the corpus. Therefore, $S_s(u_{i-1}, E_s(u_{i-1}), t_i)$ denotes the degradation caused by the mismatch with the succeeding environment in the phoneme for u_{i-1}, i.e. replacing $E_s(u_{i-1})$ with the phoneme for t_i, and $S_p(u_i, E_p(u_i), t_{i-1})$ denotes the degradation caused by the mismatch with the preceding environment in the phoneme u_i, i.e. replacing $E_p(u_i)$ with the phoneme for t_{i-1}. The sub-cost functions S_s and S_p are determined from the results of perceptual experiments described in **Appendix C** .

Even if a mismatch of phonetic environments does not occur, the sub-cost does not necessarily become 0 because this sub-cost reflects the difficulty of concatenation caused by the uncertainty of segmentation. When the segments u_{i-1} and u_i are connected in the corpus, this sub-cost is set to 0.

3.2.5 Sub-cost on spectral discontinuity: C_{spec}

This sub-cost captures the degradation of naturalness caused by the spectral discontinuity at a segment boundary. This sub-cost is calculated as the weighted sum of mel-cepstral distortion between frames of a segment and those of the preceding segment around the boundary. The sub-cost C_{spec} is given by

$$C_{spec}(u_i, u_{i-1}) = c_s \cdot \sum_{f=-w/2}^{w/2-1} h(f) MCD(u_i, u_{i-1}, f), \tag{3.7}$$

where h denotes the triangular weighting function of length w. $MCD(u_i, u_{i-1}, f)$ denotes the mel-cepstral distortion between the f-th frame from the concatenation frame ($f = 0$) of the preceding segment u_{i-1} and the f-th frame from the concatenation frame ($f = 0$) of the succeeding segment u_i in the corpus. Concatenation is performed between the -1-th frame of u_{i-1} and the 0-th frame of u_i. c_s is a coefficient to normalize the dynamic range of the sub-cost. The mel-cepstral distortion calculated in each frame-pair is given by

$$\frac{20}{\ln 10} \cdot \sqrt{2 \cdot \sum_{d=1}^{40} (mc_\alpha^{(d)} - mc_\beta^{(d)})^2}, \tag{3.8}$$

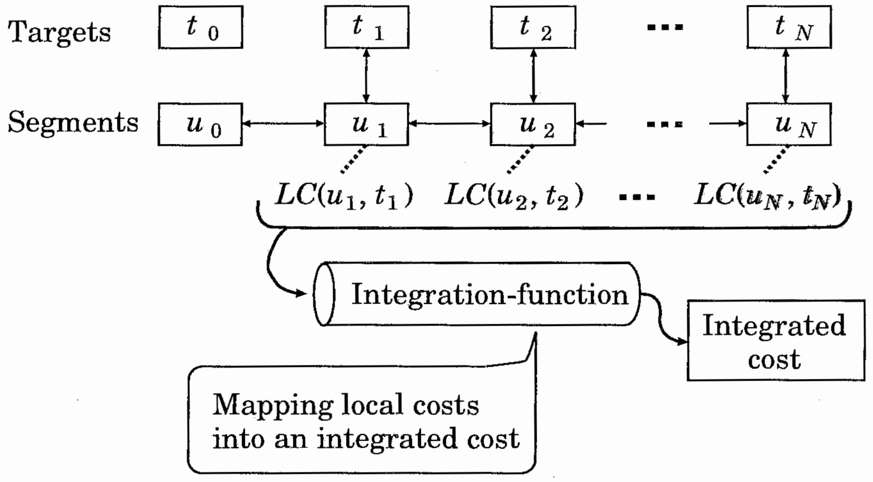

Figure 3.3: Schematic diagram of function to integrate local costs LC.

where $mc_\alpha^{(d)}$ and $mc_\beta^{(d)}$ show the d-th order mel-cepstral coefficient of a frame α and that of a frame β, respectively. Mel-cepstral coefficients are calculated from the smoothed spectrum analyzed by the STRAIGHT analysis-synthesis method [51][52]. Then, the conversion algorithm proposed by Oppenheim et al. is used to convert cepstrum into mel-cepstrum [66]. When the segments u_{i-1} and u_i are connected in the corpus, this sub-cost becomes 0.

3.2.6 Sub-cost on phonetic appropriateness: C_{app}

This sub-cost denotes the phonetic appropriateness and captures the degradation of naturalness caused by the difference in mean spectra between a candidate segment and the target. The sub-cost C_{app} is given by

$$C_{app}(u_i, t_i) = c_t \cdot MCD(CEN(u_i), CEN(t_i)), \tag{3.9}$$

where CEN denotes a mean cepstrum calculated at the frames around the phoneme center. MCD denotes the mel-cepstral distortion between the mean cepstrum of the segment u_i and that of the target t_i. c_t is a coefficient to normalize the dynamic range of the sub-cost. The mel-cepstral distortion is given by Equation (3.8). We utilize the mel-cepstrum sequence output from context-dependent HMMs in the HMM synthesis method [90] in calculating the mean cepstrum of the target $CEN(t_i)$. In this report, this sub-cost is set to 0 in the unvoiced phoneme.

3.2.7 Integrated cost

In segment selection, the optimum set of segments is selected from a speech corpus. Therefore, we integrate local costs for individual segments into a cost for a segment sequence as shown in **Figure 3.3**. This cost is defined as an integrated cost. The optimum segment sequence is selected by minimizing the integrated cost.

The average cost AC is often used as the integrated cost [8][18][21][42][80], and it is given by

$$AC = \frac{1}{N} \cdot \sum_{i=1}^{N} LC(u_i, t_i), \tag{3.10}$$

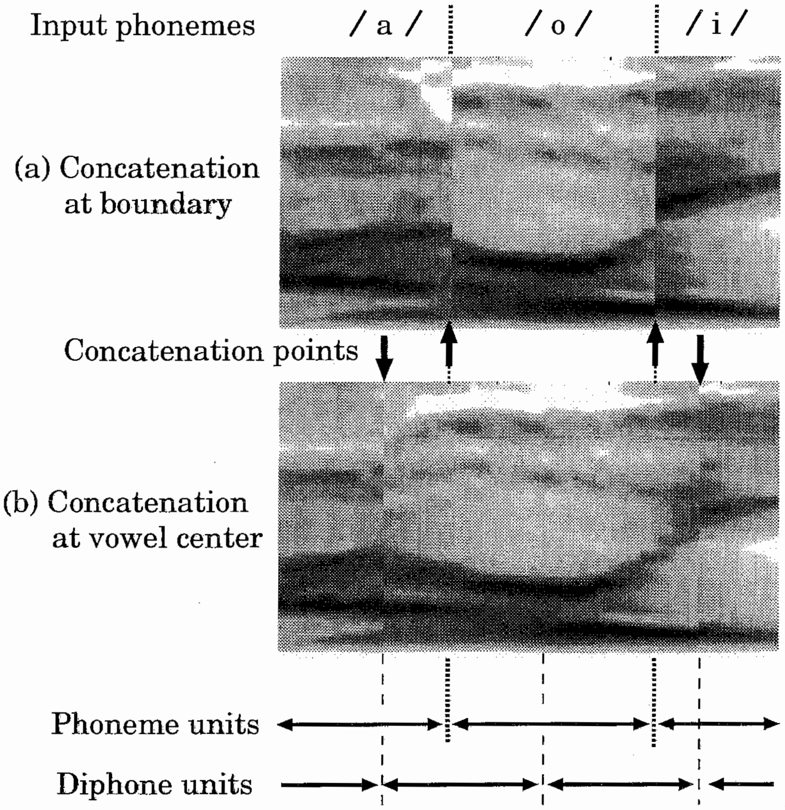

Figure 3.4: Spectrograms of vowel sequences concatenated at (a) a vowel boundary and (b) a vowel center.

where N denotes the number of targets in the utterance. t_0 (u_0) shows the pause before the utterance and t_N (u_N) shows the pause after the utterance. The sub-costs C_{pro} and C_{app} are set to 0 in the pause. Minimizing the average cost is equivalent to minimizing the sum of the local costs in the selection.

3.3 Concatenation at Vowel Center

Figure 3.4 compares spectrograms of vowel sequences concatenated at a vowel boundary and at a vowel center. At vowel boundaries, discontinuities can be observed at the concatenation points. This is because it is not easy to find a synthesis unit satisfying continuity requirements for both static and dynamic characteristics of spectral features at once in a restricted-sized speech corpus. At vowel centers, in contrast, finding a synthesis unit involves only static characteristics, because the spectral characteristics are nearly stable. Therefore, it is expected that more synthesis units reducing the spectral discontinuities can be found. As a result, the formant trajectories are continuous at the concatenation points, and their transition characteristics are well preserved.

In order to investigate the instability of spectral characteristics in the vowel, the distances of static and dynamic spectral features were calculated between centroids of individual vowels and all segments of each vowel in a corpus described in the following subsection. As the spectral feature, we used the mel-cepstrum described in **Section 3.2.5**. The results are

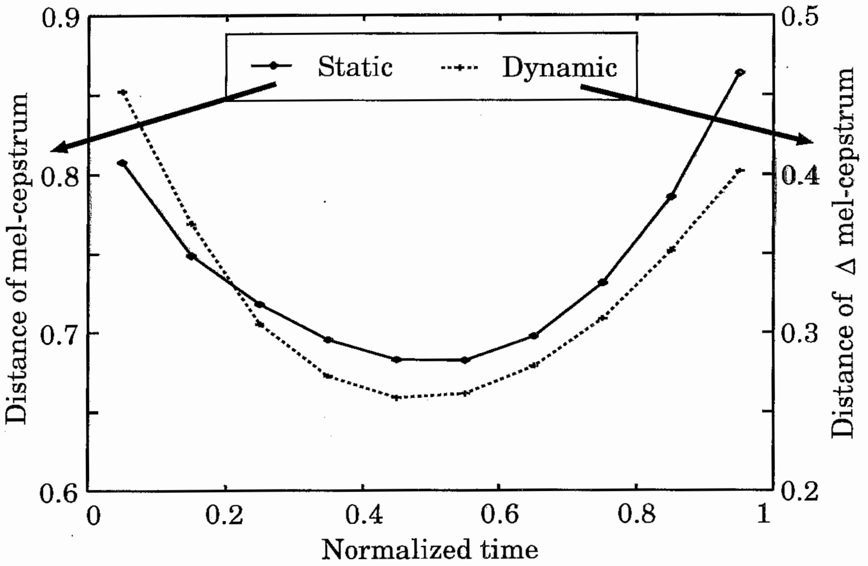

Figure 3.5: Statistical characteristics of static feature and dynamic feature of spectrum in vowels. "Normalized time" shows the time normalized from 0 (preceding phoneme boundary) to 1 (succeeding phoneme boundary) in each vowel segment.

shown in **Figure 3.5**. It is obvious that the spectral characteristics are stabler around the vowel center than those around the boundary.

From these results, it is assumed that the discontinuities caused by concatenating vowels can be reduced if the vowels are concatenated at their centers. In order to clarify this assumption, we need to investigate the effectiveness of concatenation at vowel centers in segment selection. However, it is difficult to directly show the effectiveness achieved by using the concatenation at vowel centers since various factors are considered in segment selection. Therefore, we first investigate this effectiveness in terms of spectral discontinuity, which is one of the factors considered in segment selection.

In this subsection, we compare concatenation at vowel boundaries with that at vowel centers by the mel-cepstral distortion. When a vowel sequence is generated by concatenating one vowel segment and another vowel segment, the mel-cepstral distortion caused by the concatenation at vowel boundaries and that at vowel centers are calculated. The vowel center shows a point of a half duration of each vowel segment.

3.3.1 Experimental conditions

The concatenation methods at a vowel boundary and at a vowel center are shown in **Figure 3.6**. We used a speech corpus comprising Japanese utterances of a male speaker, where segmentation was performed by experts and F_0 was revised by hand. The utterances had a duration of about 30 minutes in total (450 sentences). The sampling frequency was 16,000 Hz. The concatenation at vowel boundaries and that at vowel centers were performed by using all of the vowel sequences in the corpus. In each segment-pair, the weighted sum of the mel-cepstral distortion given by Equation (3.7) was calculated, and then the coefficient c_s was set to 1.

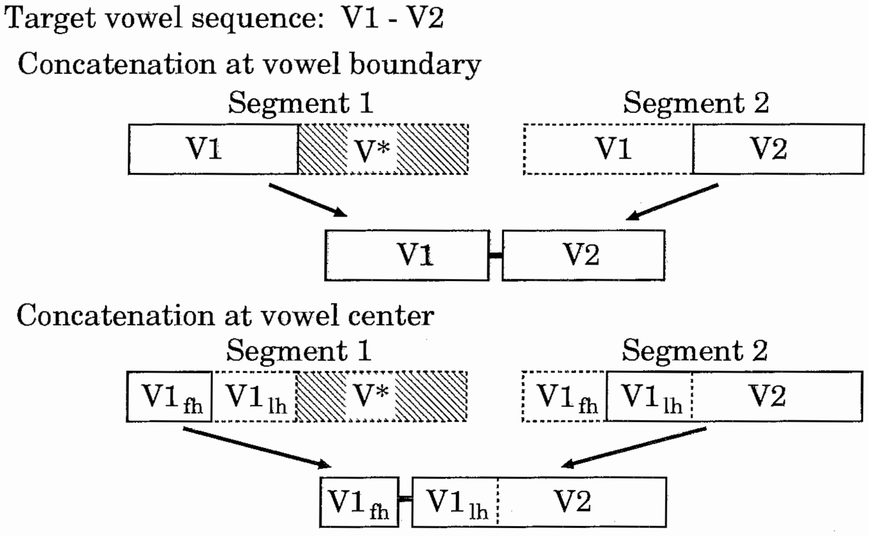
Target vowel sequence: V1 - V2
Concatenation at vowel boundary
Segment 1
Segment 2
V1
V*
V1
V2
V1
V2
Concatenation at vowel center
Segment 1
Segment 2
V1fh
V1lh
V*
V1fh
V1lh
V2
V1fh
V1lh
V2

Figure 3.6: Concatenation methods at a vowel boundary and a vowel center. "V*" shows all vowels. "V_{fh}" and "V_{lh}" show the first half-vowel and the last half-vowel, respectively.

3.3.2 Experiment allowing substitution of phonetic environment

In this experiment, substitution of phonetic environments was not prohibited. All segments of "V1" having a vowel as the succeeding phoneme in the corpus were used, i.e. "V*" in **Figure 3.6** shows all vowels.

