Internal Use Only (非公開)

002

TR-IT-0281

# PSOLAによるCHATRの韻律改善について
# Improving Quality and Prosody of CHATR Synthesis Speech Based on PSOLA

丁 文　　　ニック・キャンベル
Wen Ding　　Nick Campbell

1998.11.11

We address two issues in this paper : (1) How to derive a perceptual discontinuity function to determine the perceptually significant amount of discontinuity between two candidate units, while (2) taking into account the constraints of possible prosodic modification (pitch/duration scaling using signal processing). Both the techniques are tested with the unit selection and synthesis modules and the changes in voice quality and prosody are evaluated.

# 目次

# 第 1 章

# Introduction

## 1 CHATR Speech Synthesis

Concatenative synthesis is widely used in TTS to generate synthetic speech with high quality and relatively natural-sounding prosody. Whatever the type of synthesis unit used, (diphone, phoneme, etc.), a large speech database is usually needed to ensure the phonetic and phonemic variation of the units in a rich variety of contexts. In the CHATR synthesis system, unit selection finds the most appropriate phoneme sequence for an input text by using a criterion of minimizing a) joint discontinuity and b) mismatch in target prosody. However, in the current unit selection module, only an objective distance function is used, and the pitch and duration are not modified to match the target prosody.

The waveform concatenation in CHATR can produce very high quality of output speech by selecting appropriate units. However, since our database can not be infinite large to cover all the combination of various phonetic units to be connected, we perceived discontinuity at the unit boundaries. For prosody of output speech, we also encountered some phrases with unnatual pitch accent due to imperfect unit selection. Considering the above problems, we address two issues in this paper : (1) how to derive a perceptual discontinuity function to determine the perceptually significant amount of discontinuity between two candidate units, while (2) taking into account the constraints of possible prosodic modification (pitch/duration scaling using signal processing).

## 2 Paper Structure

The structure of this report is arranged as follows: (1) constructing two kinds of perceptual discontinuity functions based on several perceptual experiments, one for V-V type of unit connection and the other for V-U type. (2) modifying f0 contour of selected units to partially or totally match target pitch contour by using PSOLA technique, discussions and evaluations. (3) conclusions and future work.

# 第 2 章

# Detecting Discontinuity at Unit Boundaries

CHATR uses phoneme units as the basic unit for waveform concatenation. The discontinuities between unit boundaries vary according to the phoneme type of phone units. But for this paper, the unit boundaries are grouped into 2 main classes : (1) Vowel-to-Vowel concatenation is the main source of perceived clipped sound in the synthesised speech, (2) Vowel-to-Nasal consonants, /m/, /n/, and /N/.

## 1 V-V Unit Connection

In order to auto-detect the discontinuity of units, we need a measure for this. And the measure should be consistent with the perceptual impression for discontinuity. Since what we can get is only the acoustic features of units in run-time, in this paper, we try to establish such a measure or function from acoustic features. The perceptual corresponding to the acoustic features are obtained from several perceptual experiments.

### 1.1 Experiment Samples

In the current implementation of CHATR, there is a cost function to measure the joint cost between unit boundaries based on acoustic features. But what we need is to establish a perception-based model to quantitize the amount of discontinuity in perception. This kind of model is not available. We try to generate it based on perceptual experiments. In this section, the experiment related to V-V connection is discussed.

The most frequently occurred case in unit discontinuity is for Vowel-Vowel connection. We generate data for the perceptual experiments from an ATR male database in Japanese with 503 sentences. The vowels in the middle of two triphones are connected together as shown in Fig. 2.1. The formant trajectory of two units are considered as an important factor to discontinuity perception. Finally, we collected samples such as /x-a-a-y/、 /x-a-i-y/, /x-a-e-y/, /x-i-a-y/, /x-i-i-y/, /x-o-i-y/, /x-u-e-y/, /x-u-u-y/ for the experiments.

Considering a target sequence /s-a-i-d/ as shown in Fig. 2.1, the V-V connection, /a-i/, is realized by concatenating the center vowels of two triphones. The boundary discontinuity of /a-i/ is affected by two triphone contexts and the acoustic features at the boundary. We aim to establish a model to link the perceptual opinion score and the acoustic features of the speech by performing a perceptual experiment.

