FarsBayan: A Unit Selection based Farsi Speech Synthesizer

M. Mehdi Homayounpour, Majid Namnabat

Laboratory for Intelligent Sound and Speech Processing
Computer Engineering and Information Technology Department,
Amirkabir University of Technology (Tehran Polytechnics), Tehran, IRAN
Email: {homayoun, maj_nam}@ce.aut.ac.ir

Abstract

In recent years, the unit selection-based concatenative speech synthesis method using a large corpus has attracted great attention. This method provides more natural quality speech compared to the parameter driven methods. The Formant Synthesis, HNM method and use of MLSA filter are the prevalent methods for synthesizing Farsi speech. In this paper, we present the structure of a proposed unit selection synthesizer for Farsi language. In the proposed system, the linear regression method has been used for determination of weights of discrete sub-costs in the target cost, while the weights of other sub-costs have been considered constant. We have also presented a pre-selection algorithm using adaptive threshold for pruning the units. In addition, the efficiency of TD-PSOLA algorithm in improvement of resulting speech quality has been studied. Informal tests show the degrading effect of this algorithm on the output quality. The output speech was found to be remarkably fluent and natural. The quality of the output speech has been evaluated using MOS subjective test, and we have obtained a MOS test value of 3.8 for overall quality.

Index Terms: speech synthesis, unit selection, Farsi language, pre-selection using adaptive threshold

1. Introduction

The concatenative speech synthesis method based on unit selection has become one of the prominent technologies for Text-to-Speech (TTS) conversion in recent years. This technique has overcome the limitations associated with the diphone based methods, arising from the use of only one instance per unit. This problem is resolved by the use of a large database of continuous readout speech, and existence of a large number of stored instances per unit. The main components of this technique include a corpus containing variant instances, two criteria, namely target cost and concatenation cost, for evaluation of the instances, and finally a search algorithm for identification and selection of the best instances.

The target cost demonstrates the extent to which an instance from the corpus matches any particular target unit, while the concatenation cost determines the extent of discontinuity due to concatenation of two instances. The target cost is calculated as the linear weighted sum of the differences between the various prosodic and phonetic features of the instances under consideration and the target unit. Similarly, the concatenation cost is estimated as the linear weighted sum of the sub-costs such as the absolute differences in amplitude, F0, and the spectral discontinuity. Viterbi Search, for finding the path with the least cost in a network of target and concatenation costs, is used to determine the path consisting of the optimal instances.

ATR, as one of the pioneers in corpus-based speech synthesis technology, has made contributions to the progress of the technology through various studies, which lead us to the development of two TTS systems, namely ν-talk [1] and CHATR [2,3,4]. CHATR produced very natural speech in limited domains; however, the quality was unstable for unrestricted domains. CHATR directly or indirectly contributed to the development of the University of Edinburgh's FESTIVAL [5] and then AT&T's Next-Gen [6,7,8].

In this paper, specifications of the proposed synthesizer for Persian language, including the selected features for cost functions, the algorithms used for determination of the weights of the cost functions and pruning of instances, have been presented. Section 2, describes the corpus used, while the structure of cost functions including the considered features for these functions are presented in Sec. 3. The methods for determination of weights of target costs are given in Sec. 4, and the pruning algorithms used in this study are presented in Sec. 5. The results of the system evaluation are given in Sec. 6. The discussion of the results and outlines for future research are summed up in Sec. 7.

2. Speech Corpus

Corpus or unit inventory is one of the most important constituting components of the unit selection-based speech synthesis systems. Researches have shown that the output speech quality improves with the increased size of the corpus used. To build the corpus for this system, parts of the texts of large FARDAT corpus [9] were read by a single male speaker, and recorded at 16 kHz sampling frequency in a sound proof room.

This corpus was labeled at phoneme level by an automatic segmentation system, and segmentation of about one hour was corrected manually. Finally, a corpus containing 2 hours of speech and 63000 phonemes was obtained. These phonemes are used as the synthesis units. For each unit instance, a feature vector containing F0 values, amplitude, and duration has been obtained. Waveform files and the feature vector related to all instances are stored in our synthesis corpus.

3. Cost Functions

As mentioned earlier, the cost functions are weighted sum of a series of sub-costs, and each sub-cost relates to a feature which can be estimated using a distance function. The features and distance functions used in cost functions will be described in this section.