Figure 3.7 shows frequency distribution of distortion caused by concatenation. Concatenation at vowel centers ("Vowel center") can generally reduce the discontinuity caused by the concatenation compared with that at vowel boundaries ("Vowel boundary"). In the segment selection, it is important to select the segments that can reduce not only the spectral discontinuity but also the distortion caused by various factors, e.g. prosodic distance. Therefore, as the frequency distribution shifts to the left side, the number of segments that can reduce distortion increases. From this point of view, it was found that segment selection was better when using concatenation at vowel centers along with substitution of phonetic environments.

3.3.3 Experiment prohibiting substitution of phonetic environment

In this experiment, substitution of phonetic environments was prohibited. All segments of "V1" having "V2" as the succeeding phoneme in the corpus were used, i.e. "V*" = "V2."

Figure 3.8 shows the frequency distribution of distortion caused by the concatenation between vowels that have the same phonetic environment. The distortion caused by the concatenation at vowel centers is almost equal to that at vowel boundaries. Therefore, the performance of concatenation at vowel centers is the same as that at vowel boundaries when substitution of phonetic environments is not performed.

Next, we selected the best type of concatenation in each segment-pair by allowing both vowel center and vowel boundary concatenations. Frequency distribution of distortion in this case is shown in **Figure 3.8**. This approach ("Vowel boundary & Vowel center") can reduce the discontinuity in concatenation compared with concatenation performed only at

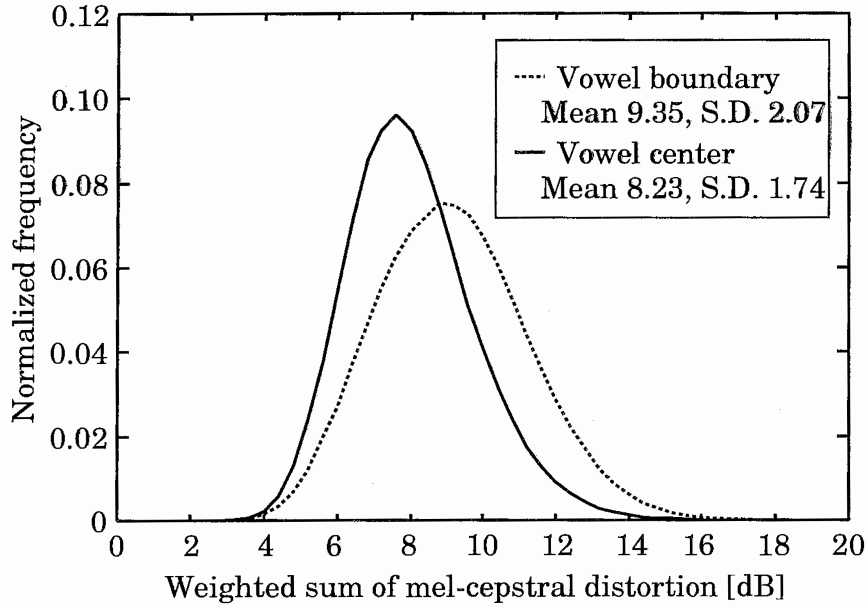

Figure 3.7: Frequency distribution of distortion caused by concatenation between vowels in the case of allowing substitution of phonetic environment. "S.D. " shows standard deviation.

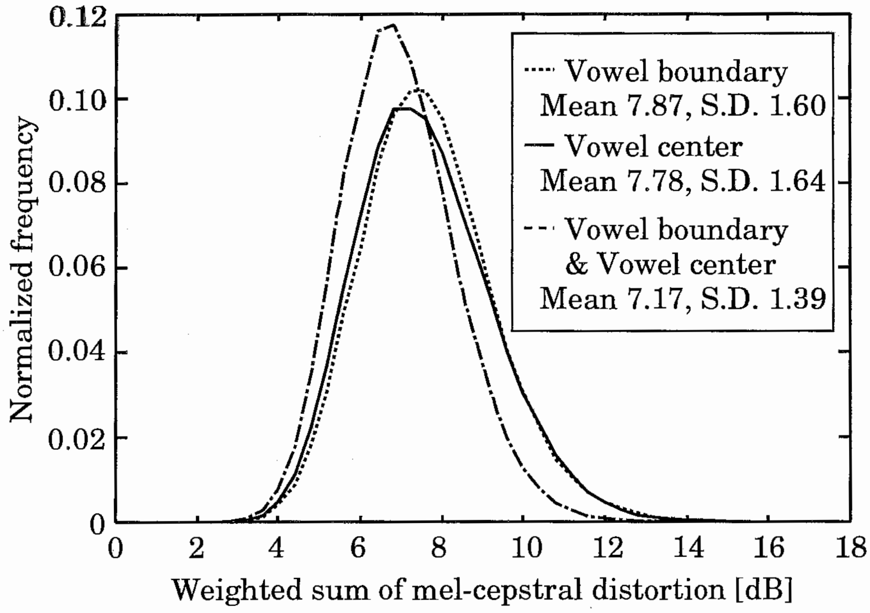

Figure 3.8: Frequency distribution of distortion caused by concatenation between vowels that have the same phonetic environment.

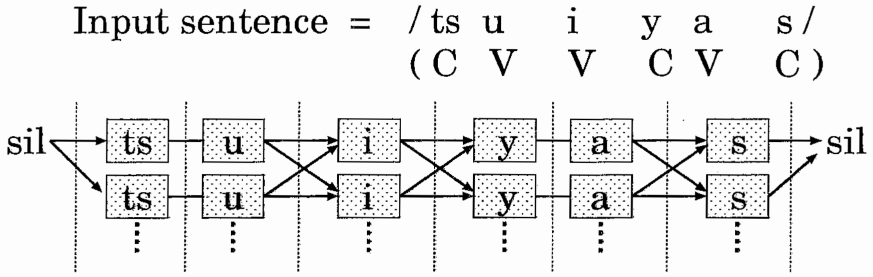

Figure 3.9: Example of segment selection based on phoneme units. The input sentence is "tsuiyas" ("spend" in English). Concatenation at C-V boundaries is prohibited.

vowel boundaries or only at vowel centers. This shows that the number of better segments increases by considering both types of concatenation.

These results clarify that better segment selection can be achieved by considering not only the concatenation at vowel centers but also that at vowel boundaries in vowel sequences.

3.4 Segment Selection Algorithm Based on Both Phoneme and Diphone Units

Motivated by the considerations in **Section 3.3**, we propose a novel segment selection algorithm based on both phoneme and diphone units. Here, we also describe the conventional segment selection algorithm based on phoneme unit for comparison with the proposed algorithm.

3.4.1 Conventional algorithm

An input sentence, i.e. a target phoneme sequence, is divided into phonemes. The local costs of candidate segments for each target phoneme are calculated by Equation (3.1). The optimum set of segments is selected from a speech corpus by minimizing the average cost given by Equation (3.10), i.e. the sum of the local costs. As a result, non-uniform units based on phoneme units can be used as synthesis units.

Figure 3.9 shows an example of the conventional segment selection based on phoneme units. In this report, we do not allow C-V concatenation since the transition from C to V is very important in auditory perception to the intelligibility in Japanese [70][82]. Therefore, the segment sequences comprised of non-uniform units based on the syllable were selected.

3.4.2 Proposed algorithm

When concatenation is allowed at the vowel centers, the half-vowel segments dividing the vowel segments are utilized to take account of diphone units. Each half-vowel segment has a half duration of the original vowel segment.

We assumed that candidate segments u_i^f and u_i^l are the first half-vowel segment of the original vowel segment u_{1i} and the last half-vowel segment of the original vowel segment u_{2i}, respectively. Sub-costs for a target phoneme t_i, which is divided into the first half-phoneme t_i^f and the last half-phoneme t_i^l, are calculated as follows:

- The C_{pro} sub-cost is calculated as the weighted sum of the sub-costs calculated at the half-vowel segments and is given by

$$w_f \cdot C_{pro}(u_i^f, t_i^f) \quad + \quad w_l \cdot C_{pro}(u_i^l, t_i^l), \tag{3.11}$$

$$w_f = \frac{dur(u_i^f)}{dur(u_i^f) + dur(u_i^l)}, \qquad w_l = \frac{dur(u_i^l)}{dur(u_i^f) + dur(u_i^l)}, \tag{3.12}$$

where the weights w_f and w_l are defined according to durations of the segments. $dur(u_i^f)$ and $dur(u_i^l)$ denote duration of the first half-vowel segment u_i^f and that of the last half-vowel segment u_i^l, respectively. In calculating the C_{pro} for the half-vowel segments, each half-vowel segment is divided into $M/2$ parts.

- The C_{F_0} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$C_{F_0}(u_i^f, u_{i-1}) + C_{F_0}(u_i^l, u_i^f). \tag{3.13}$$

- The C_{env} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$\begin{aligned} C_{env}(u_i^f, u_{i-1}) \quad + \quad & C_{env}(u_i^l, u_i^f) \\ = \quad & C_{env}(u_{1i}, u_{i-1}) + \{S_s^d(u_{1i}, E_s(u_{1i}), t_{i+1}) \\ & + S_p^d(u_{2i}, E_p(u_{2i}), t_{i-1})\}/2, \end{aligned} \tag{3.14}$$

where the phonetic environments of the half-vowel segment are equal to those of the original vowel segment divided into the half-vowel segments. On the other hand, the sub-cost functions, S_s^d and S_p^d, for the concatenation at vowel centers are not equal to the sub-cost functions, S_s and S_p, for the concatenation at phoneme boundaries, which are given by Equation (3.6).

- The C_{spec} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$C_{spec}(u_i^f, u_{i-1}) + C_{spec}(u_i^l, u_i^f). \tag{3.15}$$

- The C_{app} sub-cost is calculated as the weighted sum of the sub-costs calculated at the original vowel segments, u_{1i} and u_{2i} divided into the half-vowel segments, u_i^l and u_i^f, and is given by

$$w_f \cdot C_{app}(u_{1i}, t_i) \quad + \quad w_l \cdot C_{app}(u_{2i}, t_i), \tag{3.16}$$

where the weights w_f and w_l are given by Equation (3.12).

Figure 3.10 shows targets and segments used to calculate each sub-cost in the calculation of the cost of candidate segments u_i^f, u_i^l for a target t_i. If a diphone unit is used, the sub-costs, $C_{env}(u_i^f, u_{i-1})$, $C_{spec}(u_i^f, u_{i-1})$, and $C_{F_0}(u_i^f, u_{i-1})$, become 0, since u_{i-1} and u_i^f are connected in the corpus. Furthermore, if a phoneme unit is used, the sub-costs, $C_{env}(u_i^l, u_i^f)$, $C_{spec}(u_i^l, u_i^f)$, and $C_{F_0}(u_i^l, u_i^f)$, also become 0, since u_i^f and u_i^l are connected in the corpus.

In the other phonemes where concatenations are not allowed at their centers, the costs are calculated in the same way as the conventional algorithm. The optimum segment sequence

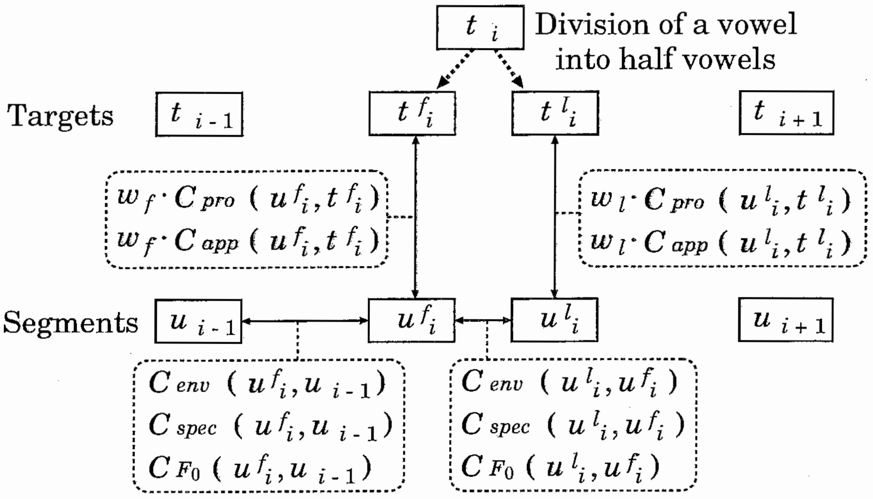
t_i
Division of a vowel into half vowels
Targets
t_{i-1}
t^f_i
t^l_i
t_{i+1}
$w_f \cdot C_{pro}(u^f_i, t^f_i)$
$w_f \cdot C_{app}(u^f_i, t^f_i)$
$w_l \cdot C_{pro}(u^l_i, t^l_i)$
$w_l \cdot C_{app}(u^l_i, t^l_i)$
Segments
u_{i-1}
u^f_i
u^l_i
u_{i+1}
$C_{env}(u^f_i, u_{i-1})$
$C_{spec}(u^f_i, u_{i-1})$
$C_{F0}(u^f_i, u_{i-1})$
$C_{env}(u^l_i, u^f_i)$
$C_{spec}(u^l_i, u^f_i)$
$C_{F0}(u^l_i, u^f_i)$

Figure 3.10: Targets and segments used to calculate each sub-cost in calculation of the cost of candidate segments u^f_i, u^l_i for a target t_i.

is selected from a speech corpus by minimizing the average cost. As a result, non-uniform units based on both phoneme units and diphone units can be used as synthesis units.

Phoneme units and diphone units have their own advantages and disadvantages. The advantage of using phoneme units is the ability to preserve the characteristics of phonemes. Moreover, it might be assumed that slight spectral discontinuities in the transitions in which spectral features change rapidly are hard to perceive. However, it is not easy to find synthesis units that can reduce the spectral discontinuities since static and dynamic characteristics of spectral features should be considered in the transitions. On the other hand, the advantage of using diphone units is the ability to preserve transitions between phonemes and to concatenate at steady parts in phonemes. Therefore, more synthesis units reducing the spectral discontinuities can be found. However, it might be assumed that spectral discontinuities in steady parts are easy to perceive. In the proposed algorithm, the cost is used to determine which of the two units is better.

We allow concatenations at vowel centers not only in transitions from V to V but also in transitions from V to a semivowel or a nasal. In the transitions from V to a semivowel or a nasal, diphone units that start from the center of a vowel in front of consonants are used. In this report, the half-vowel segments are not used except for the segments having silences as phonetic environments. Therefore, minimum units preserve either the important transitions between phonemes or the characteristics of Japanese syllables.

An example of the proposed segment selection algorithm is shown in **Figure 3.11**. Diphone units such as /ts-u/, /u-i/, and /i-y/ as well as phoneme units are considered in the segment selection.