The total number of samples are 120 and sampled in 12 kHz, quatitized in 16-bit. The samples are evaluated by 5 listeners with headphone to judge the degree of discontinuity of units. MOS values are obtained after the experiments :

| | |
|---|---|
| 2: | very good |
| 1: | good |
| 0: | normal |
| -1: | bad |
| -2: | very bad |

After the MOS values obtained, we must consider the factors from which the MOS values can be predicted. For the each boundary sides of the two units to be connected, the following factors are computed: cepstrum distance, $F0.D = logF0^{(1)} - logF0^{(2)}$, $PWR.D = logPWR^{(1)} - logPWR^{(2)}$, and $f0_{zs}$ :

$$f0_{zs} = \frac{f0 - \overline{f0}}{\sigma_{f0}} \tag{2.1}$$

where $\overline{f0}$、 $\sigma_{f0}$ denotes the f0 mean and std of the phoneme, respectively, and $f0_{zs}$ is the larger value of two joint phones. The $f0_{zs}$ is selected as the higher one from the connected two units.

## 1.2 Decision Tree for MOS Prediction

Before constructing the MOS prediction model, the MOS values are normalized using variance of MOS from five listeners so that some outliers are stripped out. This normalization

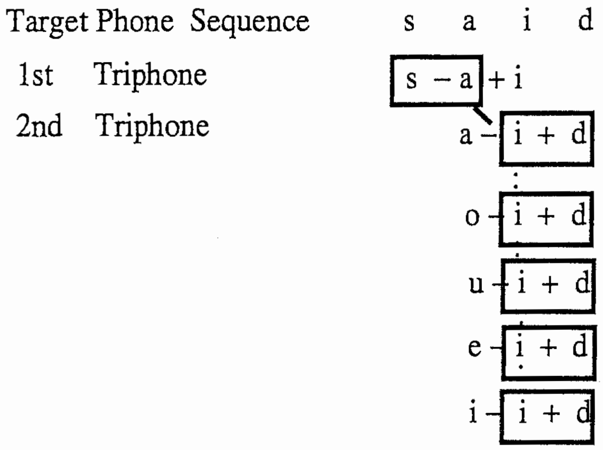

図 2.1: Description of triphone concatenation.

results in 95 samples to establish a prediction model.

Then a decision tree is made to predict the MOS value from acoustic factors : *f0dist*, *pwrdist*, *f0zs*, *cepdist*. The tree obtained from 95 samples is represented as : equation :

MOS ¡- $(cepdist + f0dist + pwrdist + f0_{zs})$

Nodes: 16

Residual mean deviance (RMD): 0.22 (18 / 84)

Distribution of RMD:

| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|
| -0.85 | -0.36 | 0.04 | 0 | 0.32 | 0.88 |

$$RMD = \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{N - number\ of\ nodes} \tag{2.2}$$

wher $N$ is total number of samples. $y_i$ is an observation value. $\tilde{y}_i$ is the predicted MOS value.

A section of the tree is shown in Fig. 2.2 (resolution at the leaves of the tree is sacrificed to provide a visual cue of split importance).

It is shown in the figure that *cepdist* is the most important factor to perception of V-V connection.

Using the training data, an open test has been carried out and the result is shown in Fig. 2.3. The correlation coefficient is 0.79.

For the open test, 10 samples selected from 95 data as the test data and using the rest

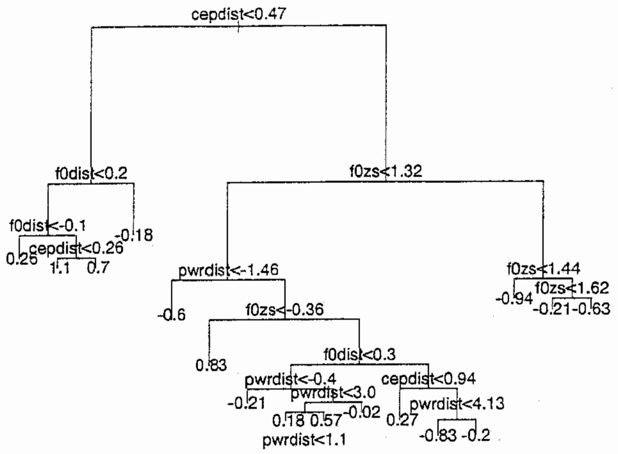

図 2.2: Dendrogram of the regression tree for prediction of MOS (node value).

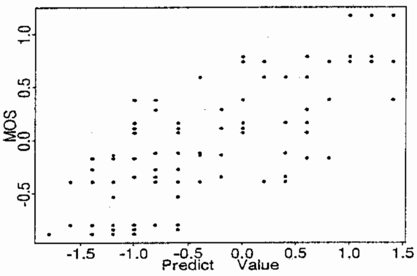

図 2.3: Relationship between MOS and the predicted value by the regression tree for the train data.

to train a decision tree, the tree is then used to predict MOS values of 10 samples. This operation is repeated 9 times for the whole data and the final result is shown in Fig. 2.4. The correlation coefficient is 0.61.

