OPTIMISING UNIT SELECTION WITH VOICE SOURCE AND FORMANTS IN THE CHATR SPEECH SYNTHESIS SYSTEM

Wen Ding and Nick Campbell

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
ding@itl.atr.co.jp

ABSTRACT

High quality corpus-based synthetic speech requires minimization of prosodic and acoustic distortions between an ideal phoneme sequence and the actual waveform segments used to reproduce it. Our synthesis system concatenates phoneme-sized waveform segments, without signal processing, selected from a large-scale speech database according to both prosodic and phonetic contextual suitability. This paper describes an approach to optimising such unit selection in speech synthesis by using voice source parameters and formant information, instead of selection based on cepstral features. We present results showing that formants and voice source parameters are more effective as acoustic features in the unit selection. These features can be estimated automatically from speech waveforms using the ARX joint estimation method. Results are compared with mel-frequency cepstrum coefficients (MFCC), previously used for unit selection, and both objective and subjective experiments showed that the new features outperformed the previous ones, and confirmed that the synthesized speech sounded much more natural.

1 INTRODUCTION

Unit selection is a key issue for concatenative TTS synthesis. For any given speech sound, a large number of potential equivalents can be found in a typical speech database, but each will be influenced by the various utterance contexts in which they are embedded, having different prosodic and glottal source characteristics, and resulting from subtly different configurations of the vocal tract. Even slightly different recording conditions can have an effect on the various waveform segments. When selecting and concatenating the required phoneme units for a given input text, the basic principle is to select units for both smooth continuity of acoustic features and maximum closeness to the desired prosodic context [1]. This type of unit selection depends on defining a cost function to minimize the distortion and optimising the kinds of measures used in the determination of weights between the different features.

Two types of distortion measure are used in unit selection for the current CHATR speech synthesis system [1, 2]. One is the unit distortion (*target cost*) which represents the distance between the predicted target segments and the candidate phoneme units in the database. The other is the continuity distortion (*concatenation cost*) which is defined as the distance between two adjacent selected units. The task of unit selection is to minimize the total distortion of the two weighted cost functions. In the *concatenation cost* and the training-weights procedure, the acoustic features previously used were MFCC, log power duration and pitch [1]. Although the acoustic distortion can be measured using many kinds of acoustic features derived from the speech waveform, the ideal features should be consistent with human perception.

Since the voice quality and phonation types are greatly influenced by the voice source, which can be characterized by the behavior of the vocal cords, the formants, and the vocal tract [4, 5], this paper focuses on using these new features in the unit selection and presents a comparison with selection using MFCC.

2 FEATURE EXTRACTION

Previously, 24 MFCC parameters were calculated with a 21.3 ms window length and a 5 ms shift length, and the first 12 parameters were considered in the cost functions. Vector quantisation was performed on every MFCC frame, to reduce to a single byte of storage space the spectral information of each frame. Power was calculated for each 20 ms windows with

a 10 ms frame shift. Log-power was used in the distance functions.

The glottal waveform $u(n)$ can be approximated by a parametric Rosenberg-Klatt (RK) source model [5]. The four parameters of the RK model include fundamental frequency (f_0), amplitude of voicing (AV), open quotient (OQ) and spectral tilt of the glottal waveform (TL). All the RK parameters and formants are estimated automatically by the ARX joint estimation method [6], which uses Kalman filtering and simulated annealing. The ARX model has the following formulation,

$$\sum_{i=0}^{p} a_i(n)s(n-i) = \sum_{j=0}^{q} b_j(n)u(n-j) + \varepsilon(n), \quad (1)$$

where $s(n)$ and $u(n)$ represent the speech signal and the glottal waveform at time n, respectively. $a_i(n)$ and $b_j(n)$ are model coefficients. p and q are analysis orders, and $\varepsilon(n)$ is an equation error. The only observed signal is speech signal $s(n)$. The input signal $u(n)$ is assumed as the glottal waveform which is generated by the RK model including four parameters. In such a case, the four unknown source parameters are optimized using simulated annealing, and the time-variant model coefficients $a_i(n)$ and $b_j(n)$ are estimated using Kalman filter. The two approaches are integrated into a recursive procedure based on a mean-square error criterion.

3 UNIT SELECTION

In our phoneme-based synthesis system, phone-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. The *concatenation cost* measures the acoustic discontinuity between two adjoining phone-units.

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:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \quad (2)$$

where p denotes the dimension of the feature vector. Currently, p is set between 20 and 30.

In this paper, in place of the mel-cepstral coefficients, we tested three sub-set feature vectors represented as

$$C_1^c(u_{i-1}, u_i) = pwr,$$
$$C_2^c(u_{i-1}, u_i) = f0,$$
$$C_{3\_new}(u_{i-1}, u_i) = \{AV, OQ, TL, F_1, \cdots, F_5, B_1, \cdots, B_4\}$$

where pwr is log power.