3.4.3 Comparison with segment selection based on half-phoneme units

Our TTS under development is mainly for Japanese speech synthesis. Since the number of Japanese syllables is much smaller than that of English syllables, we can construct a speech corpus containing all of the syllables. Therefore, we restrict minimum synthesis units to

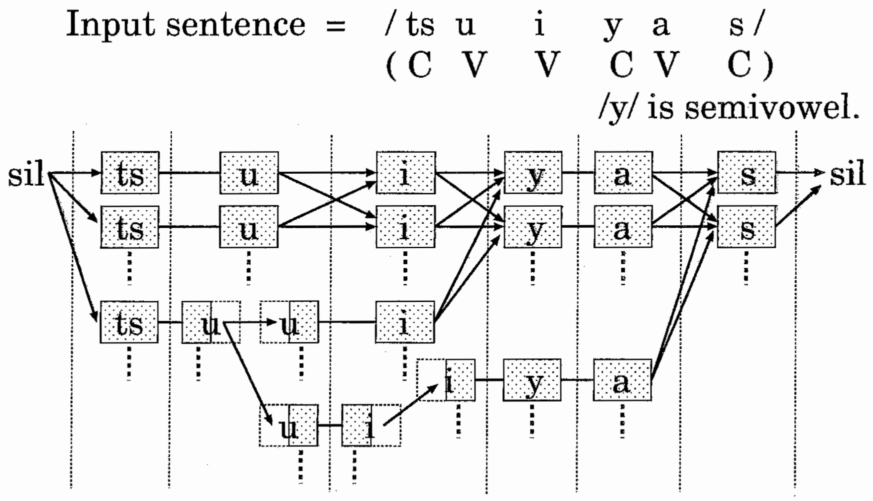

Figure 3.11: Example of segment selection based on phoneme units and diphone units. Concatenation at C-V boundaries and selection of isolated half-vowels are prohibited.

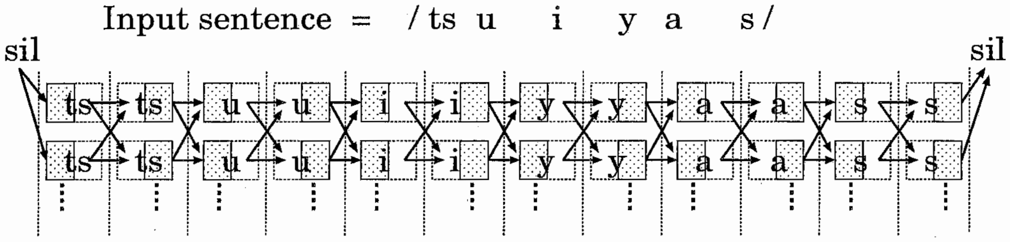

Figure 3.12: Example of segment selection based on half-phoneme units.

syllables or diphones in order to preserve important transitions. Namely, we don't use half-phoneme units as used in the AT&T NextGen TTS system [24]. An example of a segment selection algorithm based on half-phoneme units is shown in **Figure 3.12**.

The proposed algorithm can be considered an algorithm based on half-phoneme units adapted for Japanese speech synthesis by restricting some types of concatenation. However, it might be assumed that half-phoneme units would also work well for Japanese speech synthesis. Therefore, we compared the proposed algorithm with the conventional algorithm based on half-phoneme units. In order to make a fair comparison, the weight w_{env} in Equation (3.1) was set to 0 since we have not yet determined S_s and S_p in sub-cost C_{env} for almost all of the half-phoneme units. As a result of a preference test on the naturalness of synthetic speech, the 95% confidence interval of the proposed algorithm's preference score was 66.67 ± 3.98%. This result shows that as the length of units used in segment selection becomes short, the risk of causing more audible discontinuities by excessive concatenations becomes high, although more combinations of units can be considered to reduce the prosodic difference. Therefore, it is important to use a cost that is accurate enough to capture the audible discontinuities, especially in segment selection based on short units.

Table 3.2: Number of concatenations in experiment comparing proposed algorithm with segment selection based on phoneme units. "S" and "N" show semivowel and nasal. "Center" shows concatenation at vowel center

Concatenation	V-C	V-V	V-S	V-N	Center
Proposed algorithm	125	6	3	11	25
Conventional algorithm	124	16	3	20	-

Table 3.3: Number of concatenations in experiment comparing proposed algorithm with segment selection allowing only concatenation at vowel center in V-V, V-S, and V-N sequences

Concatenation	V-C	V-V	V-S	V-N	Center
Proposed algorithm	137	10	6	23	22
Conventional algorithm	143	-	-	-	62

3.5 Experimental Evaluation

In order to evaluate the performance of the proposed algorithm, we compared the proposed algorithm with the conventional algorithm, which allows concatenation only at phoneme boundaries. Moreover, we also compared the proposed algorithm with another conventional algorithm, which allows concatenation only at vowel centers in V-V, V-S, and V-N sequences. We call the former comparison Experiment A and the latter comparison Experiment B.

3.5.1 Experimental conditions

We used the speech corpus comprising Japanese utterances of a male speaker, which is described in **Section 3.3.1**.

A preference test was performed with synthesized speech of 10 Japanese sentences. The sentences used in Experiment A were different form those in Experiment B. These sentences were not from the speech corpus used in the segment selection. The speech was synthesized by the proposed segment selection algorithm or the conventional algorithm in each experiment. The natural prosody and the mel-cepstrum sequences extracted from the original utterances were used to investigate the performance of the segment selection algorithms. In Experiment A, all of the synthesized speech were comprised of 366 phonemes, and the number of concatenations in each algorithm is shown in **Table 3.2**. In the other experiment, Experiment B, all of the synthesized speech were comprised of 453 phonemes, and the number of concatenations in each algorithm is shown in **Table 3.3**.

The speech was synthesized with prosody (F_0 contour, duration, and power) modification by using STRAIGHT. Ten listeners participated in the experiment. In each trial, a pair of utterances synthesized with the proposed algorithm and the conventional algorithm was presented in random order, and the listeners were asked to choose either of the two types of synthetic utterances as sounding more natural.

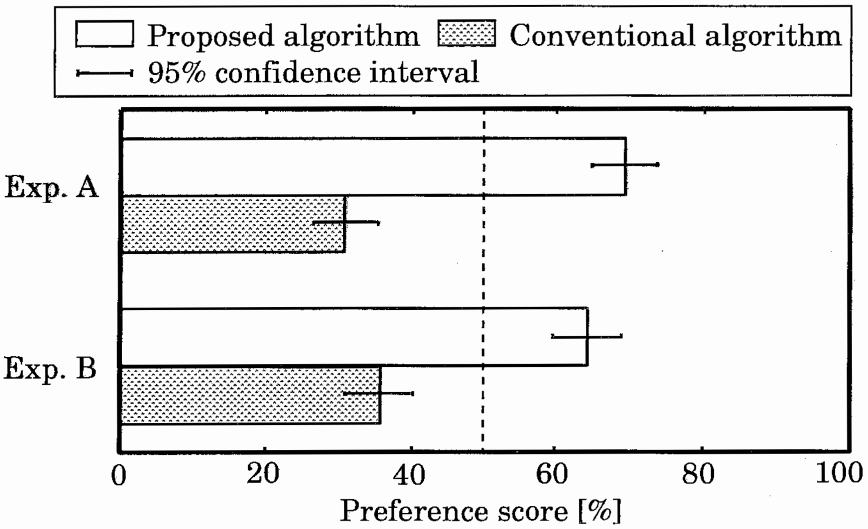

Figure 3.13: Results of comparison with the segment selection based on phoneme units ("Exp. A") and those of comparison with the segment selection allowing only concatenation at vowel center in V-V, V-S, and V-N sequences ("Exp. B").

3.5.2 Experimental results

Results of Experiment A and those of Experiment B are shown in **Figure 3.13**. The preference score of the proposed algorithm in Experiment A was 69.25%, and that in Experiment B was 64.25%. In both experiments, the preference scores of the proposed algorithm exceeded 50% by a large margin. These results demonstrate that the proposed algorithm can synthesize speech more naturally than the conventional algorithms.

3.6 Summary

In this section, we proposed a novel segment selection algorithm for Japanese speech synthesis with both phoneme and diphone units to avoid the degradation of naturalness caused by concatenation at perceptually important transitions between phonemes. In the proposed algorithm, non-uniform units allowing concatenation not only at phoneme boundaries but also at vowel centers can be selected from a speech corpus. The experimental results of concatenation of vowel sequences clarified that better segments reducing the spectral discontinuities increases by considering both types of concatenation. We also performed perceptual experiments. The results showed that speech synthesized with the proposed algorithm has better naturalness than that of the conventional algorithms.

Although we restrict minimum synthesis units to syllables or diphones, these units are not always the best units. The best unit definition is expected to be determined according to various factors, e.g. corpus size, correspondence of the cost to perceptual characteristics, synthesis methods, and the kinds of languages. Consequently, we need to investigate various units based on this view. Moreover, we need to clarify the effectiveness of searching optimal concatenation frames [23] after segment selection, although we have no acoustic measure that is accurate enough to capture perceptual characteristics.

Chapter 4

An Evaluation of Cost Capturing Both Total and Local Degradation of Naturalness for Segment Selection

In segment selection for concatenative TTS, it is important to utilize a cost that corresponds to the perceptual characteristics. In this section, we clarify correspondence of the cost to the perceptual scores and then evaluate various functions to integrate the costs. The perceptual scores are determined from results of perceptual experiments on the naturalness of synthetic speech. The results show that the average cost, which shows the degradation of naturalness over the entire synthetic speech, has better correspondence to the perceptual scores than the maximum cost, which shows the local degradation of naturalness. Furthermore, it is shown that RMS (Root Mean Square) cost, which is affected by both the average cost and the maximum cost, has the best correspondence. We also clarify that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost. Then, we investigate the effect of using the RMS cost for segment selection. From the results of experiments comparing this approach with segment selection based on the conventional average cost, it is found that (1) in segment selection based on the RMS cost a larger number of concatenations causing slight local degradation are performed to avoid concatenations causing greater local degradation, and (2) the effect of the RMS cost has little dependence on the size of the corpus.

4.1 Introduction

In corpus-based concatenative TTS, speech synthesis based on segment selection has recently become the focus of much work on synthesis [20][80]. In segment selection, the optimum set of segments is selected from a speech corpus by minimizing the integrated cost for a segment sequence, which is described in **Section 3.2.7**. Therefore, it is important to utilize a cost that corresponds to the perceptual characteristics to synthesize speech naturally [61][87]. However, such a cost has not been found so far [26][59][81][88]. To realize TTS with high quality and robustness, it is necessary to improve this cost function [21][67].

In the design process of the cost function, however, it is doubtful whether this correspondence is preserved, since there are some approximations, e.g. utilization of acoustic measures that are not accurate enough to capture perceptual characteristics [57][79], and independence among various factors. Moreover, it might be assumed that local degradation of naturalness would have a great effect on the naturalness of synthetic speech, although the average cost, which shows the degradation of naturalness over the entire synthetic speech, is often used

[8][18][42]. Therefore, a direct investigation of the relationship between a perceptual measure and the cost is worthwhile.

Here, we clarify the correspondence of our cost described in **Section 3.2** to the perceptual scores. Then various functions to integrate the costs are evaluated in terms of correspondence to the MOS (Mean Opinion Score) determined from the results of perceptual experiments on the naturalness of synthetic speech. As a result, we show that the RMS (Root Mean Square) cost, which is affected by both the average cost and the maximum cost showing the local degradation of naturalness, has the best correspondence to the perceptual scores. We also clarify that the naturalness of synthetic speech can be slightly improved by using the RMS cost in segment selection.

In order to investigate the effect of considering not only the total degradation of naturalness of synthetic speech but also the local degradation in segment selection, we compare segment selection based on the RMS cost with that based on the average cost. Selected segment sequences are analyzed from various points of view to clarify how the local degradation of naturalness can be alleviated by utilizing the RMS cost. We also clarify the relationship between the effectiveness of the RMS cost and the size of the corpus.

This section is organized as follows. In **Section 4.2**, various integrated costs for the segment selection are described. In **Section 4.3**, we present perceptual evaluations of the costs. In **Section 4.4**, the effectiveness of utilizing the RMS cost in segment selection is discussed. Finally, we summarize this section in **Section 4.5**.

4.2 Various Integrated Costs

In the conventional segment selection [8][18][21][42][80], the optimum set of segments is selected from a speech corpus by minimizing the average cost AC given by Equation (3.10). The average cost shows the degradation of naturalness over the entire synthetic utterance. Therefore, a segment with a large cost can be included in the output sequence of segments even if it is optimal in view of the average cost.

It might be assumed that the largest cost in the sequence, i.e. the local degradation of naturalness, would have much effect on the degradation of naturalness in synthetic speech. To investigate this issue, let us define the maximum cost MC as the integrated cost given by

$$MC = \max_i \{WC(u_i, t_i)\}, \quad 1 \leq i \leq N, \tag{4.1}$$

where N denotes the number of targets in the utterance.

In order to evaluate the various integrated costs, we utilize the norm cost, NC_p, given by

$$NC_p = \left[\frac{1}{N} \cdot \sum_{i=1}^{N} \{WC(u_i, t_i)\}^p \right]^{\frac{1}{p}}, \tag{4.2}$$

where p denotes a power coefficient. When p is set to 1, the norm cost is equal to the average cost. When p is set to infinity, the norm cost is equal to the maximum cost. Thus, this norm cost takes into account both the mean value and the maximum value by varying the power coefficient. In the following subsection, we find an optimum value of the power coefficient from the results of perceptual experiments.

4.3 Perceptual Evaluation of Cost

4.3.1 Correspondence of cost to perceptual score

We performed an opinion test on the naturalness of the synthetic speech. In order to select a proper set of test stimuli, a large number of utterances were synthesized by varying the corpus size from 0.5 to 32 hours (0.5, 0.7, 1, 1.4, 2, 2.8, 4, 5.7, 8, 11.3, 16, 22.6, 32). Each utterance consisted of a part of a sentence that was divided by a pause. We synthesized 14,926 utterances that were not included in the corpus, from which we selected a set of 140 stimuli so that the set covers a wide field in terms of both average cost and maximum cost. This selection was performed under the restriction that the number of phonemes in an utterance, the duration of an utterance, and the number of concatenations are roughly equal among the selected stimuli. The distribution of the average cost and the maximum cost for all synthetic utterances and selected test stimuli are shown in **Figure 4.1** and **Figure 4.2**, respectively.

Natural prosody and the mel-cepstrum sequences extracted from the original utterances were used as input information for segment selection. In order to alleviate audible discontinuity at the boundary between vowel and voiced phoneme, concatenation at the preceding vowel center is also allowed in the segment selection. In the waveform synthesis, signal processing for prosody modification was not performed except for power control. Therefore, we should use a different sub-cost on prosody C_{pro} from that described in **Section 3.2.2**, where the function P has been determined in the case of performing prosody modification. In the case of not performing prosody modification, however, we have not determined any function P. Therefore, we approximate C_{pro} by the function P described in **Appendix C** .