## 2 Vowel-Nasal Connection

In the same way, we investigate the vowel-nasal case.

### 2.1 Experiment Samples

Besides V-V connection, we need to consider discontinuity between vowel and voiced consonants, since they have a quite different formant structure compared to the vowels. For

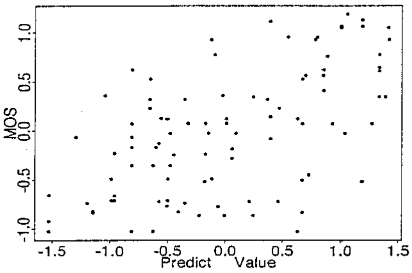

図 2.4: Relationship between MOS and the predicted value by the regression tree for the test data.

Japanese, we use /m/, /n/, and nasal vowel /N/ in this paper. There are two categories of data in the perception experiment, the first one is connection between /a/ and /'N' + '#'/ and the second one is connection between /a/ and /'m' + 'o'/, as shown in Figs. 2.5 and 2.6.

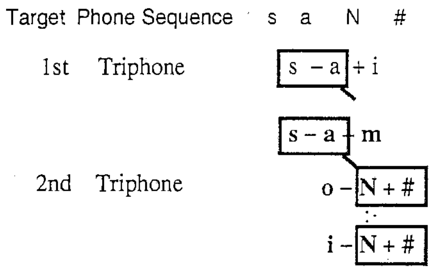

図 2.5: Description of /s-a : N-#/ concatenation.

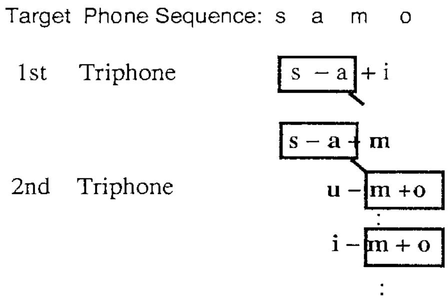

図 2.6: Description of /s-a : m-o/ concatenation.

The same database MHT is used and 73 samples are generated. Five listener evaluated the discontinuity perception with MOS values.

## 2.2 Decision Tree for MOS Prediction

A decision tree is made to predict the MOS value from acoustic factors : *f0dist*, *pwrdist*, *f0zs*, *cepdist*. The tree obtained from 72 samples is represented as : Equation : MOS ¡- $(cepdist + f0dist + pwrdist + f0_{zs})$

Nodes of tree: 11

Residual mean deviance (RMD): 0.15 = 9.27 / 61

Distribution of residual:

| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|
| -1.1 | -0.24 | -0.01 | 0 | 0.19 | 1.13 |

Figure 2.7 shows the decision tree for the V-N case.

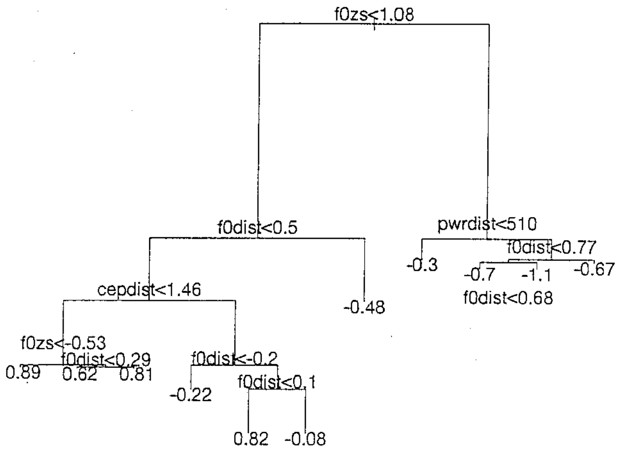

図 2.7: Dendrogram of the regression tree for prediction of MOS (node value).

For the training data, the predicted result of a close-test is shown in Fig. 2.8, and the correlation coefficient is 0.88. The correlation coefficient of an open test is 0.73.

