IMPROVING SPEECH SYNTHESIS OF CHATR USING A PERCEPTUAL DISCONTINUITY FUNCTION AND CONSTRAINTS OF PROSODIC MODIFICATION

Weń Ding, Ken Fujisawa and Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Kyoto 619-02, Japan
E-mail: ding@itl.atr.co.jp

ABSTRACT

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.

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.

1 INTRODUCTION

In any concatenative TTS system, methods must be found to reduce discontinuities at unit boundaries. For the first issue, we believe that the use of perceptual-based unit selection is necessary since objective distance measure do not always coincide with human hearing. To do this, we first need a prediction model to quantitize the joint discontinuity in perceptual space. Then we utilize the quantitized perceptual distances directly in the unit selection.

In this paper, we describe results using a decision tree as the prediction model, taking multiple opinion score (MOS) values as input and predicting them from acoustic features of the candidate segments.

In order to train the decision tree, we use acoustic factors related to the joint discontinuity, difference of f0 and power, distance of cepstrum coefficients, and *z_score* of f0. The relevance of these factors was determined by separate experiments and the corresponding MOS values were derived by using tests with an ATR Japanese speech database (MHT) as described below.

For the second issue, we considered the effects of pitch scaling using signal processing to improve the accent perception of CHATR output. To avoid degrading the natural voice quality, the scale factor for signal processing should be within a reasonable range. In the CHATR unit selection process, we already select units from within a similar range based on absolute difference in F0 alone. In the present paper we report on the difference in perceived quality of the resulting speech when selection is performed by allowing free prosody targets and introducing limited signal processing.

The cost function of this unit selection module is represented as a sum of the perceptual discontinuity predicted by the above decision tree and the penalty for pitch modification; i.e., by adding the signal processing distance to the absolute F0 distance in the candidate selection. This procedure allows us to modify the selected units slightly, to bring them within perceptual limens, without introducing unnecessary degradation to the signal.

2 BOUNDARY DISCONTINUITY

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/.

2.1 V-V Connection

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.

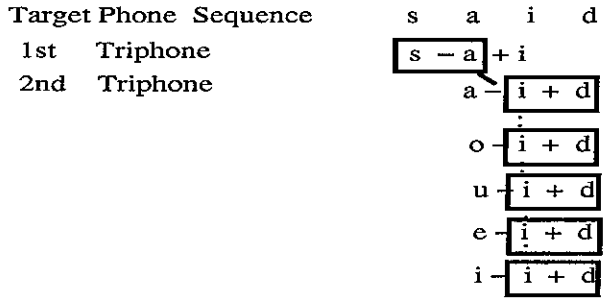

Figure 1: Description of triphone concatenation.

Considering a target sequence /s-a-i-d/ as shown in Fig. 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.

2.1.1 Experiment

A male Japanese database MHT containing 503 phoneme balance sentences was used. Considering the triphone context coverage and available data, we produced 120 samples including V-V connection, e.g., /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/, where x, y denotes other phoneme respectively. Five listeners judged the discontinuity in MOS value (-2, -1, 0, 1, 2) for all the samples.

The acoustic features at the phone boundary include $logF0$ distance ($f0dist$), power distance ($pwrdist$), cepstrum distance ($cepdist$), $F0$ $zscore$ ($f0_{zs}$) at the unit boundaries.

$$f0_{zs} = \frac{f0 - \overline{f0}}{\sigma_{f0}} \quad (1)$$

where $\overline{f0}$、σ_{f0} denotes the f0 mean and std of the phoneme, respectively, and $f0_{zs}$ is the larger value of two joint phones.

2.1.2 MOS Prediction

Then a decision tree model is established to predict the MOS value using the above acoustic features:

Equation : MOS ~ $(cepdist + f0dist + pwrdist + f0_{zs})$

Nodes of tree: 16

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

Distribution of residual:

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

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

where N denotes the number of observations, y_i denotes the known MOS value、$\tilde{y}_i$ denotes the predicted MOS value.

A section of the tree is shown in Fig. 2 (resolution at the leaves of the tree is sacrificed to provide a visual cue of split importance).

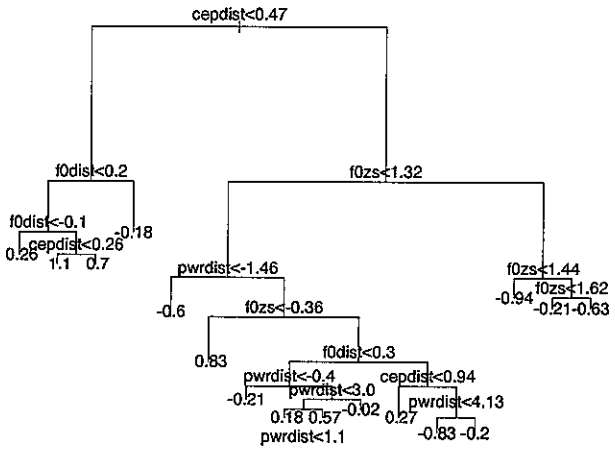

Figure 2: Dendrogram of the regression tree for prediction of MOS (node value).