Eight listeners participated in the experiment. They evaluated the naturalness on a scale of seven levels, namely 1 (very bad) to 7 (very good). These levels were determined by each listener, so the scores were distributed widely among the stimuli. The perceptual score, here the MOS, was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

Figure 4.3 shows the correlation coefficient between the norm cost and the perceptual score as a function of the power coefficient. The average cost ($p = 1$) has better correspondence to the perceptual scores (correlation coefficient = -0.808) than does the maximum cost (correlation coefficient = -0.685). Therefore, the naturalness of synthetic speech is better estimated by the degradation of naturalness over the entire synthetic utterance than by using only the local degradation of naturalness. **Figure 4.4** shows the correlation between the average cost and the perceptual score, and **Figure 4.5** shows the correlation between the maximum cost and the perceptual score.

Moreover, when the power coefficient is set to 2, the norm cost, called the Root Mean Square (RMS) cost, has the best correspondence to the perceptual scores (correlation coefficient = -0.840). The absolute value of the correlation coefficient in the case of the RMS cost is statistically larger than those in the cases of the average cost ($t = 2.4696, df = 137, p < 0.05$). Therefore, the naturalness of synthetic speech is better estimated by considering both the degradation of naturalness over the entire synthetic speech and the local degradation of naturalness, since the RMS cost is affected by both types of degradation. **Figure 4.6** shows the correlation between the RMS cost and the perceptual score.

In order to estimate the perceptual scores more accurately, we also performed multiple linear regression analysis by utilizing the norm costs while varying the power coefficient from 1 to 10 and the maximum cost as predictor variables. As a result, the correspondence to the perceptual scores was not improved statistically (multiple correlation coefficient =

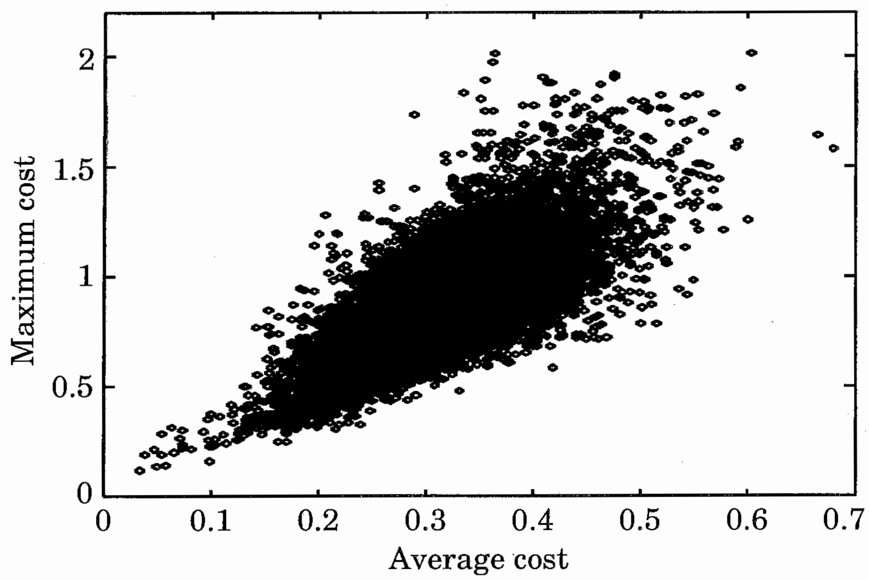

Figure 4.1: Distribution of average cost and maximum cost for all synthetic utterances.

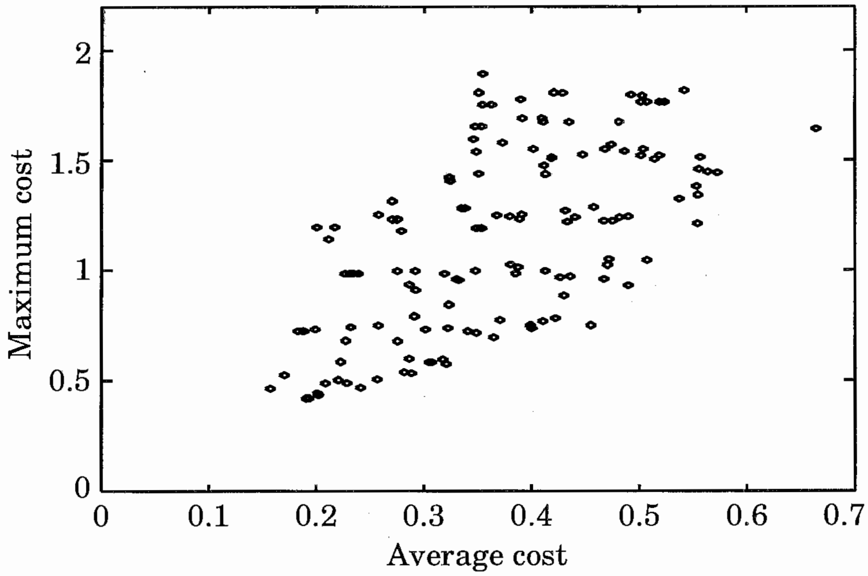

Figure 4.2: Scatter chart of selected test stimuli.

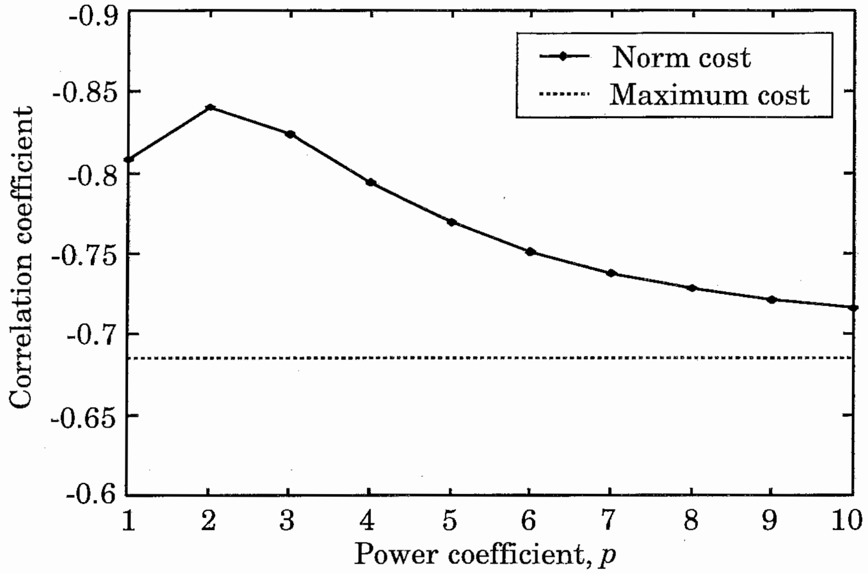

Figure 4.3: Correlation coefficient between norm cost and perceptual score as a function of power coefficient, p.

0.846) compared with the correspondence of the RMS cost.

Although the RMS cost is the best integrated cost in this experiment, it is expected that the best power coefficient depends on the correspondence of the local cost to perceptual characteristics. It is worth noting that the naturalness of synthetic speech is better estimated by using both the total degradation of naturalness and the local degradation than by using only the total degradation.

4.3.2 Preference test on naturalness of synthetic speech

Figure 4.7 shows an example of local costs of a segment sequence selected by the conventional average cost and that of another segment sequence selected by the novel RMS cost. Some large local costs surrounded by circles are shown in the case of the average cost. On the other hand, such large local costs are alleviated in the case of the RMS cost. In order to clarify which of the two costs could select the best segment sequence, we performed a preference test on the naturalness of synthetic speech. The corpus size was 32 hours, and utterances used as test stimuli were not included in the corpus. Natural prosody and the mel-cepstrum sequences extracted from the original utterances were used as input information for segment selection. Signal processing for prosody modification was not performed except for power control.

The naturalness of synthetic speech was expected to be nearly equal between segment sequences having similar costs. Therefore, we used pairs of segment sequences that had greater cost differences in the test. Each pair is comprised of the segment sequence selected by the RMS cost and that by the average cost. In order to fairly compare the performances of these two costs, we selected the pairs with the larger differences in the average cost as well as those with the larger differences in the RMS cost. A scatter chart of the test stimuli is shown in **Figure 4.8**. The difference in the RMS cost and that in the average cost were

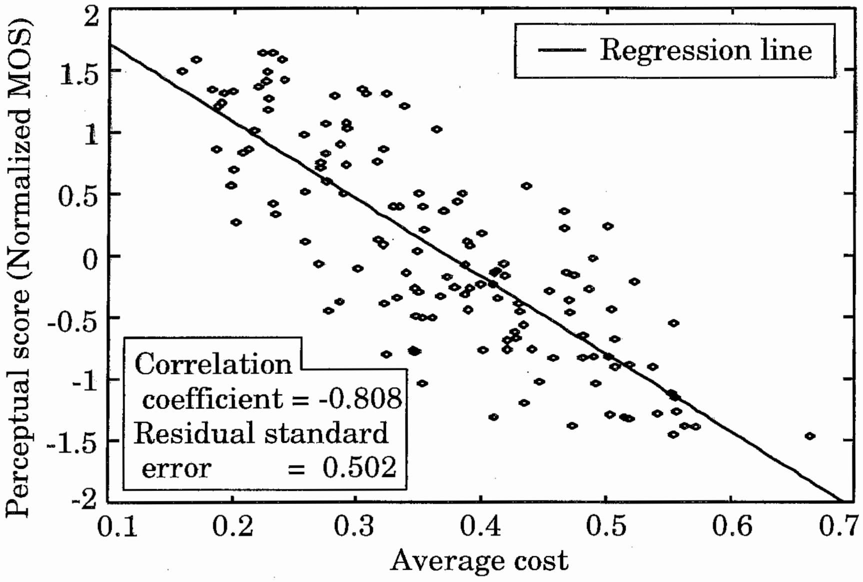

Figure 4.4: Correlation between average cost and perceptual score.

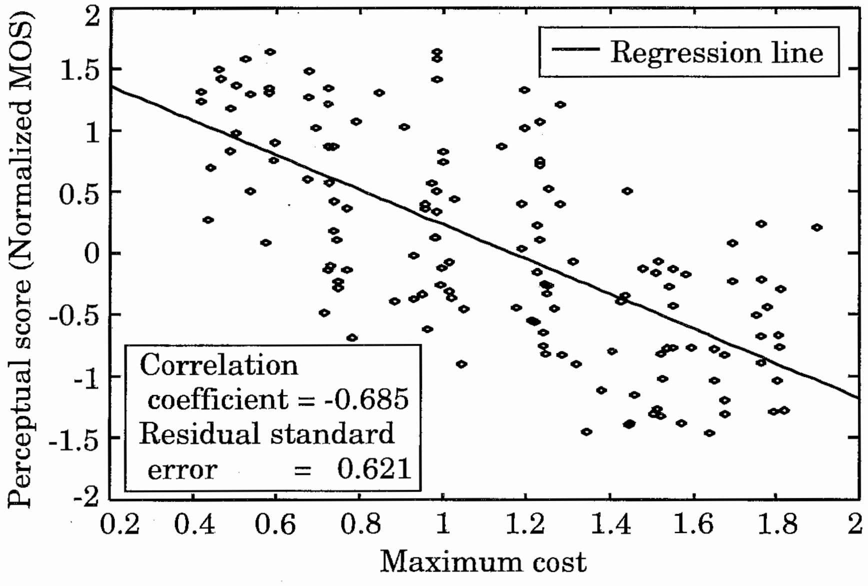

Figure 4.5: Correlation between maximum cost and perceptual score.

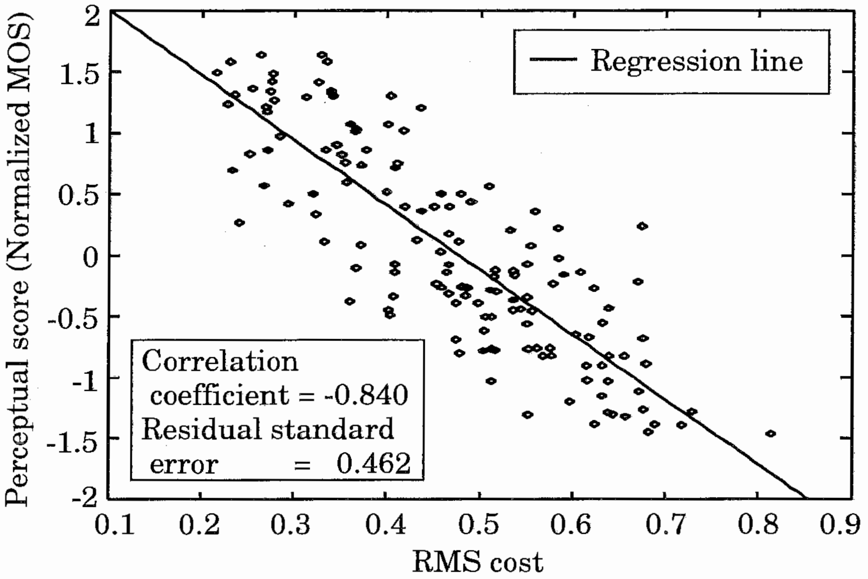

Figure 4.6: Correlation between RMS cost and perceptual score. The RMS cost can be converted into a perceptual score by utilizing the regression line.