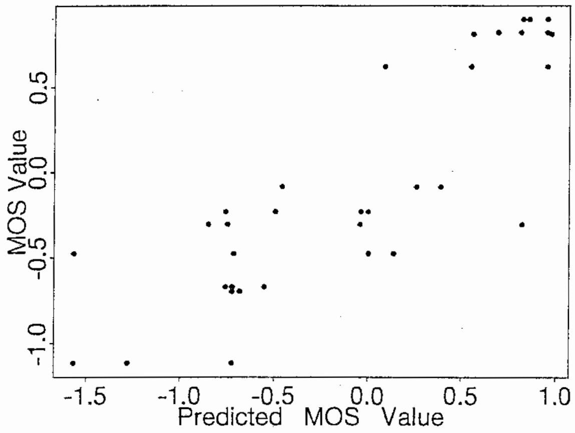

図 2.8: Relationship between MOS and the predicted value by the regression tree for the train data.

# 第 3 章

# Modifying Chatr F0 Pattern with PSOLA

Concatenative synthesis speech usually has two kinds of discontinuity: (1) Discontinuity at unit joint boundaries, which has been stated in the previous chapter. (2) Prosodic mismatch between unit and target f0 pattern, which will be explained in this chapter.

The technology to modify f0 pattern used this chapter is divided into three categories: (1) partial modification with f0 slope detection, (2) partial modification with DP search, (3) point-to-point f0 modification to target f0 pattern. All the tree methods will be described in the following sections.

## 1 Partial Modification with f0 Slope Detection

An example is given in Fig. 3.1 to show the target f0 pattern and unit f0 of CHATR speech. The difference between unit f0 and target f0 could be the course of wrong accent in the output speech.

The philosophy of partial f0 modification is to maintain natural quality of most part of units and use PSOLA to correct f0 values of those mora units which causing wrong accent type in their corresponding accent phrases. For Japanese, we perform partial f0 modification based on mor units, while in English, syllable structure is used. Then an objective criterion to detect the above "partial" mora units is needed. Here, a slope detection based measure is used because it is simple and efficient.

The flowchar of the approach is given in Fig. 3.2.

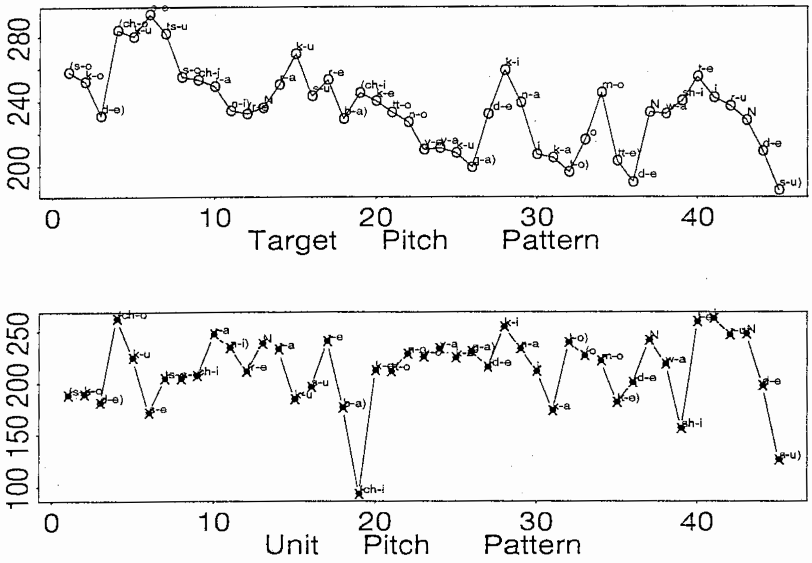

図 3.1: Target f0 and unit f0 of Chatr.

Text Input
↓
Text analysis & grapheme to phoneme
↓
Prosody Prediction (duration, f0, power)
↓
Unit Selection
↓
Partial f0 Modification
↓
Output Speech

図 3.2: The flowchart with the proposed method.

## 1.1 Objective Criterion of f0 Slope Detection

We try to detect mora pair with wrong f0 slope compared with its predicted f0 partner. The function could be the following equation:

$$f0_{slope\_range} = \left\{ \begin{array}{l} logf0_{slope} + \delta_{up} \\ logf0_{slope} + \delta_{down} \end{array} \right. \qquad (3.1)$$

$$\delta_{up} \quad = \quad f(Phr_{comm}, Acc_{comm}, Pos_{phr}) \qquad (3.2)$$

$$\delta_{down} \quad = \quad g(Phr_{comm}, Acc_{comm}, Pos_{phr}) \tag{3.3}$$

where $Phr_{comm}$ : phrase command, $Acc_{comm}$ : accent command, $Pos_{phr}$ : position in the phrase.