For the training data, we performed a close-test using the decision tree and the result is shown in Fig. 3. The correlation coefficient is 0.79. The correlation coefficient for open test data is 0.58.

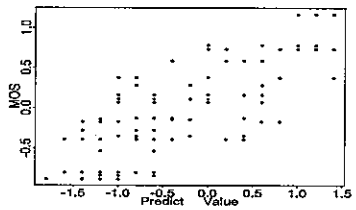

Figure 3: Relationship between MOS and the predicted value by the regression tree for the train data.

2.2 Vowel-Nasal Connection

In the same way, we investigate the vowel-nasal case. Since /N/ is a nasal vowel in Japanese, there are two subcategories, V-Nasal consonants and V-Nasal vowel, as shown in Figs. 4 and 5.

2.2.1 Experiment

The same database MHT was used and 72 samples were generated for experiments. The MOS values were obtained by the five listeners. And the acoustic features at the unit boundaries were the same as the V-V case.

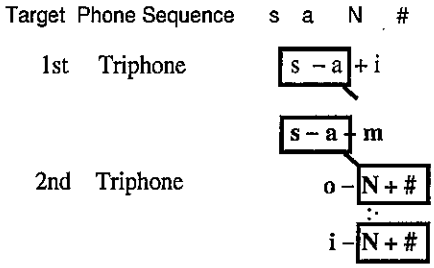

Figure 4: Description of /s-a : N-#/ concatenation.

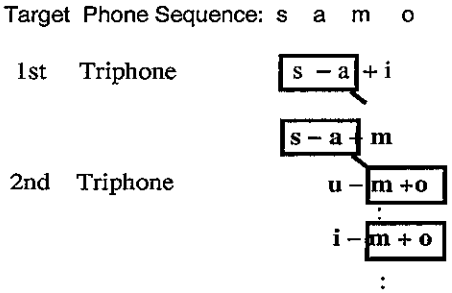

Figure 5: Description of /s-a : m-o/ concatenation.

2.2.2 MOS Prediction

A decision tree was trained to predict the MOS values from the acoustic values as follows:
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 6 shows the decision tree for the V-N case.

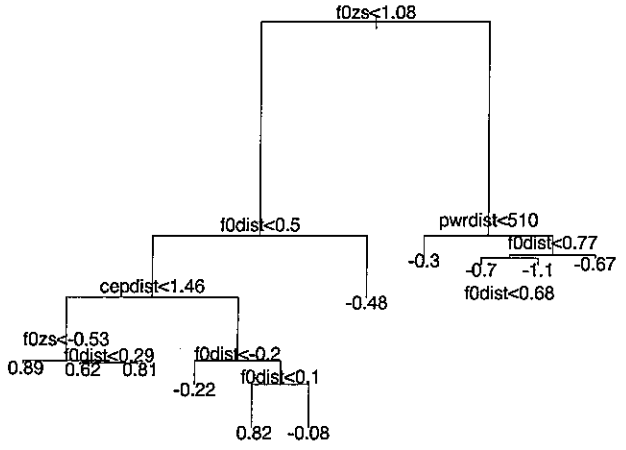

Figure 6: Dendrogram of the regression tree for prediction of MOS (node value).

For the training data, the predicted result of a close-test is shown in Fig. 7, and the correlation coefficient is 0.88. The correlation coefficient of an open test is 0.73.

2.3 Discussion

The vertical length of the two decision trees in Figs. 2 and 7, indicates the split importance of the acoustic factor. Therefore, it is shown in Fig. 2 that *cepdist* and $f0_{zs}$ are important to predict MOS values in V-V case, while Fig. 7 shows $f0_{zs}$ the most important factor to MOS prediction. Then the next issue is how to use these two models in unit selection.

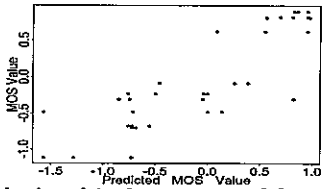

Figure 7: Relationship between MOS and the predicted value by the regression tree for the train data.

3 UNIT SELECTION

In phoneme based synthesis system CHATR, phoneme sized units which make up the phone sequence for an intended utterance are selected from a speech database based on a *target cost* and a *concatenation cost*. The *target cost* measures to what degree the selected units match the predicted ones in perceptual terms.

3.1 Concatenation Cost

In the current system, the *concatenation cost* measures the acoustic discontinuity between two adjoining phone-units, which is computed from distance of f0, ceptral coefficients and power. In this paper, we propose to use the perceptual based cost to compute *concatenation cost* and this can be done by using the above MOS prediction decision tree. Then the cost is decided by the predicted MOS value :

$$concat\_cost = \begin{cases} mos1 & \text{if } V-V \ \ case \\ mos2 & \text{if } V-N \ \ case \\ 0 & \text{other cases} \end{cases}$$

In such a way, concatenation cost incorporating the MOS value replaces the previously used acoustic vector and shows a perception-based performance. Also only one weight needed for the new cost compared with several weights for the previous one.