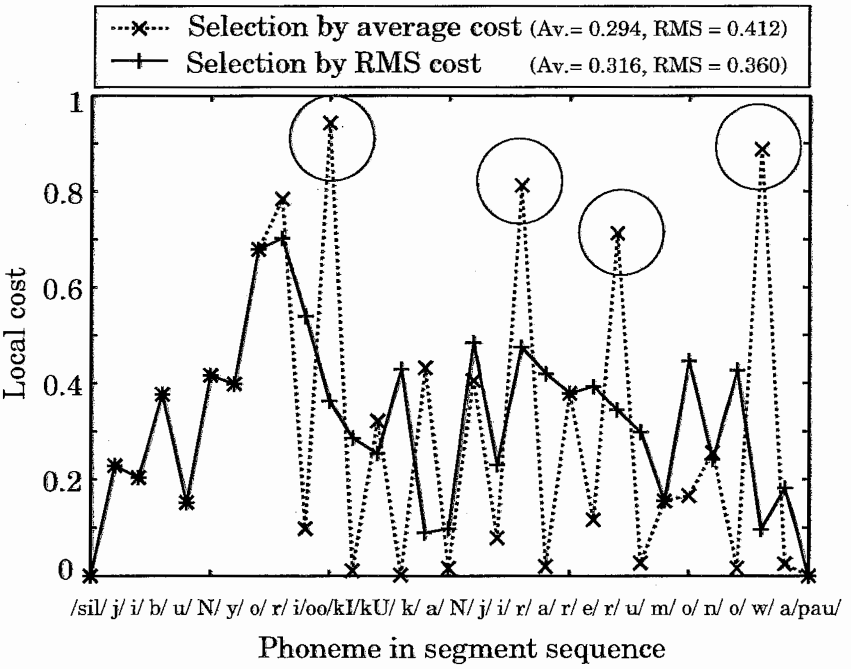

Figure 4.7: Examples of local costs of segment sequences selected by the average costs and by the RMS cost. "Av." and "RMS" show the average and the root mean square of local costs, respectively.

calculated as follows:

$$RMSC_{AC_{sel}} - RMSC_{RMSC_{sel}}, \tag{4.3}$$

$$AC_{AC_{sel}} - AC_{RMSC_{sel}}, \tag{4.4}$$

where $RMSC_{AC_{sel}}$ and $AC_{AC_{sel}}$ denote the RMS cost and the average cost of the segment sequence selected by minimizing the average cost, respectively. $RMSC_{RMSC_{sel}}$ and $AC_{RMSC_{sel}}$ denote the RMS cost and the average cost of the segment sequence selected by minimizing the RMS cost, respectively. "Sub-set A" includes stimulus pairs with larger differences in the RMS cost. On the other hand, "sub-set B" includes stimulus pairs with larger differences in average cost. There were 20 stimulus pairs in each sub-set, and the total number of stimulus pairs was 35, since 5 pairs were included in both sub-sets.

Eight Japanese listeners participated in the experiment. In each trial, synthetic speech by the segment selection based on the average cost and that by the segment selection based on the RMS cost were presented in random order, and listeners were asked to choose either of the two types of synthetic speech as sounding more natural.

The results in **Figure 4.9** show that the segment selection based on the RMS cost can synthesize speech more naturally than that based on the average cost in all cases: utilizing all stimuli, stimuli in sub-set A only, and stimuli in sub-set B only. However, this improvement is only slight.

4.3.3 Correspondence of RMS cost to perceptual score in lower range of RMS cost

We clarified the correspondence of the RMS cost to the perceptual scores when the size of the corpus was varied in **Section 4.3.1**. However, our TTS system utilizes a large-sized corpus that includes many segments with high coverage on both phonetic environment and prosody to synthesize speech more naturally and consistently. In the case of a large-sized corpus, the RMS costs are expected to be distributed not in a wide range but in a lower range, since segments causing only a slight degradation of naturalness can usually be selected. Thus, it is worthwhile to investigate the correspondence to the perceptual scores in a range of lower RMS costs.

We performed an opinion test on the naturalness of the synthetic speech to clarify the correspondence of the RMS cost to the perceptual scores in a lower range. Test stimuli were included in the region covered in utilizing the 32-hour corpus, in which the RMS costs were less than 0.4. They were selected from a large number of utterances synthesized by varying the corpus size. This selection was performed under the restriction that the number of phonemes in an utterance, the duration of an utterance, and the number of concatenations were roughly equal among the selected stimuli. The number of selected stimuli was 160. Eight Japanese listeners participated in the experiment. They evaluated the naturalness on a scale of seven levels. These levels were determined by each listener, so the scores were distributed widely among the stimuli. The perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The correspondence of the RMS cost to the perceptual scores is shown in **Figure 4.10**. The correspondence is much worse (correlation coefficient = −0.400) than that in the case of utilizing stimuli that cover a wide range of the cost (correlation coefficient = −0.840). Therefore, it is obvious that the correspondence of the RMS cost is inadequate and that we should improve the cost function.

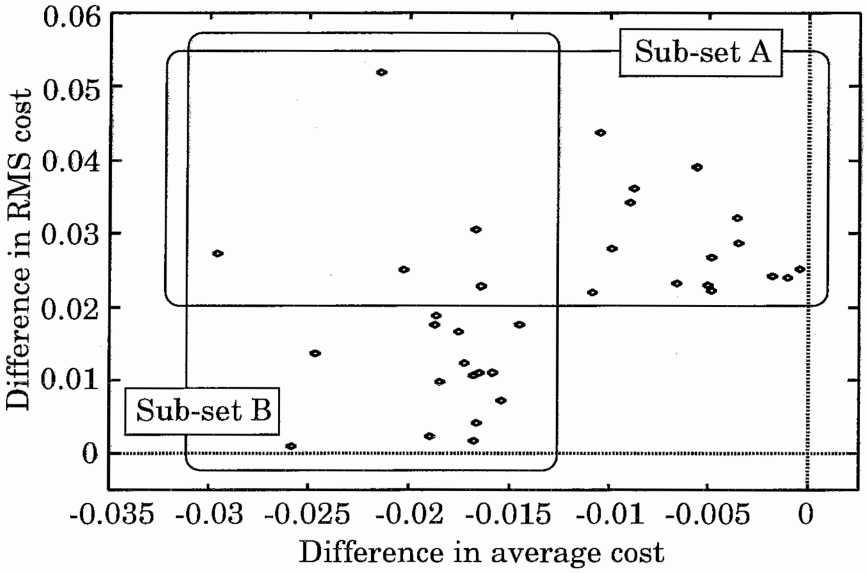

Figure 4.8: Scatter chart of selected test stimuli. Each dot denotes a stimulus pair.

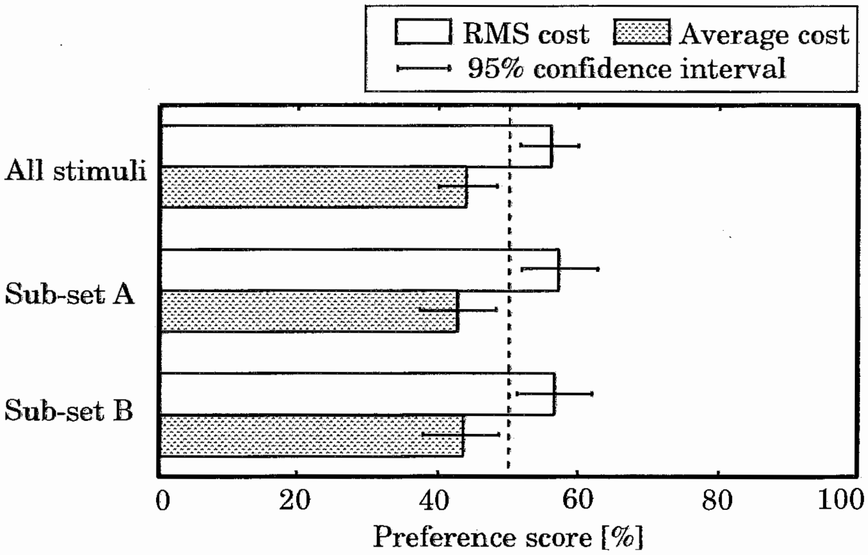

Figure 4.9: Preference score.

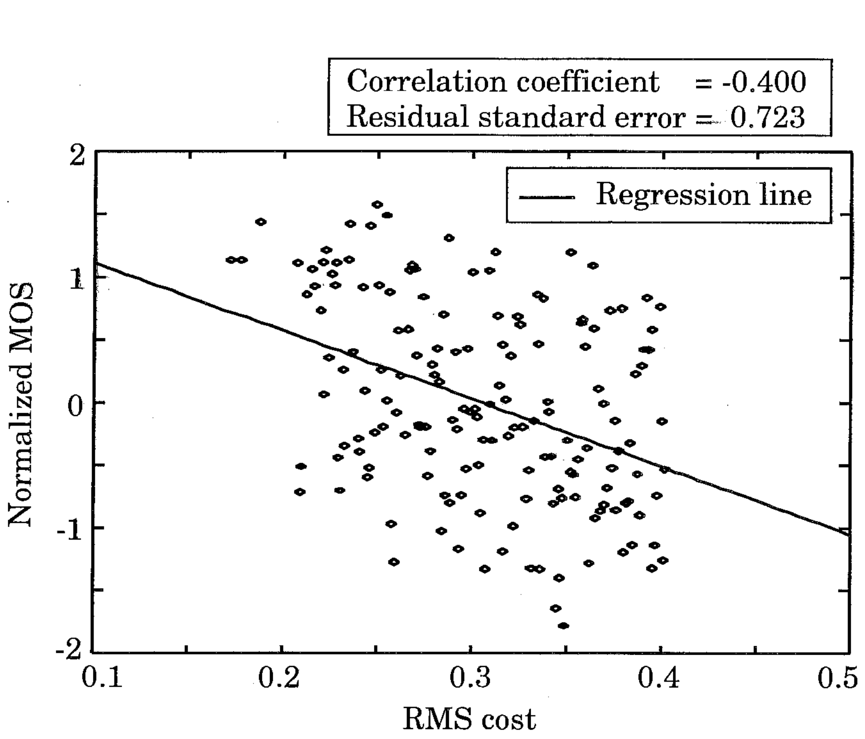

Figure 4.10: Correlation between RMS cost and perceptual score in lower range of RMS cost.

4.4 Segment Selection Considering Both Total and Local Degradation of Naturalness

In the conventional segment selection based on the average cost, the optimum segment sequence is selected while taking into account only the total degradation. In order to consider not only the total degradation but also the local degradation, we incorporated the RMS cost in segment selection. In this selection, the RMS cost $RMSC$ is minimized and given by

$$RMSC = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \{LC_i(u_i, t_i)\}^2}. \tag{4.5}$$

Actually, only the sum of the square local costs is calculated for the selection.

In this subsection, in order to represent the costs more easily, we divide the five sub-costs described in **Section 3.2** into two commonly used costs, i.e. a target cost C_t and a concatenation cost C_c [8][18][42][80]. These costs are given by

$$\begin{aligned} C_t(u_i, t_i) &= w_{pro}/w_t \cdot C_{pro}(u_i, t_i) \\ &\quad + w_{app}/w_t \cdot C_{app}(u_i, t_i), & (4.6) \\ C_c(u_i, u_{i-1}) &= w_{env}/w_c \cdot C_{env}(u_i, u_{i-1}) \\ &\quad + w_{spec}/w_c \cdot C_{spec}(u_i, u_{i-1}) \\ &\quad + w_{F_0}/w_c \cdot C_{F_0}(u_i, u_{i-1}), & (4.7) \\ w_t &= w_{pro} + w_{app}, & (4.8) \\ w_c &= w_{F_0} + w_{env} + w_{spec}, & (4.9) \end{aligned}$$

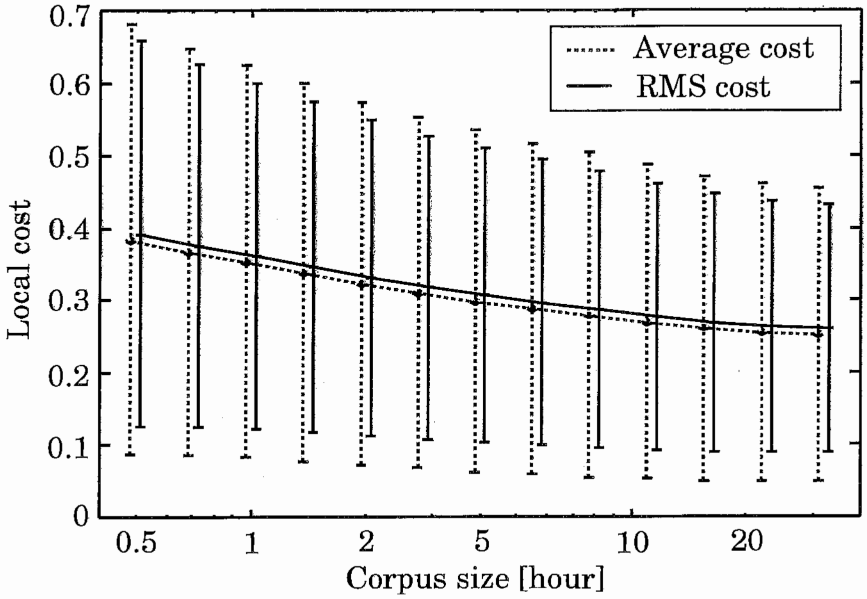

Figure 4.11: Local costs as a function of corpus size. Mean and standard deviation are shown.

and then the local cost is written as

$$LC(u_i, t_i) = w_t \cdot C_t(u_i, t_i) + w_c \cdot C_c(u_i, u_{i-1}), \tag{4.10}$$

$$w_t + w_c = 1. \tag{4.11}$$

We compared the segment selection based on the RMS cost with that based on the average cost. We utilized 1,131 utterances as an evaluation set. These utterances were not included in the corpus used for segment selection. In the segment selection, concatenations at certain phoneme centers, i.e. preceding vowel centers of voiced phonemes and unvoiced fricative centers, were also allowed in order to alleviate audible discontinuity.

4.4.1 Effect of RMS cost on various costs

We investigated the effect of the RMS cost on the local cost. **Figure 4.11** shows the local costs as a function of corpus size. In the segment selection based on the RMS cost ("RMS cost"), the standard deviation of the local cost is smaller than that of the segment selection based on the average cost ("Average cost"), although the mean of this cost is slightly worse. This is a consequence of the large penalty imposed on a segment with a large local cost in the case of the RMS cost.

In order to clarify what causes the decrease in the standard deviation, we investigated the effects of the RMS cost on both the target cost and the concatenation cost. The target cost is shown in **Figure 4.12** as a function of corpus size. The mean of the target costs is degraded, and the standard deviation increases slightly by utilizing the RMS cost. **Figure 4.13** shows the concatenation cost as a function of corpus size. Although the means of concatenation costs are equal between the average cost and the RMS cost, the standard deviation becomes smaller by utilizing the RMS cost. The increase in the standard deviation of the target cost is much smaller than the decrease in the standard deviation of the concatenation cost.