But we can not get these values from current prosody module in Chatr. In the current experiment, $\delta_{up}, \delta_{down}$ are set by manual. It can be improved only if we perform some perceptual experiments.

A more detailed description of this method is shown in Fig. 3.3. After detecting the mora pair $(m_{i-1}, m_i)$ whose f0 slope needs to be modified, we also need to know which mora should be repaired. Current approach is to modify the mora with f0 lower than $average_f0$ of units, otherwise modify $m_i$.

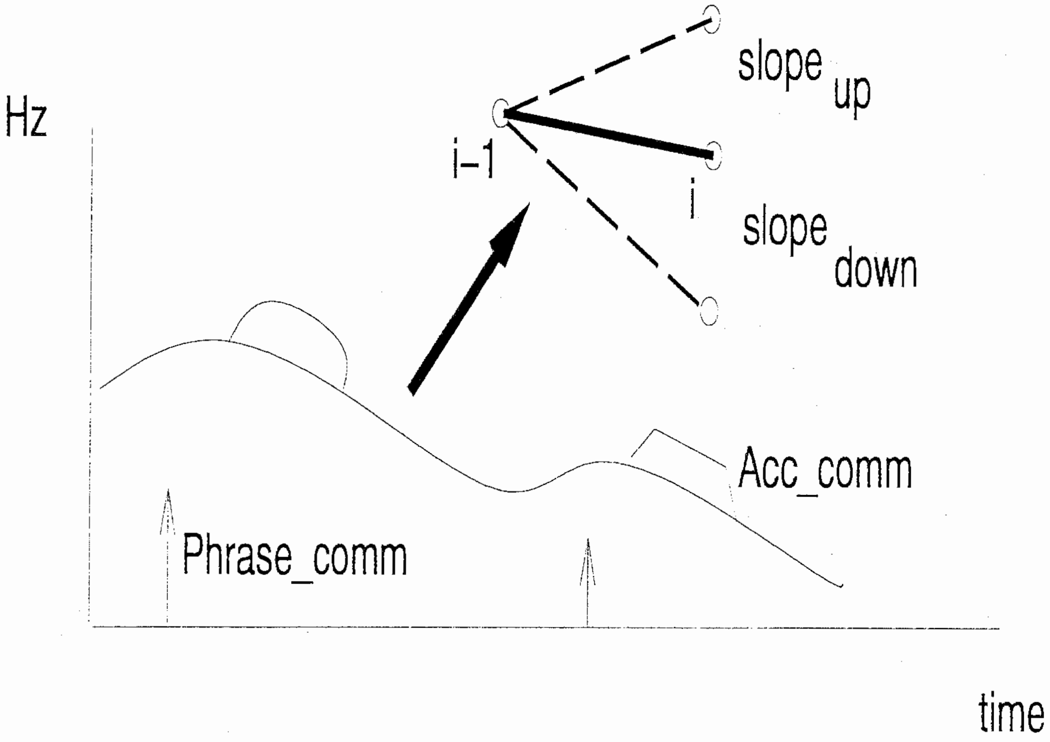

図 3.3: Detection of "partial" mora units.