3.1. Target Cost

The target cost C^C, consists of two costs corresponding to phonetic features C_D^C, and prosodic features C_P^C. The features of the phonemes neighboring the main phoneme, and the position of the phoneme in syllable, were used as phonetic features. Position of a given phoneme in syllable containing this phoneme is one of phonetic features. Different features were used for consonants and vowels. For each neighboring phoneme, three features consisting of manner of articulation, place of articulation and voiced/unvoiced parameter are considered as features for consonants. There are 6 vowels in Farsi language, which are further subdivided into short and long vowels. Phonetic class of a vowel was used as one of the features for vowels. This feature determines whether this vowel is a short or a long vowel. Some other features such as being front or back and being high, mid or low are also used for vowels. F0 values, energy and duration were used as prosodic features. Three feature values were obtained for the F0. These values are estimated by dividing each phoneme into three equal parts and the average values of F0 for each part is estimated. Phoneme energy feature is estimated using equation (1). U_S is the number of first sample in a given phoneme signal, U_F is the number of the last sample, and S_i is the value of the i^{th} sample in the given phoneme.

$$E = \log\left(\sqrt{\frac{\sum_{i=U_S}^{U_F}(S_i)^2}{U_F - U_S}}\right) \tag{1}$$

3.2. Concatenation Cost

The concatenation cost, consists of three sub-costs related to discontinuities in F0, amplitude and spectrum, caused by concatenation of two instances at their boundary. The F0 and energy discontinuities are measured as absolute difference of pitches and amplitudes of the two instances at their boundary. The spectral discontinuity is calculated by determining the similarity measure between the end frames of the phoneme on the left side of the boundary and the boundary region of the previous context of right side phoneme of the concatenation. In this case, if U1 and U2 are two instances in the corpus, and PU2 is U2's previous instance in the corpus, then the discontinuity can be calculated by measuring the similarity between the two frames in the Overlap region. Figure 1 demonstrates this method. Mahalanobis distance is used as similarity measure between two overlapped regions. Each frame is represented using 12 MFCC coefficients and their first derivatives.

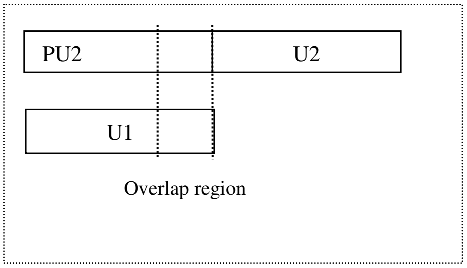

Figure 1: The method of spectral discontinuity calculation

Let N be the number of cepstral coefficients, $\hat{x}_i^j$ be the i^{th} normalized coefficient of the j^{th} instance, and μ_i and σ_i be the average and the standard deviation of the i^{th} coefficient, respectively. Then the spectral discontinuity is given by [10]:

$$Dis_{Spec}(U_1, U_2) = \frac{1}{N}\sum_i (\hat{x}_i^1 - \hat{x}_i^2)^2 \quad 0 \le i \le N \tag{2}$$

$$\hat{x}_i^t = \frac{x_i^t - \mu_i}{\sigma_i} \tag{3}$$

It should be noted that all the continuous features used in the target and concatenation costs are normalized, having zero averages and unit standard deviations. Application of average and standard deviation for normalization of the duration feature, results in obtaining small values for this sub-cost compared to other sub-costs. This is due to the similarity between duration of the instances of one phoneme, and the large difference between the intrinsic duration of various phonemes. Therefore, this feature is normalized separately for the instances of each phoneme, using the averages and standard deviations of the duration of instances of that phoneme.

4. Optimization of the Weights

Relying on experience, two sets of weights have been considered for the concatenation cost: one set for concatenation of every two instances of voiced phonemes, and another set for other concatenations. In order to determine the weights of the target cost, linear regression method has been used [3]. However, in the proposed system the sub-costs related to prosodic features are considered to be continuous, and the regression method is not very effective for the determination of weights of these sub-costs. Also, in this case the normalization of target cost values is difficult. In addition, it seems that the effect of the prosodic differences between an instance of a phoneme and a target unit, on the system output quality is a weak function of the phoneme type, when phoneme instance and target unit are both voiced or are both unvoiced. Hence, in the proposed system, regression is used only for determination of the weights of C_D^C cost. For C_P^C, depending on whether the phoneme is voiced or unvoiced, two sets of constants weights have been empirically obtained and used. For the regression purpose, the mean Euclidean distance between the time aligned vectors of 12 MFCC coefficients was used as objective distance measure. The following equation was used for combining the phonetic and prosodic costs and calculating the overall target cost:

$$C^{C} = \alpha * C_{P}^{C} + (1-\alpha) * C_{D}^{C} \tag{4}$$

5. Pruning Algorithms

In order to obtain a real-time synthesizer, the search domain is pruned in four steps. The first and second steps are the phonetic context pruning and pre-selection [11,12]. These prunings are performed at first inspection of the instances. The third pruning step is performed after calculation of the target cost. A beam width pruning is performed as the forth step [3]. In the phonetic context pruning, the phonemes are classified into four different groups according to the number of instances assigned to them in the corpus. Then at the time of searching for the corpus instances for a target phoneme with high occurrence frequency, if the phoneme on its left (or right) has a high occurrence frequency, only the instances which have the same left (or right) context will be analyzed. Also, if both phonemes adjacent to the target phoneme have high occurrence frequencies, only instances with the same left and right contexts will be analyzed.

In the pre-selection, an instance is accepted if the calculated value for C_{P}^{C} part of the target cost is less than a particular threshold. For each target unit, the number of similar instances in the corpus is different. In this case, if a constant threshold is used for some units, many of similar instances will be accepted. Therefore, each threshold was determined adaptively in accordance to the acceptance rate of the instances. At the time of pruning, considering the rate of acceptance, the threshold is changed in a manner that a fixed number of instances (C_{PS}) are always accepted for each target unit. A suitable rate of acceptance (SI_{j}) for any target unit j is calculated on the basis of number of instances $(NCps_{j})$ in the corpus for that target unit, using equation (5).

$$SI_{j} = \frac{C_{PS}}{NCps_{j}} \tag{5}$$

The actual acceptance rate of instances at the pre-selection step, after the acceptance of every N instances, can be calculated using equation (6).

$$SR_{j} = \frac{Number\ of\ accepted\ instances}{Number\ of\ instances\ processed\ for\ this\ unit} \tag{6}$$

The threshold adapts itself in a way that SR_{j} and SI_{j} always coincide with each other. Therefore, at each step (after the acceptance of every N instances), if the instances acceptance rate is bigger (or smaller) than the ideal rate for its target phoneme, the threshold is decreased (or increased) by a certain constant amount.

Four pruning algorithms were evaluated and the percentages of increase in the overall optimum path cost were calculated. It should be mentioned that pruning algorithms were applied successively and not separately. Pruning algorithms are listed in table 1. As it can be seen, pruning with target cost results a high increase in cost or a decrease in output signal quality. This is probably due to the incapability of the target cost in predicting the costs of concatenation of an instance with other instances, and the extent of smoothness of concatenations. In addition, pre-selection using adaptive threshold leads to a minimum quality degradation.

Table 1: Increase in the optimum path cost due to the application of different pruning steps

used pruning algorithm	Percentage of increase in the optimum path cost
Phonetic Context	1.1%
Pre-selection using adaptive threshold	0.3%
target cost	3.1%
beam width	0.6%

6. Implementation and Evaluation

Quality of output speech can be improved by using prosodic modification algorithms. In this research, Time-domain Pitch Synchronous Overlap and Add Method (TD-PSOLA) is used for prosodic modifications [13]. In this system, a better performance was achieved with no prosody modifications. However, using Overlap and Add method (OLA) in phonemes boundaries and amplitude corrections, improve output speech quality.

In order to evaluate the performance of the system, five-point MOS tests are used to measure the four criteria: overall speech quality, intelligibility, naturalness and pleasantness of the synthesized speech. For this purpose, 50 speech utterances, each of about 10 to 15 seconds duration, were re-synthesized using prosodic features of their original natural utterances. These utterances have been classified randomly into 4 groups, each consisting of 22 utterances. In addition to the synthesized utterances, five natural speech utterances were also added to each group. The utterances in each group were evaluated by 7 listeners. During tests, the listeners were asked to listen to each utterance only twice, and they were prevented from going back to the previous utterances and correcting the previous scores. After listening to each utterance, the listeners were asked to give a score to each utterance, decimal or integer and ranging from 1 to 5. The score of 4 listeners were omitted since their scores to the natural speech were found to be inappropriate i.e. their scores were less than 4 [14].

The average and standard deviation for the natural and synthesized utterances are given in Table 2 and Figure 2. The standard deviations show that a maximum difference of opinion exists among the test participants, as far as the scoring procedures for pleasantness and naturalness criteria are concerned. MOS standard deviations for natural speech is less than standard deviations for synthesized speech.