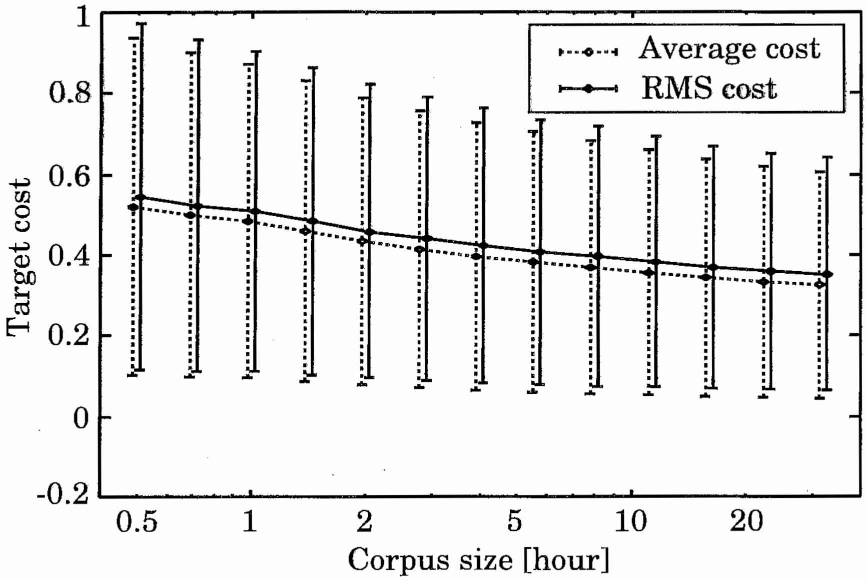

Figure 4.12: Target cost as a function of corpus size.

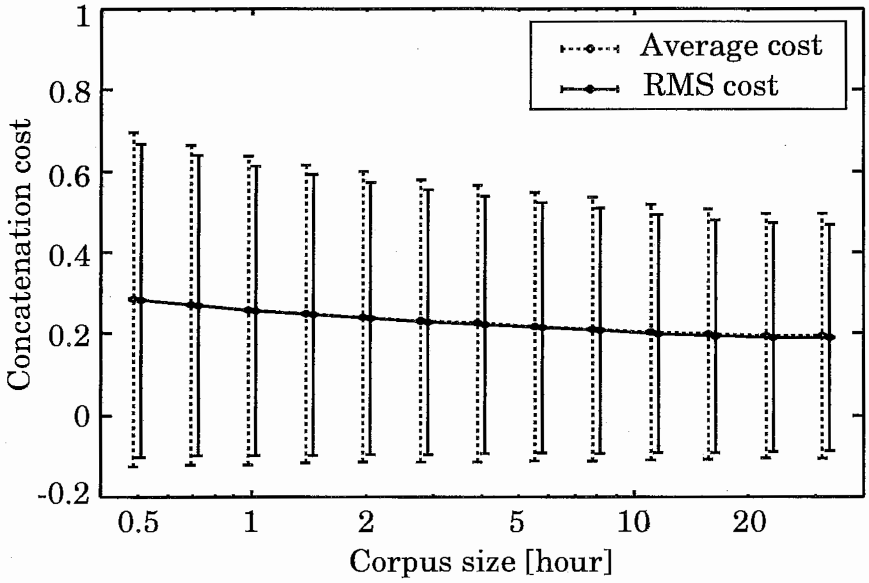

Figure 4.13: Concatenation cost as a function of corpus size.

These results show that the effectiveness of decreasing the standard deviation of the local cost is dependent on the concatenation cost. On the other hand, the mean of the local cost is slightly worse as a consequence of the degradation of the target cost.

However, it might be assumed that these results would be influenced by the weights for sub-costs rather than by the local degradation, since we utilized a weight set in which the weight for the target cost was smaller than that for the concatenation cost, i.e. $w_t = 0.4$, $w_c = 0.6$ in Equation (4.10). Therefore, we tried to analyze the effects of utilizing other weight sets. The same results were obtained for all weight sets, in which the ratios of the target cost to the concatenation cost were set to 1 to 2, 1 to 1.5, 1 to 1, 1.5 to 1, and 2 to 1. Therefore, the effectiveness mentioned above depends not on the weights for sub-costs but on the function used to integrate the local costs, which can take account of the local degradation as well as the the total degradation.

4.4.2 Effect of RMS cost on selected segments

In order to clarify what causes the decrease in the standard deviation of the concatenation cost, we investigated characteristics of the segments selected by minimizing the RMS cost. The segment length in the number of phonemes is shown in **Figure 4.14** as a function of corpus size. The segment length is shorter in the selection with the RMS cost compared with that in the case of the average cost, while the standard deviation is nearly equal.

Moreover, **Figure 4.15** shows the segment length in number of syllables as a function of corpus size for reference, since our segment selection is essentially based on syllable units, i.e. concatenations at C-V boundaries (C: Consonant and V: Vowel) are prohibited. In calculating the segment length, the number of syllables for half-phonemes in a diphone unit is set to 0.5 when the syllable is V, or 0.25 when the syllable is comprised of CV. It can be seen that the segment length is unexpectedly short, only less than 1.4 syllables, although the standard deviation of the segment length is large. In the segment selection, it is not necessarily best to select the longest segments, since such segments often cause a decrease in the number of candidate segments and thus cannot always synthesize speech naturally. The important point is to select the best segment sequence by considering not only the degradation caused by concatenation but also that caused by various factors, e.g. prosodic distance. It is also shown that the segment length becomes shorter as the corpus size increases to a level over 20 hours. This result is caused by pruning candidate segments to reduce the computational complexity of segment selection. We perform the pruning process, called pre-selection [25], by considering the target cost and the mismatch of phonetic environments. Namely, we do not consider whether segments are connected in the corpus. Therefore, when we use the large-sized corpus that includes many candidate segments having target phonetic environments, remaining candidate segments do not always connected in the corpus even if these segments have target phonetic environments.

The rate of increase in the number of concatenations is shown in **Figure 4.16** as a function of corpus size. This rate is calculated by dividing the number of concatenations in the case of utilizing the RMS cost by that in the case of utilizing the average cost. By utilizing the RMS cost, the concatenation at a boundary between any phoneme and a voiced consonant ("* - Voiced consonant") decreases in any corpus size. However, the concatenations at both a phoneme center ("Phoneme center") and a boundary between any phoneme and an unvoiced consonant ("* - Unvoiced consonant") increase. **Figure 4.17** shows the concatenation cost in each type of concatenation when the corpus size is 32 hours. The concatenation between any phoneme and an unvoiced consonant can often reduce the concatenation cost compared with that between any phoneme and a voiced consonant, since the former type of concate-

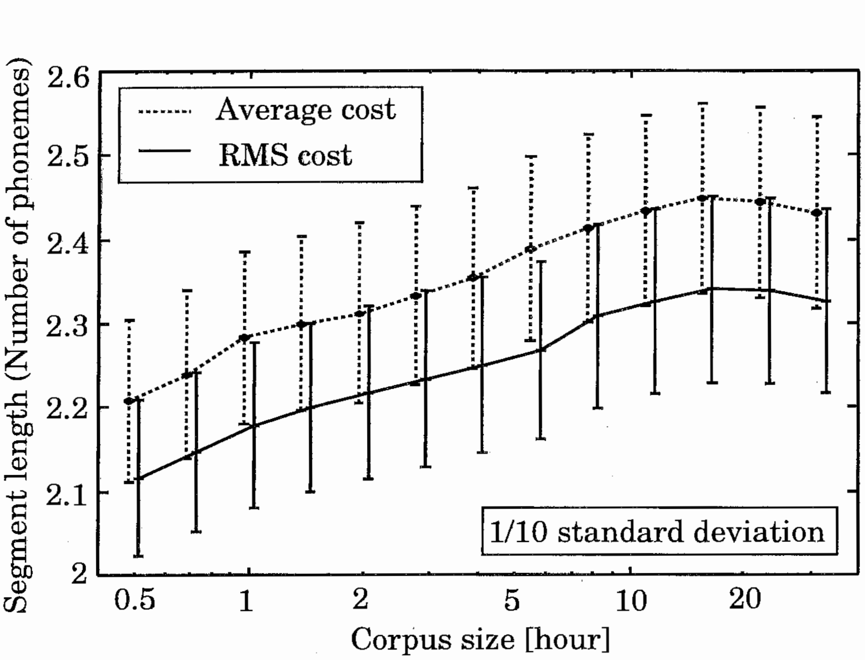

Figure 4.14: Segment length in number of phonemes as a function of corpus size.

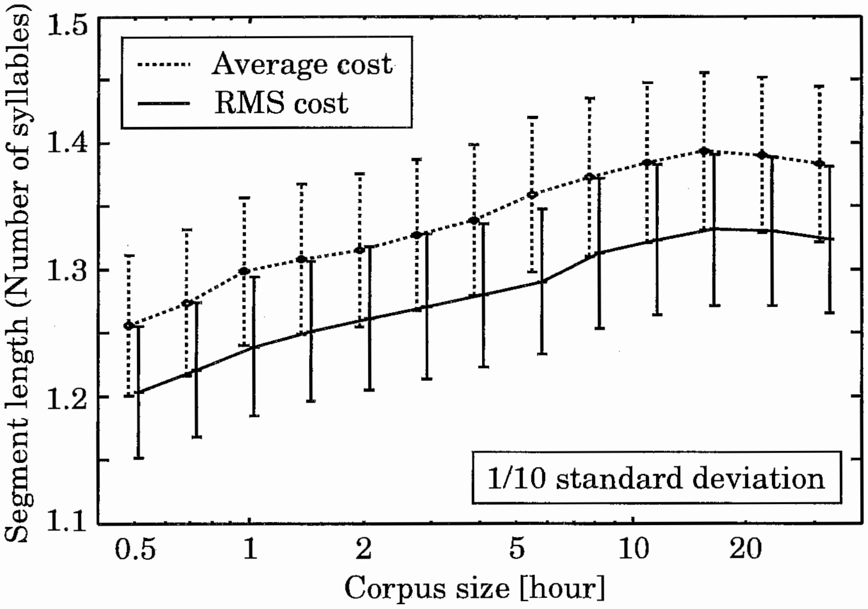

Figure 4.15: Segment length in number of syllables as a function of corpus size.

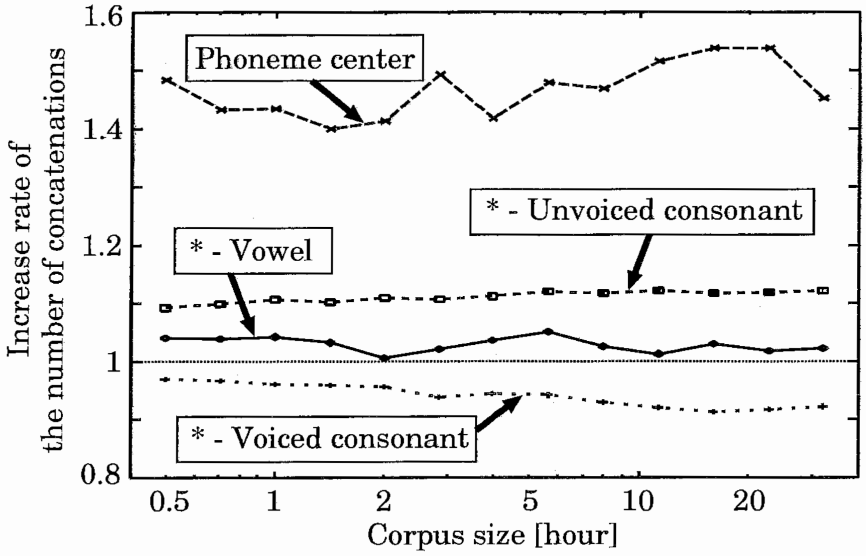

Figure 4.16: Increase rate in the number of concatenations as a function of corpus size. "*" denotes any phoneme.

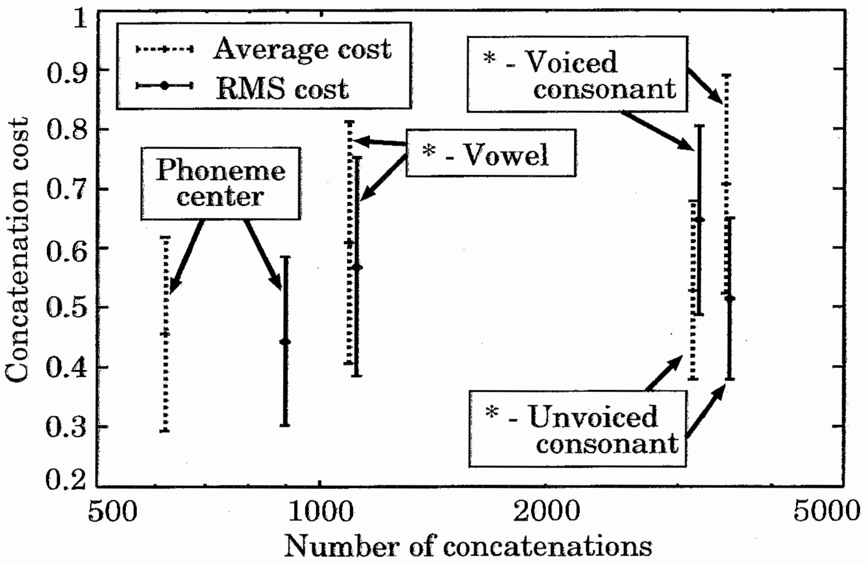

Figure 4.17: Concatenation cost in each type of concatenation. The corpus size is 32 hours.

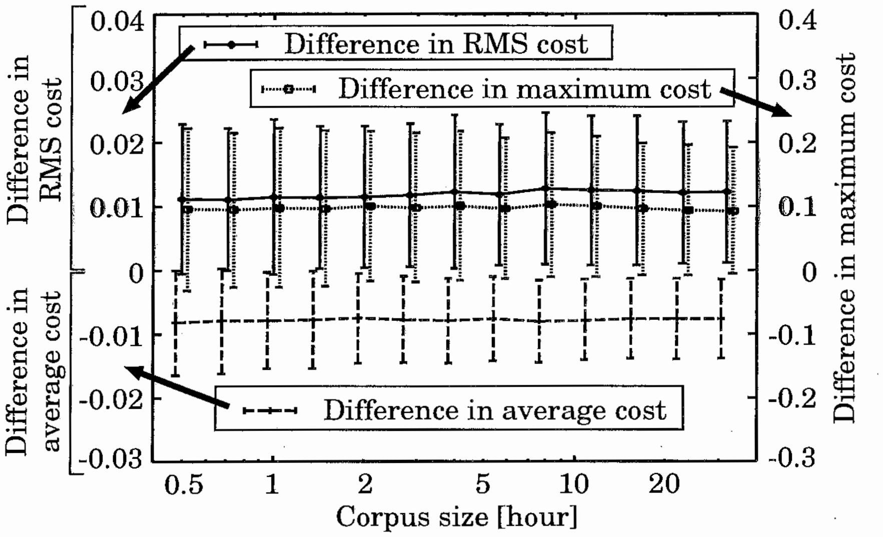

Figure 4.18: Differences in costs as a function of corpus size.

nation has no discontinuity caused by concatenating F_0s at a segment boundary. It was also found that the concatenation at the phoneme center tends to reduce the concatenation cost compared with the other types of concatenation, although the number of concatenations is small.