## 1.2 Examples

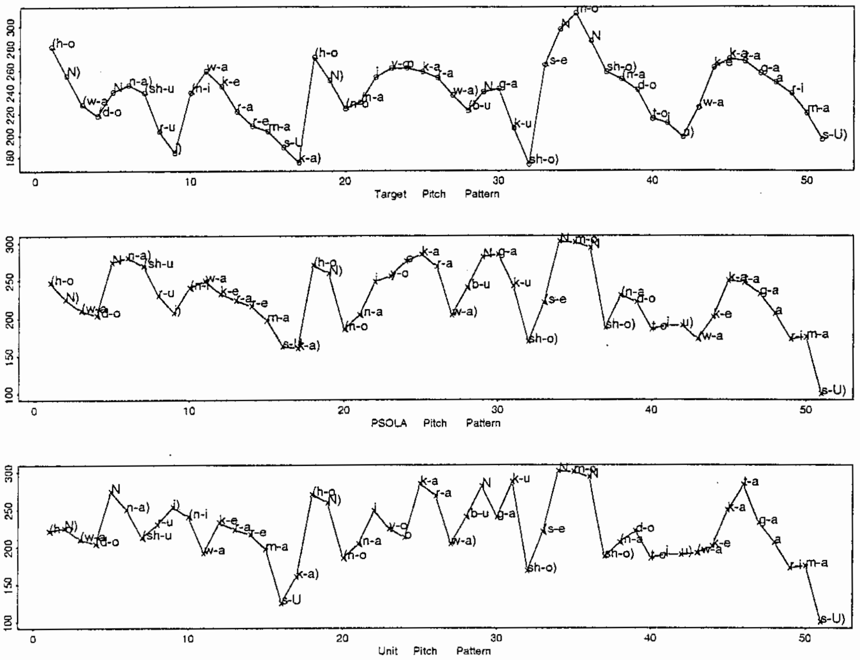

図 3.4: 本はどんな種類に分けられますか？本の内容からは、文学書、専門書などという分け方があります。

## 2 Partial Modification with DP Search

The second method to improve prosody of CHATR output speech is to use a MOS decision tree and DP search within reasonable f0 modification range of PSOLA.

The basic idea of this approach is illustrated in Fig. 3.6. Here the f0 modification range is set to 20%. Any unit f0 is supposed to be modified within this range and several discrete values in the range become f0 candidates of the unit. Then DP is used to find the best path through these candidates to minimize a predefined cost of the path.

The cost function consists two parts : (1) a quality cost introduced by PSOLA, named *psola_cost*, (2) a prosody cost means goodness of f0 pattern of the underlying phrase, which is measured by a MOS decision tree, named *prosody_cost*. They are explained in the following subsections.

Therefore, DP search is employed to minimize the following cost function:

$$cost = psola\_cost * weight + prosody\_cost \tag{3.4}$$

$$min\_cost = \min_{i=1,N^M} cost \tag{3.5}$$

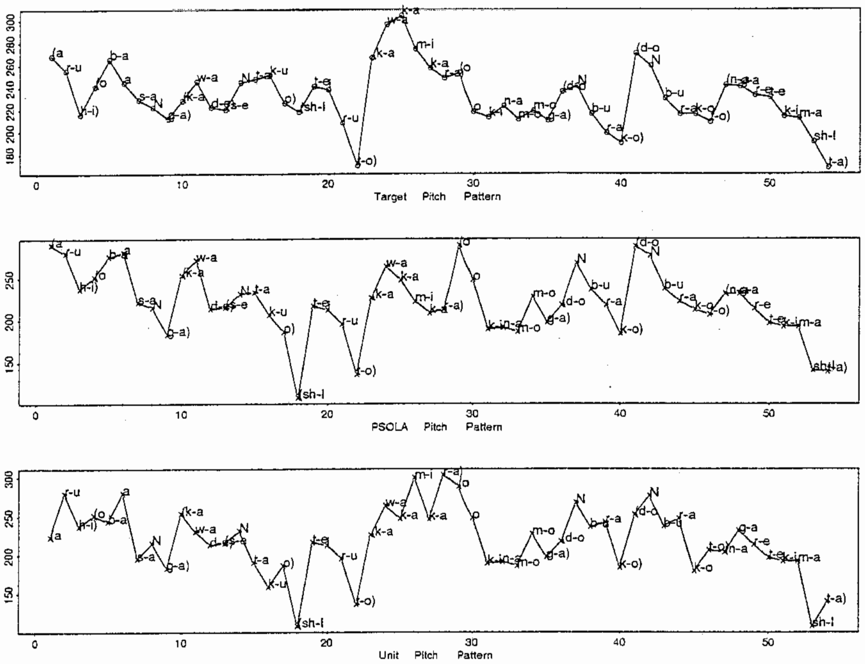

図 3.5: ある日、お婆さんが川で洗濯をしていると、川上から大きな桃がドンブラコ、ドンブラと流れてきました。

where *weight* is used to take the balance between prosody and quality.

## 2.1 PSOLA Cost

The relationship between f0 modification rate and quality degradation is appropriated by $y = ax^2$. where $x$ means maximum f0 modification rate by PSOLA (±0.25). $y$ means quality degradation caused by f0 modification(MOS_cost), which is the PSOLA cost. $a$ is the coefficient.