For comparison, the old feature vector including 12-MFCC is represented as

$$C_{3\_old}(u_{i-1}, u_i) = \{c_1, \cdots, c_{12}\}.$$

Then the *concatenation cost* is calculated as

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{3} w_j^c C_j^c(u_{i-1}, u_i). \quad (3)$$

The unit selection is performed with a pruned Viterbi search based on minimization of the total cost function:

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i). \quad (4)$$

4 WEIGHT TRAINING

The remained important issue concerns how to train the weights w_j^t and w_j^c. We assume that every acoustic feature is linearly related and contributes somehow to the prediction of voice quality in the concatenative synthesis. The weights w_j^t and w_j^c are optimized using linear regression separately [2]. In a previous experiment [3], various MFCC features were used to predict the quality of concatenation of units in the CHATR system. The analysis suggested that a linear combination of cepstral distance and power difference at the join point may be a useful predictor of the perceptual quality of isolated words. The weights of concatenation cost, w_j^c, were determined in the experiment.

For the weights of the target cost, w_j^t, an objective distance measure D_{obj}^n is defined to approximate the difference between natural and synthesized utterances from speech database, with the constraint that the objective distance measure should be as consistant with human perception as possible. In the current CHATR synthesis system, the cepstral distance between the waveforms is used as the objective distance measure [1].

$$D_{obj}^n = \sum_{j=1}^{n} (MFCC_j^{org} - MFCC_j^{syn})^2. \quad (5)$$

The weights are grouped into different sets according to the class of phoneme (e.g. all nasals, back plosives, low vowels). In this paper, the objective function is defined based on the coefficients of glottal source and formant (CSF) parameters:

$$D_{obj}^n = \sum_{j=1}^{n} (CSF_j^{org} - CSF_j^{syn})^2, \quad (6)$$

where
$CSF = \{AV, OQ, TL, F_1, \cdots, F_5, B_1, \cdots, B_4\}$. All the weights are determined based on the objective function and multiple linear regression:

$$D^n_{obj} = \sum_{j=1}^{p} w^t_j C^t_j(t_i, u_i) \quad (7)$$

The weights are trained by minimizing the prediction error of D^n_{obj} based on mean-square sense.

5 EVALUATION

Evaluation experiments have been carried out to verify the validity of the new features, i.e., formants and glottal source parameters $C_{3\_new}(u_{i-1}, u_i)$ compared with mel-cepstral features $C_{3\_old}(u_{i-1}, u_i)$ according to objective criteria and a subjective hearing test.

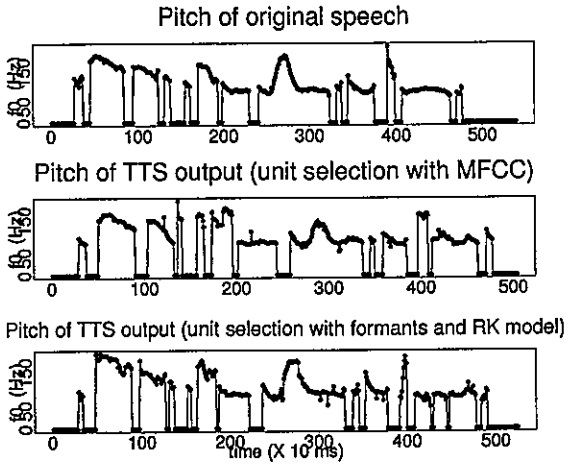

Figure 1: Pitch contours of original utterance and two TTS output speech.

Speech databases from two male Japanese speakers (MHN and MHT) were used for training with both the new and old acoustic features. Each database consists of 503 sentence-length phoneme-balanced utterances. We prepared two different databases for each speaker, where the weights were trained by using two kinds of acoustic features separately. The synthesised sentences were generated from each trained speech database.

The target utterances consisted of the original sentences in the database, each being excluded in turn from the corpus during synthesis so as to avoid selecting the original unit segments. The CHATR system resynthesized each target sentence to produce 503 synthesized sentences for objective comparison with the corresponding original natural speech.

As an example, Fig. 1 illustrates the pitch contours of CHATR output waveforms generated by the two kinds of unit selection. It can be seen that the pitch contour of the old features has some large jumps that are not present in the original. This kind of pitch pattern jump generally produces an unnatural perception of intonation and can result in ambiguity of word meaning in Japanese. The pitch contour of the new features has an envelope close to the original pitch, but some still shows mismatch in fine detail since no pitch modification is performed in our system.

To perform an objective evaluation, we calculated the $f0$ differences for the 503 sentences between the target pitch and the pitch of the synthesized sentence. Each sentence in turn is selected as the target and then excluded from the source speech database. The target utterance is synthesised using CHATR and the pitch difference of the two utterances is obtained. The distributions of average f_0 error computed from 503 utterances are shown in Fig. 2. It is shown that waveforms concatenated by unit selection with the new features gave better f_0 reproduction and result in smoother-sounding speech.