These results show a tendency to avoid performing concatenations that cause much local degradation of naturalness by instead performing more concatenations that cause slight audible discontinuity. As a whole, the number of concatenations increases rather than decreases. Therefore, a larger number of segments with shorter lengths, which only cause slight local degradation, are selected by utilizing the RMS cost.

4.4.3 Relationship between effectiveness of RMS cost and corpus size

In order to clarify the relationship between the effectiveness of using RMS cost and the corpus size, we investigated the differences in average costs, RMS costs, and maximum costs between the segment sequences selected by utilizing the average cost and those by utilizing the RMS cost.

The results are shown in **Figure 4.18**. The cost differences are calculated by subtracting the costs of the segment sequences selected by utilizing the RMS cost from those of the segment sequences selected by utilizing the average cost as described in **Section 4.3.2**. From the results, the RMS cost works well for alleviating the local degradation of naturalness, since the maximum cost becomes small, i.e. the differences in the maximum cost are positive. Moreover, the differences in all costs have little dependence on the corpus size. Therefore, the effectiveness of utilizing the RMS cost can be found in a corpus of any size.

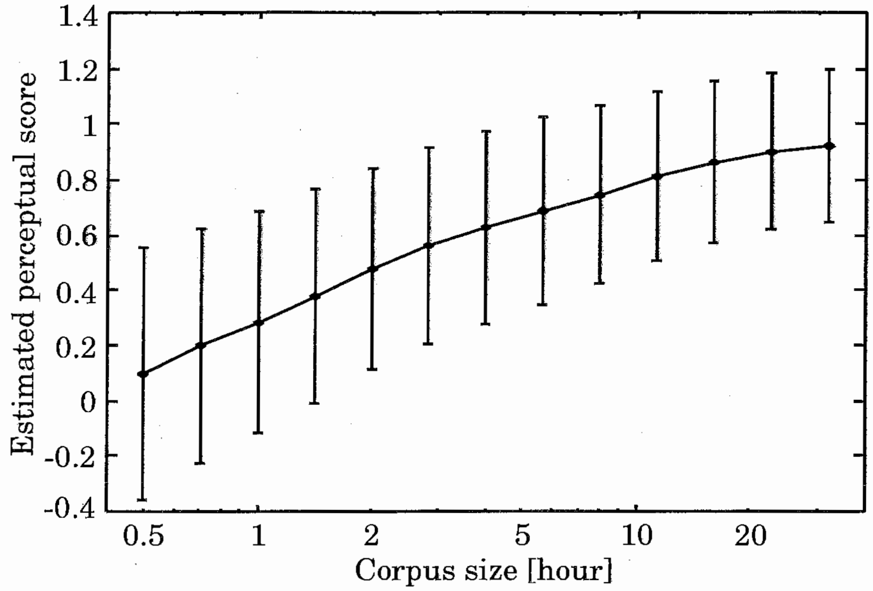

Figure 4.19: Estimated perceptual score as a function of corpus size.

4.4.4 Evaluation of segment selection by estimated perceptual score

The performance of segment selection is shown in the cost. However, it is difficult to estimate the naturalness of synthetic speech from the cost value directly. In order to indicate the performance of segment selection in a more intuitive quantity than the cost, we converted the cost value into a perceptual score by using the regression line on the RMS cost shown in **Figure 4.6**. Chu and Peng have also estimated the MOS from their cost [21][67].

Estimated perceptual score is shown in **Figure 4.19** as a function of corpus size. Small-sized corpora have been constructed so that various phonetic contexts are included. As the corpus size becomes larger, the estimated perceptual score becomes higher and its standard deviation becomes smaller. This result means that the quality of the segment selection is higher and more consistent by utilizing the larger corpus.

4.5 Summary

In segment selection for concatenative TTS, it is important to utilize a cost that corresponds to the perceptual characteristics. In this section, we evaluated a cost for segment selection based on a comparison with perceptual scores determined from the results of perceptual experiments on the naturalness of synthetic speech. As a result, we clarified that the average cost, which captures the total degradation, has a better correspondence to the perceptual scores than does the maximum cost, which captures the local degradation. Furthermore, we found that the RMS (Root Mean Square) cost, which takes into account both the average cost and the maximum cost, has the best correspondence. We also clarified that the naturalness of synthetic speech could be slightly improved by utilizing the RMS cost.

We investigated the effect of considering not only the degradation of naturalness over the entire synthetic speech but also local degradation in segment selection. In this selection, the optimum segment sequences are selected by minimizing the RMS cost instead of the conventional average cost. From the results of experiments comparing this approach with segment selection based on the average cost, it was found that segment selection based on RMS cost performed a larger number of concatenations that caused slight local degradation in order to avoid concatenations causing greater local degradation. Namely, a larger number of segments with shorter units were selected. Moreover, the effectiveness of this selection was found for any size of corpus.

When the RMS costs were distributed widely, the correspondence of the RMS cost seemed to be good. Namely, it was possible to accurately estimate the perceptual score from the RMS cost for synthetic speech of various qualities. Therefore, we evaluated the performance of segment selection by the estimated perceptual score by varying the corpus size. As a result, the quality of the segment selection was higher and more consistent by utilizing the larger corpus.

We also performed the perceptual evaluation of the RMS cost in a lower range of the RMS cost. The results clarified that the correspondence of the RMS cost to the perceptual scores is inadequate in this case. Therefore, it is obvious that the RMS cost is not accurate enough for making comparisons between similar segments, which is naturally a difficult problem. However, since our TTS does not consistently synthesize sufficiently natural speech, we should further improve the cost function based on perceptual characteristics. In particular, it is necessary to determine the optimum weight set for sub-costs. We will determine this weight set from the results of perceptual experiments on the naturalness of synthetic speech with a set of stimuli covering a wide range in terms of individual sub-costs.

Chapter 5

Conclusions

5.1 Summary of the Report

Corpus-based Text-to-Speech (TTS) enables us to dramatically improve the naturalness of synthetic speech over that of rule-based TTS. However, so far no general-purpose TTS has been developed that can consistently synthesize sufficiently natural speech. In this report, we addressed the problem how to improve the naturalness of synthetic speech in corpus-based TTS.

We first described the structure of a corpus-based TTS system in **Chapter 2**. Almost all corpus-based TTS systems have been developed on the basis of this structure. The various techniques in each module were reviewed.

In **Chapter 3**, we proposed a novel segment selection algorithm for Japanese speech synthesis in order to improve the naturalness of synthetic speech. Since Japanese syllables consist of CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) or V, except when a vowel is devoiced, CV units are often used in concatenative TTS systems for Japanese. However, speech synthesized with CV units sometimes have auditory discontinuity due to V-V and V-semivowel concatenations. Since various vowel sequences appear frequently in Japanese, it is not realistic to prepare long units that include all possible vowel sequences to avoid V-V concatenation. In order to address this problem, we proposed a novel segment selection algorithm that does not avoid the concatenation of the vowel sequences but alleviates the discontinuity by utilizing both phoneme and diphone units. In the proposed algorithm, non-uniform units allowing concatenation not only at phoneme boundaries but also at vowel centers can be selected from a speech corpus. The experiments on concatenation of vowel sequences clarified that the number of better candidate segments increases by considering concatenations both at phoneme boundaries and at vowel centers. We also performed perceptual experiments. The results showed that speech synthesized with the proposed algorithm has better naturalness than that of the conventional algorithms. We also compared the proposed algorithm with the algorithm based on half-phoneme units. Moreover, a cost function for selecting the optimum waveform segments in our TTS system was described.

In **Chapter 4**, we performed a perceptual evaluation of costs for segment selection. In order to achieve high-quality segment selection for concatenative TTS, it is important to utilize a cost that corresponds to the perceptual characteristics. From the results of perceptual experiments, we clarified the correspondence of the cost to the perceptual scores and then evaluated various functions to integrate local costs capturing degradation in individual segments. As a result, it was clarified that the average cost, which captures the degradation of naturalness over the entire synthetic speech, has better correspondence to the perceptual scores than the maximum cost, which captures the local degradation of naturalness. Furthermore, the RMS (Root Mean Square) cost taking account of both average cost and maximum cost has the best correspondence. We also showed that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost. Then, we investigated the effect of using the RMS cost for segment selection. From the results of experiments comparing this approach with segment selection based on the conventional average cost, it was found that (1) in segment selection based on the RMS cost, a larger number of concatenations causing slight local degradation were performed in order to avoid concatenations causing greater local degradation, and (2) the effect of the RMS cost has little dependence on the size of the corpus. We also clarified that utilizing a larger corpus improves the quality and consistency of the segment selection.

In summary, we confirmed that the proposed segment selection algorithm and the proposed cost function based on perceptual evaluation are effective for improving the naturalness of synthetic speech.

5.2 Future Work

Although we have improved the corpus-based TTS system, a number of problems still remain to be solved.

Effective search algorithm: The computational complexity of segment selection needs to be reduced while maintaining the naturalness of synthetic speech. As an approach to this problem, some clustering algorithms have been proposed to decrease the number of candidate segments [11][30]. Decision trees are constructed by utilizing target features in segment selection in advance in order to cluster similar candidate segments into the same classes. Other approaches have proposed pre-selection by sub-costs with small computational cost [25], which is applied in our TTS, and construction of a practical and efficient cache of sub-costs with high computational cost [6]. Moreover, a pruning algorithm that considers concatenation between candidate segments has been proposed [15]. It is expected that a larger-sized corpus will be used in the future to synthesize high-quality speech more consistently. More approaches from various viewpoints are needed to address this problem.

Measures against voice-quality variation: Variation in voice quality is caused by recording the speech of a speaker for a long time. Some approaches as described in **Section 2.2.5** have been proposed [56][75]. However, much more research is needed to solve this problem.

Utilization of multiple targets: Almost all algorithms use the most suitable target information predicted from contextual information. However, the predicted target information is not always the best in cases where only a small number of candidate segments having the predicted target exist in the corpus. It is assumed that segment sequences with smooth concatenations cannot be selected under such conditions. As an interesting approach to this problem, Bulyko et al. proposed a selection algorithm that considers multiple targets [12]. Moreover, Hirai et al. proposed a selection algorithm based on acceptable prosodic targets in Japanese speech synthesis [37]. Especially in Japanese, since accent information is crucial, it is important to avoid selecting segment sequences with unacceptable accent information. Moreover, it might be promising to utilize not only accent information but also other contextual information, e.g. syntactic structure in segment selection.

Improvement of cost: Cost functions for segment selection should be improved based on perceptual characteristics [21][67]. In this report, we did not determine the optimum weight set for sub-costs. Although a determination algorithm based on linear regression has been proposed, it uses an acoustic measure, e.g. cepstral distortion, as an objective measure in the regression [42]. However, the correspondence to perceptual characteristics of acoustic measures is not sufficient. Therefore, it is necessary to explore a measure having better correspondence to perceptual characteristics.

Various applications of TTS: Not only general-purpose TTS but also limited domain TTS have been studied [12][22]. As an effective approach in the case of a small-sized corpus, spectral modification has been applied to concatenation speech synthesis [89]. Moreover, it is necessary to synthesize not only speech in a reading style but also speech in various speaking styles [48][58] as well as expressive speech [13][43] in order to realize rich communication between man and machine. A number of these problems still remain to be solved.

Appendix

A Frequency of Vowel Sequences

Table A .1 shows the frequency of vowel sequences in newspaper articles comprised of 571,283 sentences. Phoneme sequences were divided into CV* units [53], and then we calculated the frequency of vowel sequences while ignoring consonants, e.g. both /kai/ (CVV) and /ai/ (VV) are considered /ai/ (length = 2). Semivowels were considered vowels. The length of long vowels is set to 1, and that of devoiced vowels is set to 0.

Table A .1: Frequency of vowel sequences

Length	Frequency	Normalized frequency [%]	Number of different sequences
0	518553	2.7590	-
1	16069643	85.4999	11
2	1677866	8.9272	99
3	459538	2.4450	575
4	57090	0.3038	1154
5	10306	0.0548	1155
6	1582	0.0084	473
7	306	0.0016	146
8	28	0.0002	30
9	4	0.0000	4

B Definition of the Nonlinear Function P

We performed perceptual experiments on the degradation of naturalness caused by prosody modification with STRAIGHT [51][52] in order to define a sub-cost function on prosodic difference. Listeners evaluated the degradation on a scale of seven levels, namely 1 (very bad) to 7 (very good). Perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The perceptual scores were modeled by a nonlinear function z as follows:

$$\begin{aligned} z(x,y) &= \min\{s, g(x,y)\}, && \text{(B .1)} \\ g(x,y) &= a \cdot \exp\left[-\left\{\left(\frac{x}{S_x}\right)^2 - 2 \cdot r \cdot \left(\frac{x}{S_x}\right) \cdot \left(\frac{y}{S_y}\right)\right.\right. \\ &\qquad \left.\left. + \left(\frac{y}{S_y}\right)^2\right\}\right] + b, && \text{(B .2)} \\ S_x &= \begin{cases} S_{xp} & (x > 0) \\ S_{xm} & (x \leq 0) \end{cases}, \quad S_y = \begin{cases} S_{yp} & (y > 0) \\ S_{ym} & (y \leq 0) \end{cases}, && \text{(B .3)} \end{aligned}$$

where, x and y denote F_0 modification ratio and duration modification ratio by octave, respectively. Parameters a, b, r, s, S_{xp}, S_{xm}, S_{yp}, and S_{ym} are given by

$$\begin{array}{rclrcl} a &=& 2.649368, & b &=& -1.574549, \\ r &=& 0.062317, & s &=& 0.942505, \\ S_{xp} &=& 0.400522, & S_{xm} &=& 0.568179, \\ S_{yp} &=& 0.924295, & S_{ym} &=& 0.913813. \end{array} \qquad \text{(B .4)}$$

These parameters were estimated by minimizing the error between the perceptual score and that estimated by the function $z(x, y)$. The nonlinear function P in Equation (3.3) is defined as follow:

$$P(x,y) = -z(x,y) + s. \qquad \text{(B .5)}$$

In the cases of 1) the sub-cost on F_0 discontinuity as described in Equation (3.4) and 2) no signal processing for prosody modification in waveform synthesis, the parameters are given by

$$\begin{array}{rclrcl} a &=& 2.649368, & b &=& -1.574549, \\ r &=& 0.0, & s &=& 1.074819, \\ S_{xp} &=& 0.400522, & S_{xm} &=& 0.400522, \\ S_{yp} &=& 0.913813, & S_{ym} &=& 0.913813. \end{array} \qquad \text{(B .6)}$$

C Sub-Cost Functions, S_s and S_p, on Mismatch of Phonetic Environment

The sub-cost functions were determined from results of perceptual experiments, in which listeners evaluated the degradation of naturalness by listening to the speech stimuli synthesized by concatenating phonemes extracted from various phonetic environments. The experimental method is described in [57]. Listeners evaluated the degradation on a scale of seven levels, namely 1 (very bad) to 7 (very good). Perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The cost S_s of capturing the degradation caused by a mismatch with the succeeding environment is given by

$$S_s(Ph, Ph_e, Ph_u) \quad = \quad -z(Ph, Ph_e, Ph_u) + b, \qquad \text{(C .1)}$$

where $z(Ph, Ph_e, Ph_u)$ denotes the perceptual score in the case of a mismatch of the succeeding environment in a phoneme Ph, i.e. replacing a phoneme Ph_e with a phoneme Ph_u. b is set to 1.5, which is utilized in order to convert the perceptual score into a positive value. The cost S_p of capturing the degradation caused by a mismatch with the preceding environment is determined in a similar way.