## 2.2 Prosody Mismatch Cost

Prosody mismatch cost is assumed to be a perception-related value obtained by experiments (refer to $TR - IT - 0276$). The acoustic features used are :

$$logslope^i_{target} = \log f0^{i+1}_{target} - \log f0^i_{target} \tag{3.6}$$

$$logslope^i_{unit} = \log f0^{i+1}_{unit} - \log f0^i_{unit} \tag{3.7}$$

$$Logslope_{distance} = \frac{\sqrt{\sum_{i=1}^{N-1}(logslope^i_{target} - logslope^i_{unit})^2}}{N-1} \tag{3.8}$$

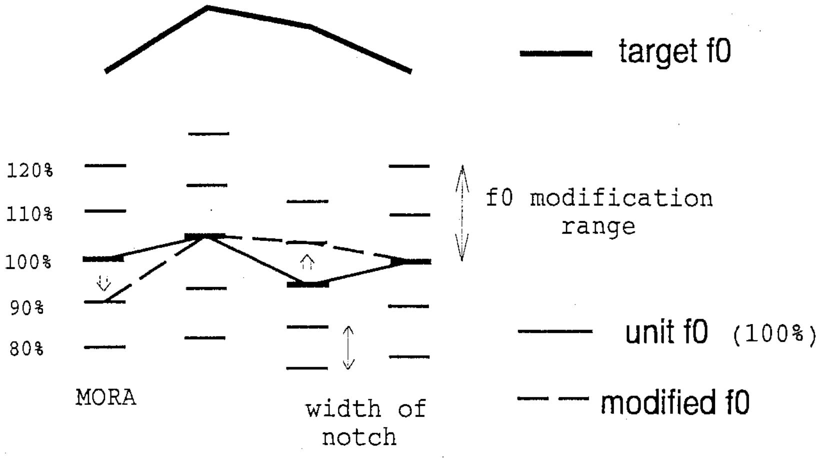

図 3.6: Searching the best path with PSOLA candidates and DP.

$$Delta = \max_{i=1,M} |logslope^i_{target} - logslope^i_{unit}| \tag{3.9}$$

where $N$ is number of mora in the phrase.

From several experiments of evaluating prosody perception in MOS values, we got a decision tree to predict the prosody mismatch cost from the above acoustic features. Figure 3.8 gives an illustration of the decision tree.

All the details of implementing this technique in CHATR system has been described in $(TR - IT - 0276)$.

## 3 Point-to-Point f0 Modification

Unlike the methods described in the previews sections, this section describes an approach of modifying f0 pitch-synchronously. Since the partial modification of f0 pattern usually modifies some parts of the selected units to maintain the voice quality and the limited f0 modification rate by PSOLA, prosody of the output speech could include some units which result in unnatural prosody. Although there exists a tradeoff between prosody and voice quality when using signal processing in f0 pattern, point-to-point f0 modification provide a powerful method to test the quality and prosody.

An example of the method is shown in Fig. 3.9.

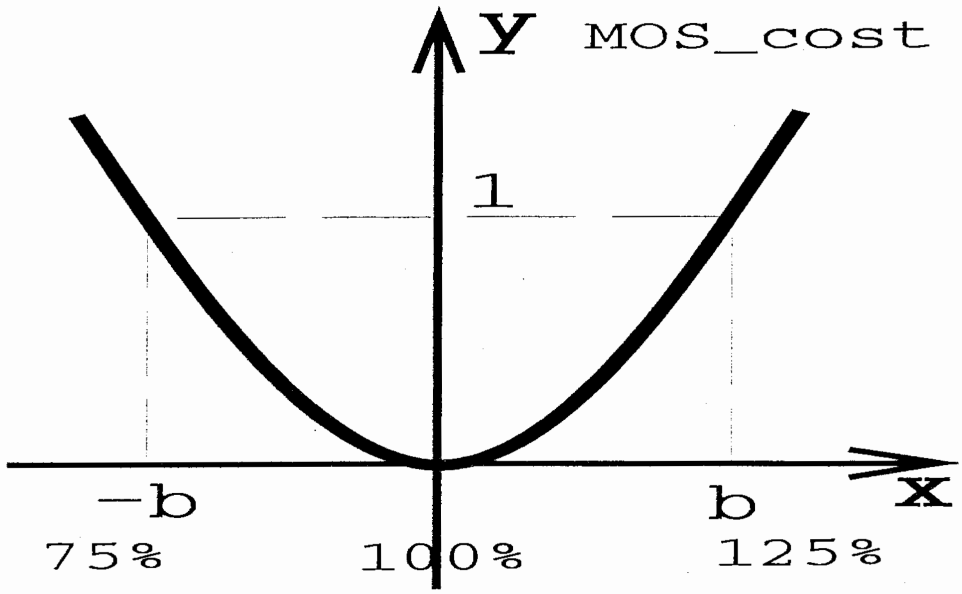

図 3.7: PSOLA degradation score of f0 modification. (after X. Marumoto, W. Ding , and N. Campbell, 1998)