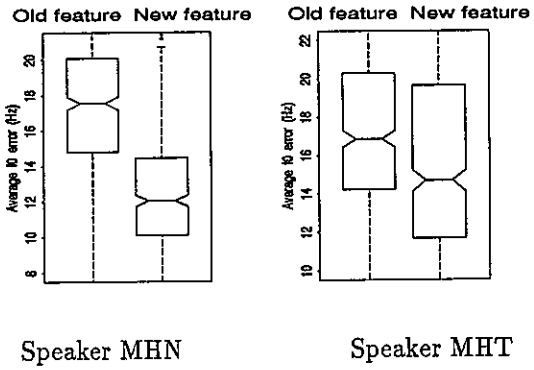

Figure 2: Average $f0$ differences between target and synthesized sentences.

Subjective evaluation was performed using 18 sentences synthesized from two male Japanese speakers MHN and MHT. An ABX listening test was carried out by presenting the 36 sentences to three subjects compared the original and synthesized sentences in order to select the ones which were perceptually closer to the original speech. The results shown in Fig. 3 confirm that sentences generated by the new features were preferred over the old ones. According to the hearing test, the new feature improved the voice quality of speaker MHN much more than that of speaker MHT. The speaking style of MHN is considered much more dynamic, while MHT speaks more smoothly in intonation, that may be the reason.

6 DISCUSSIONS

The comparison performed between the two kind

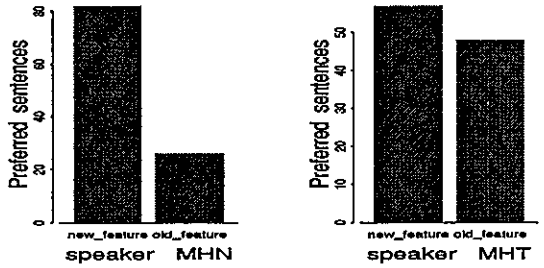

Figure 3: Hearing test (total 108 sentences) of two speakers: MHN and MHT.

of acoustic features is based on both subjective and objective measurements. The new features, glottal source and formants parameters, are well-known to possess a clear relationship to voice quality and human perception. It is clear from this study that the pitch reproduction in the synthesised sentences can be improved by using the new acoustic features but this improvement could not be obtained just by resetting the global selection weighting of pitch for all the phonemes to a larger value and using the old features or by training with other weights.

In the experiments, for simplicity, and in order to ensure that the new features be of the same dimension (12-dim) as the mel-cepstral parameters, formant bandwidths were included. Generally speaking, the difference of the formant frequencies can be considered to have more influence on voice quality than their bandwidth. This means that more weighting on frequencies may improve unit selection in future systems.

Although some pitch pattern jumps between units were observed when using the new features, the envelope of the pitch contour over the various phone units can be generated more naturally than before. The detailed pitch jumps could still be found at the join position of units, but on the whole the utterance seemed to be heard as more natural than when using the mel-cepstral parameters for unit selection. This kind of micro-discontinuity of pitch units is now being considered as suitable for localised pitch modification techniques.

In the current system, the unit selection is the key component and should be treated carefully. The weights for the target cost are trained with a form of linear regression, while the smaller number of weights for the concatenation cost function are now set heuristically by hand-tuning. Changing the acoustic features may change most of the weights of target cost, which results in better phoneme unit candidates be selected. More natural synthesis speech may be expected if the training of weights for concatenation cost can also be performed automatically,

7 CONCLUSIONS

We proposed the use of formant and voice-source information derived from the glottal source model as new acoustic features to be used in the unit selection module of the CHATR system. The new features including amplitude of voicing, glottal open quotient and glottal spectral tilt; formant parameters are used as the objective distance measure. The features they replace are 12 mel-scaled cepstral parameters. Two Japanese male speech databases were trained using the two features, and compared for selection by the new features vs. mel-scaled cepstral coefficients. The objective test using 503 sentences showed that using the new features the pitch contour of the synthesis utterance was more natural and the pitch differences between synthesis and original speech were reduced. Subjective hearing experiments also confirmed that significantly more of the sentences synthesized with the new features were preferred.

8 ACKNOWLEDGEMENTS

The authors would like to thank Dr. Yasuhiro Yamazaki, President of ITL-ATR, and Dr. Norio Higuchi, Head of Department 2 of ITL, for their kind support.

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 et al, , Olive, Hirschberg & Sproat, Springer New York, 1996

[2] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *ICASSP'96*, pp.373-376 (1996).

[3] A.J. Hunt and A.W. Black, "An investigation of the quality of concatenation of speech waveforms." Technical report, ATR interpreting Telecommunications Research Laboratories: TR-IT-0137, 1995.

[4] G. Fant , "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, pp. 119-156 (1996).

[5] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, Vol. 87, pp. 820-857 (1990).

[6] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model", *IEICE Trans. Inf. & Syst.*, E78-D, pp. 738-743 (1995).