3.2 Target Cost

The features used to compute the *target cost* include phonetic context, duration, log power and mean f0. The *target cost* is the weighted sum of the difference between the feature vector of the target segments and candidate phoneme units.

In this paper, we consider the use of signal processing to modify the selected units with poor prosody score. The valid candidates of unit are constrained by a reasonable range of f0 and duration modification for the signal processing, e.g., within 20% f0 modification for PSOLA. So in the first stage of unit selection, i..e, determination of target cost, we use a pre-selection technique to select the units within the desirable modification range.

4 PARTIAL PROSODY MODIFICATION

Without signal processing, the concatenated units could generate natural sound but the prosody is uncontrollable. With the current system, a limited database can result in some sentences with wrong accent types and strange prosody. This is the motivation of partial prosody modification, which modifies the f0 and duration of only those mora units that cause wrong accent type in their corresponding accent phrases.

We adopt mora (CV or V in Japanese) as the basic unit for partial modification, since mora is widely accepted as the basic prosodic unit in Japanese.

The partial f0 modification is positioned as a post-processing in the whole process as follows,

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

The next issue is how to determine the faulty accent type and position.

4.1 Objective Criterion of Slope Detection

Although a target pitch contour for input text can be predicted correctly, the pitch of selected units does not always follow it because of limited database and no signal processing.

Detecting a mora pair with the wrong f0 slope compared with the predicted f0 contour is a straightforward way to detect wrong accent type and position:

$$f0_{slope\_range} = \begin{cases} logf0_{slope} + \delta_{up} \\ logf0_{slope} + \delta_{down} \end{cases} \quad (3)$$

$$\delta_{up} = f(Phr_{pos}, Acc_{pos}, f0_{zscore}) \quad (4)$$

$$\delta_{down} = g(Phr_{pos}, Acc_{pos}, f0_{zscore}) \quad (5)$$

where $logf0_{slope} = logf0_i^{targ} - logf0_{i-1}^{targ}$, Phr_{comm} : phrase position, Acc_{comm} : accent position, $f0_{zscore}$: relative height of f0 of the mora.

But we can not get these values directly from the current prosody module. In this paper, $\delta_{up}, \delta_{down}$ are set manually. These values can be improved only if we perform some perceptual experiments.

When $logf0_{slope}$ of units goes beyond the above slope range, the mora pair is found for partial f0 modification.

Figure 8 gives an example of a Japanese sentence with and without partial f0 modification with PSOLA. The mora units within a parenthesis represent a accent phrase. After pitch modification, most pitch accents have been improved and the output speech perceived to have a good prosody, and almost no quality degradation has been introduced. Because of the f0 modification range of PSOLA (20% is the current setting), the target pitch contour is not realized exactly by the partial modification.

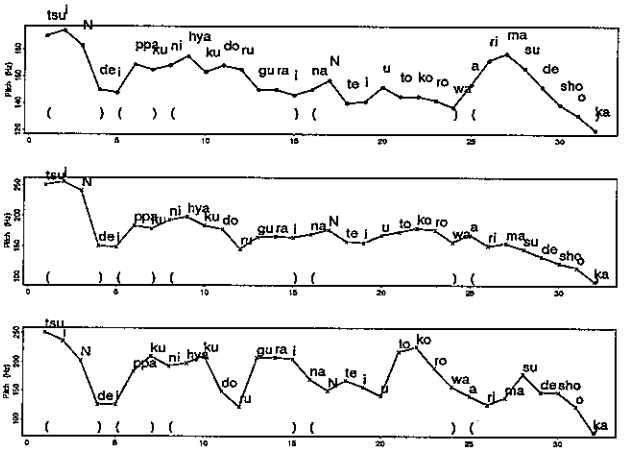

Figure 8: Partial f0 modification with PSOLA. (top: target pitch, middle: after partial modification (R43_01.WAV), bottom: no signal processing (R43_02.WAV).

5 CONCLUSIONS

This paper proposed methods to deal with two kinds of synthesis problems : boundary discontinuity and wrong accent perception. A perception based measure for unit selection was proposed to solve the first issue. The concatenation cost function was represented as the MOS values predicted by decision trees. Experiments were carried out to construct these MOS prediction models. For the second issue, we tried to maintain the natural quality of output speech while improving the accent perception by using partial pitch modifications. An objective criterion was used to detect the position of partial modification and showed a improved results in prosody of the output speech. Further work is required on autodetermine the threshold for the partial modification and on novel synthesis techniques to enlarge the acceptable modification range of prosody.

References

[1] W. N. Campbell & A. W. Black, "Prosody and the selection of units for speech synthesis", pp 279-292 in *Progress in Speech Synthesis*, eds Santen , Olive, Hirschberg & Sproat, Springer New York, 1996

[2] W. Ding and N. Campbell, "Optimising Unit Selection with Voice Source and Formants in the CHATR Speech Synthesis System", *Proc. EuroSpeech*, Rhodes, Greece, 1997.