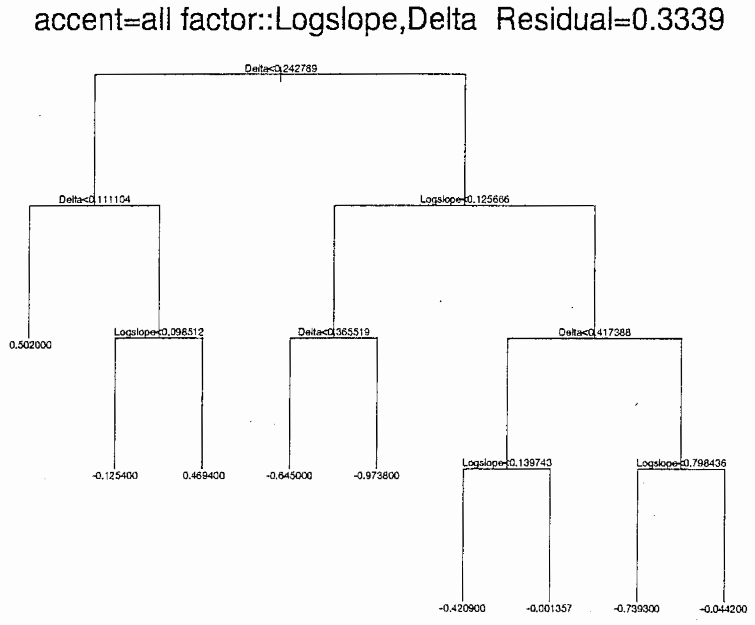

図 3.8: The decision tree used in Chatr.

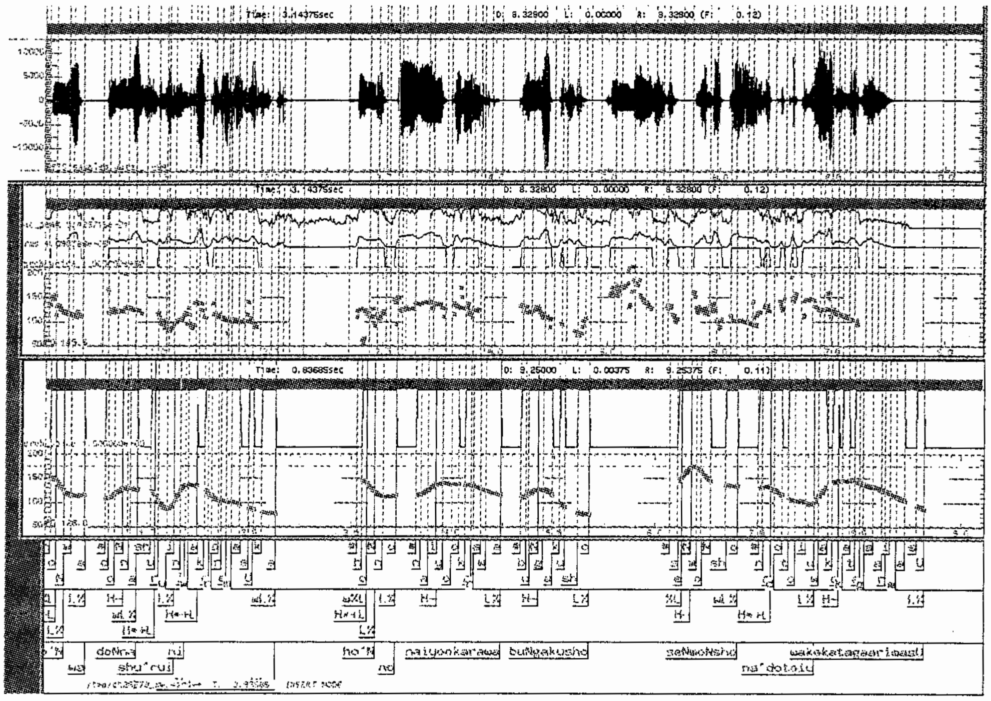

図 3.9: Point-to-Point f0 modification : middle panel is after modified f0 pattern, bottom panel is the target f0 pattern.)

# 謝辞

Many people deserve thanks for their kind support and useful discussions during my stay in ATR. In particular, I wish to express my sincere gratitude to Nick Campbell, supervisor & dept2 head, Norio Higuchi, previous dept2 head of ITL, for the guidance and support. Special thanks must also go to Yoshinori Sagisaka, dept1 head of ITL, for invaluable comments and guidance to me.