References

[1] M. Abe, Y. Sagisaka, T. Umeda, H. Kuwabara. Speech database user's manual. *ATR Technical Report*, TR-I-0166, 1990.

[2] M. Akamine and T. Kagoshima. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS). *Proc. ICSLP*, pp. 1927–1930, Sydney, Australia, Dec. 1998.

[3] C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House. Reduction of speech spectra by analysis-by-synthesis techniques. *J. Acoust. Soc. Am.*, Vol. 33, No. 12, pp. 1725–1736, 1961.

[4] M. Beutnagel, A. Conkie, and A.K. Syrdal. Diphone synthesis using unit selection. *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 185–190, Jenolan Caves, Australia, Nov. 1998.

[5] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A.K. Syrdal. The AT&T Next-Gen TTS system. *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, Mar. 1999. http://www.research.att.com/projects/tts/pubs.html

[6] M. Beutnagel, M. Mohri, and M. Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. *Proc. EUROSPEECH*, pp. 607–610, Budapest, Hungary, Sep. 1999.

[7] A.W. Black and P. Taylor. CHATR: a generic speech synthesis system. *Proc. COLING*, pp. 983–986, Kyoto, Japan, Aug. 1994.

[8] 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.

[9] A.W. Black. Building practical speech synthesis systems. *ATR Technical Report*, TR-IT-0163, 1996.

[10] A.W. Black and A.J. Hunt, Generating F_0 contours from ToBI labels using linear regression. *Proc. ICSLP*, pp. 1385–1388, Philadelphia, U.S.A., Oct. 1996.

[11] A.W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. *Proc. EUROSPEECH*, pp. 601–604, Rhodes, Greece, Sep. 1997.

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

[13] M. Bulut, S.S. Narayanan, A.K. Syrdal. Expressive speech synthesis using a concatenative synthesizer. *Proc. ICSLP*, pp. 1265–1268, Denver, U.S.A., Sep. 2002.

[14] I. Bulyko and M. Ostendorf. Efficient integrated response generation from multiple targets using weighted finite state transducers. *Computer Speech and Language*, Vol. 16, No. 3–4, pp. 533–550, 2002.

[15] I. Bulyko, M. Ostendorf, and J. Bilmes. Robust splicing costs and efficient search with BMM models for concatenative speech synthesis. *Proc. ICASSP*, pp. 461–464, Orlando. U.S.A., May 2002.

[16] W.N. Campbell and C.W. Wightman. Prosodic encoding of syntactic structure for speech synthesis. *Proc. ICSLP*, pp. 369–372, Banff, Canada, Oct. 1992.

[17] W.N. Campbell. Prosody and the selection of units for concatenation synthesis. *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 61–64, New York, U.S.A., Sep. 1994.

[18] W.N. Campbell and A.W. Black. Prosody and the selection of source units for concatenative synthesis. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 279–292, 1997.

[19] W.N. Campbell. Processing a speech corpus for CHATR synthesis. *Proc. ICSP*, pp. 183–186, Seoul, Korea, Aug. 1997.

[20] M. Chu, H. Peng, H. Yang, and E. Chang. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. *Proc. ICASSP*, pp. 785–788, Salt Lake City, U.S.A., May 2001.

[21] M. Chu and H. Peng. An objective measure for estimating MOS of synthesized speech. *Proc. EUROSPEECH*, pp. 2087–2090, Aalborg, Denmark, Sep. 2001.

[22] M. Chu, C. Li, H. Peng, and E. Chang. Domain adaptation for TTS systems. *Proc. ICASSP*, pp. 453–456, Orlando. U.S.A., May 2002.

[23] A. Conkie and S. Isard. Optimal coupling of diphones. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 293–304, 1997.

[24] A. Conkie. Robust unit selection system for speech synthesis. *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, Mar. 1999.
http://www.research.att.com/projects/tts/pubs.html

[25] A. Conkie, M. Beutnagel, A.K. Syrdal, and P.E. Brown. Preselection of candidate units in a unit selection-based text-to-speech synthesis system. *Proc. ICSLP*, Vol. 3, pp. 279–282, Beijing, China, Oct. 2000.

[26] 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.

[27] N.R. Dixon and H.D. Maxey. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 1, pp. 40–50, 1968.

[28] R.E. Donovan and E.M. Eide. The IBM trainable speech synthesis system. *Proc. ICSLP*, pp. 1703–1706, Sydney, Australia, Dec. 1998.

[29] R.E. Donovan and P.C. Woodland. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, Vol. 13, No. 3, pp. 223-241, 1999.

[30] R.E. Donovan. Segment pre-selection in decision-tree based speech synthesis systems. *Proc. ICASSP*, pp. 937–940, Istanbul, Turkey, June 2000.

[31] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169-177, 1939.

[32] K.E. Dusterhoff and A.W. Black. Generating F_0 contours for speech synthesis using the Tilt intonation theory. *Proc. ESCA Workshop on Intonation*, pp. 107–110, Athens, Greece, Sep. 1997.

[33] K.E. Dusterhoff, A.W. Black, and P. Taylor. Using decision trees within the Tilt intonation model to predict F_0 contours. *Proc. EUROSPEECH*, Budapest, Hungary, pp. 1627–1630, Sep. 1999.

[34] H. Fujisaki, K. Hirose. Analysis of voice fundamental frequency contours for declarative sentence of Japanese. *J. Acoust. Soc. Jpn. (E)*, Vol. 5, No. 4, pp. 233–242, 1984.

[35] K. Hakoda, S. Nakajima, T. Hirokawa, and H. Mizuno. A new Japanese text-to-speech synthesizer based on COC synthesis method. *Proc. ICSLP*, pp. 809–812, Kobe, Japan, Nov. 1990.

[36] T. Hirai, N. Iwahashi, N. Higuchi, and Y. Sagisaka. Automatic extraction of F_0 control rules using statistical analysis. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 333–346, 1997.

[37] T. Hirai, S. Tenpaku, and K. Shikano. Speech unit selection based on target values driven by speech data in concatenative speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[38] T. Hirokawa. Speech synthesis using a waveform dictionary. *Proc. EUROSPEECH*, pp. 140–143, Paris, France, Sep. 1989.

[39] T. Hirokawa and K. Hakoda. Segment selection and pitch modification for high quality speech synthesis using waveform segments. *Proc. ICSLP*, pp. 337–340, Kobe, Japan, Nov. 1990.

[40] K. Hirose, H. Fujisaki, and H. Kawai. Generation of prosodic symbols for rule-synthesis of connected speech of Japanese. *Proc. ICASSP*, pp. 2415–2418, Tokyo, Japan, Apr. 1986.

[41] K. Hirose, M. Eto, and N. Minematsu. Improved corpus-based synthesis of fundamental frequency contours using generation process model. *Proc. ICSLP*, pp. 2085–2088, Denver, U.S.A., Sep. 2002.

[42] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP*, pp. 373–376, Atlanta, U.S.A., May 1996.

[43] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura. A speech synthesis system with emotion for assisting communication. *Proc. ISCA Workshop on Speech and Emotion*, pp. 167–172, Belfast, Northern Ireland, Sep. 2000.

[44] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (MLSA) filter for speech synthesis. *IEICE Trans.*, Vol. J66-A, No. 2, pp. 122–129, 1983 (in Japanese).

[45] M. Isogai and H. Mizuno. A new F_0 contour control method based on vector representation of F_0 contour. *Proc. EUROSPEECH*, pp. 727–730, Budapest, Hungary, Sep. 1999.

[46] K. Itoh, S. Nakajima, and T. Hirokawa. A new waveform speech synthesis approach based on the COC speech spectrum. *Proc. ICASSP*, pp. 577–580, Adelaide, Australia, Apr. 1994.

[47] N. Iwahashi, N. Kaiki, and Y. Sagisaka. Speech segment selection for concatenative synthesis based on spectral distortion minimization. *IEICE Trans. Fundamentals*, Vol. E76-A, No. 11, pp. 1942–1948, 1993.

[48] K. Iwano, M. Yamada, T. Togawa, and S. Furui, Speech-rate-variable HMM-based Japanese TTS system. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[49] T. Kagoshima and M. Akamine. Automatic generation of speech synthesis units based on closed loop training. *Proc. ICASSP*, pp. 963–966, Munich, Germany, Apr. 1997.

[50] T. Kagoshima and M. Akamine. An F_0 contour control model for totally speaker driven text to speech system. *Proc. ICSLP*, pp. 1975–1978, Sydney, Australia, Dec. 1998.

[51] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.

[52] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity. *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.

[53] H. Kawai, N. Higuchi, T. Shimizu, and S. Yamamoto. Development of a text-to-speech system for Japanese based on waveform splicing. *Proc. ICASSP*, pp. 569–572, Adelaide, Australia, Apr. 1994.

[54] H. Kawai, K. Hirose, and H. Fujisaki. Rules for generating prosodic features for text-to-speech synthesis of Japanese. *J. Acoust. Soc. Jpn. (J)*, Vol. 50, No. 6, pp. 433–442, 1994 (in Japanese).

[55] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu. A design method of speech corpus for text-to-speech synthesis taking account of prosody. *Proc. ICSLP*, Vol. 3, pp. 420–425, Beijing, China, Oct. 2000.

[56] H. Kawai and M. Tsuzaki. A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[57] 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.

[58] H. Kawanami, T. Masuda, T. Toda, and K. Shikano. Designing Japanese speech database covering wide range in prosody. *Proc. ICSLP*, pp. 2425–2428, Denver, U.S.A., Sep. 2002.

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

[60] D.H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, Vol. 82, No. 3, pp. 737–793, 1987.

[61] M. Lee. Perceptual cost functions for unit searching in large corpus-based concatenative text-to-speech. *Proc. EUROSPEECH*, pp. 2227–2230, Aalborg, Denmark, Sep. 2001.

[62] N. Minematsu, R. Kita, and K. Hirose. Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[63] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, Vol. 9, No. 5–6, pp. 453–467, 1990.

[64] S. Nakajima and H. Hamada. Automatic generation of synthesis units based on context oriented clustering. *Proc. ICASSP*, pp. 659–662, New York, U.S.A., Apr. 1988.

[65] J.P. Olive. Rule synthesis of speech from diadic units. *Proc. ICASSP*, pp. 568–570, Hartford, U.S.A., May. 1977.

[66] A.V. Oppenheim and D.H. Johnson. Discrete representation of signals. *Proc. IEEE*, Vol. 60, No. 6, pp. 681–691, 1972.

[67] H. Peng, Y. Zhao, and M. Chu. Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation. *Proc. ICSLP*, pp. 2613–2616, Denver, U.S.A., Sep. 2002.

[68] G. Peterson, W. Wang, and E. Silvertsen. Segmentation techniques in speech synthesis. *J. Acoust. Soc. Am.*, Vol. 30, No. 8, pp. 739–742, 1958.

[69] Y. Sagisaka and H. Satoh. Accentuation rules for Japanese word concatenation. *IEICE Trans.*, Vol. J66-D, No. 7, pp. 849–856, 1983 (in Japanese).

[70] 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.

[71] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR ν-talk speech synthesis system. *Proc. ICSLP*, pp. 483–486, Banff, Canada, Oct. 1992.

[72] Y. Sagisaka, Natural language processing in speech synthesis. *IPSJ Journal*, Vol. 34, No. 10, pp. 1281–1286, 1993 (in Japanese).

[73] H. Sato. Speech synthesis on the basis of PARCOR-VCV concatenation units. *IEICE Trans.*, Vol. J61-D, No. 11, pp. 858–865, 1978 (in Japanese).

[74] H. Sato. Speech synthesis using CVC concatenation units and excitation waveform elements. *Trans. Comm. Speech Res., Acoust. Soc. Jpn.*, S83-69, pp. 541–546, 1984 (in Japanese).

[75] Y. Shi, E. Chang, H. Peng, and M. Chu. Power spectral density based channel equalization of large speech database for concatenative TTS system. *Proc. ICSLP*, pp. 2369–2372, Denver, U.S.A., Sep. 2002.

[76] T. Shimizu, N. Higuchi, H. Kawai, and S. Yamamoto. A morphological analyzer for a Japanese text-to-speech system based on the strength of connection between words. *J. Acoust. Soc. Jpn. (J)*, Vol. 51, No. 1, pp. 3–13, 1995 (in Japanese).

[77] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling English prosody. *Proc. ICSLP*, pp. 867–870, Banff, Canada, Oct. 1992.

[78] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 21–29, 2001.

[79] 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.

[80] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP*, Vol. 3, pp. 410–415, Beijing, China, Oct. 2000.

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

[82] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima. A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 3–10, 2001.

[83] K. Takeda, K. Abe, and Y. Sagisaka. On the basic scheme and algorithms in non-uniform unit speech synthesis. *Talking machines: theories, models, and designs*, G. Bailly, C. Benoit, and T. Sawallis, North-Holland, Elsevier Science, pp. 93–105, 1992.

[84] P. Taylor and A.W. Black. Synthesizing conversational intonation from a linguistically rich input. *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 175–178, New York, U.S.A., Sep. 1994.

[85] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. *Proc. ICASSP*, pp. 229–232, Phoenix, U.S.A., May 1999.

[86] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[87] M. Tsuzaki and H. Kawai. Feature extraction for unit selection in concatenative speech synthesis: comparison between AIM, LPC, and MFCC. *Proc. ICSLP*, pp. 137–140, Denver, U.S.A., Sep. 2002.

[88] J. Wouters and M.W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. *Proc. ICSLP*, pp. 2747–2750, Sydney, Australia, Dec. 1998.

[89] J. Wouters and M.W. Macon. Control of spectral dynamics in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, Vol.'9, No. 1, pp. 30–38, 2001.

[90] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Proc. EUROSPEECH*, pp. 2347–2350, Budapest, Hungary, Sep. 1999.