Table 2: Results of the system evaluation

Measure	Natural Samples		Synthesized Samples	
	MOS	Std	MOS	Std
Overall quality	4.7	0.34	3.8	0.56
Intelligibility	4.9	0.23	3.9	0.46
Naturalness	4.7	0.32	3.6	0.58
Pleasantness	4.5	0.45	3.5	0.59

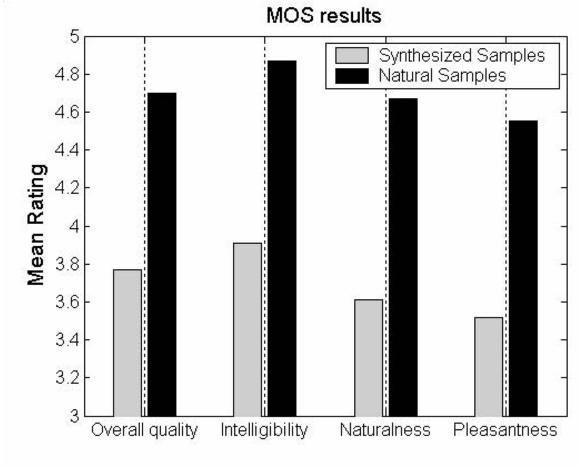

Figure 2: Mean of opinion scores for synthesized and natural samples

7. Conclusion and Future Research

In this paper, the structure of a synthesis engine for Farsi language, using unit selection method has been described. The system output quality is nearly natural and the MOS quantity for the overall quality criteria was found to be 3.8. In addition, a version of the pre-selection algorithm using adaptive threshold was introduced. It is believed that a consideration of the phonemes types in concatenation cost, and application of more appropriate criteria for evaluation of the spectral discontinuity will result in further improvements in the synthesized speech quality. It is also possible to have a more intelligent pruning algorithm, through the addition of sub-costs, such as splicing cost [15], to the target cost. Sub-costs are used for prediction of the extent of concatenation smoothness of one instance with respect to the other instances**.** Further improvement in the system output quality is also possible, through more precise manual segmentation of the corpus, and use of half-phone or di-phone synthesis units instead of phoneme.

8. Acknowledgement

This work was supported by Iran Telecommunication Research Center (ITRC).

9. References

[1] Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K., “ATR $-\nu-$TALK speech synthesis system”, ICSLP, 1: 483–486, 1992.

[2] Black, A., and Campbell, N., “Optimizing Selection of Units from Speech Databases for Concatenative Synthesis”, ICSLP, 581-584, 1995.

[3] Hunt, A., and Black, A., “Unit Selection in A Concatenative Speech Synthesis System Using A Large Speech Database”, ICASSP, 373-376, 1996.

[4] Campbell, N., and Black, A., “CHATR: a multi-lingual speech re-sequencing synthesis system”, Institute of Electronic, Information and Communication Engineers, 1996.

[5] Black, A., and Taylor, P., “Festival Speech Synthesis System: system documentation (1.1.1)” Human Communication Research Centre Technical Report HCRC/TR-83, 1997.

[6] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A., “The AT&T Next-Gen TTS System“, 137th Acoustical Society of America Metting, 1999.

[7] A. Conkie, “A robust unit selection system for speech synthesis,” 137th meet. ASA/Forum Acusticum, 1999.

[8] M. Beutnagel, A. Conkie, and A. Syrdal, “Diphone synthesis using unit selection,” 3rd International Workshop on Speech Synthesis, pp. 185-190, 1998.

[9] Bijankhan, M., Great Farsdat database, Technical Report, Research Center on Intelligent Signal Processing, 2002.

[10] Nukaga, N., Kamashida, R., and Nagamatsu, K., ”Unit Selection Using Pitch Synchronous Correlation for Japanese Concatenative Speech Synthesis”, 5th ISCA Speech Synthesis Workshop, 43-48, 2005.

[11] Conkie, A., Beutnagel, M. C., Syrdal, A., and Brown, P. E., “Pre selection of candidate units in a unit selection-based Text-To-Speech Synthesis System”, ICSLP, III: 314-317, 2000.

[12] Hamza1, W., and Donovann, R., “Data-Driven Segment Preselection in the IBM Trainable Speech Synthesis System”, ICSLP, 2609-2612, 2002.

[13] Moulines, E., and Charpentier, F., “Pitch-Synchronous Waveform Processing for Text-To-Speech Synthesis Using Diphones”, Speech Communication, 9: 453-467, 1990.

[14] Black, A., and Tokuda, K., “Blizzard Challenge-2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets”, Interspeech, 2005.

[15] Bulyko, I., Ostendorf, M., and Bilmes, J., “Robust Splicing Costs and Efficient Search with BMM Models for Concatenative Speech Synthesis”, ICASSP, 1: 461-464, 2002.