A QUANTITATIVE METHOD FOR MODELING CONTEXT IN CONCATENATIVE SYNTHESIS USING LARGE SPEECH DATABASE

Wael Hamza[1]*, Mohsen Rashwan*[2]*, Mohamed Afify*[3]

[1]IBM European Speech Research, IBM Egypt.

[2]Assoc. Prof., Cairo University.

[3]Bell Laboratories, Lucent Technologies.

ABSTRACT

Modeling phonetic context is one of the key points to get natural sounding in concatenative speech synthesis. In this paper, a new quantitative method to model context has been proposed. In the proposed method, the context is measured as the distance between leafs of the top-down likelihood-based decision trees that have been grown during the construction of acoustic inventory. Unlike other context modeling methods, this method allows the unit selection algorithm to borrow unit occurrences from other contexts when their context distances are close. This is done by incorporating the measured distance as an element in the unit selection cost function. The motivation behind this method is that it reduces the required speech modification by using better unit occurrences from near context. This method also makes it easy to use long synthesis units, e.g. syllables or words, in the same unit selection framework.

1. INTRODUCTION

Concatenative speech synthesis becomes the most common way to generate natural speech [1,2,3,4]. Using concatenative synthesis makes it easy to generate sounds that were very difficult to generate by rules. Concatenative speech synthesis is performed by pre-recording a set of speech segments and recalling them during synthesis. The problem with concatenative synthesis is that the acoustic features of the synthesis target often differ from those of the stored speech units. These acoustic features comprise the following two major features: (1) the vocal-tract shape which differs as a function of the neighboring context, (2) the pitch contour, the duration information, and the energy information that differ as a function of the overall prosodic information.

Two major techniques are proposed to solve this problem. The first technique is to modify the stored speech units to have the target acoustic features. Synthesis systems that use this technique store a small amount of synthesis segments and are known as small database synthesis systems. The second technique stores a large amount of synthesis units in order to cover all possible acoustic features and selects the best unit occurrences in order to resemble the required target. Systems that use this technique are known as large database synthesis systems [1,3,4]. In this work, we will concentrate on large database synthesis systems and we will propose a new method to model the context, the first acoustic feature mentioned above.

Most of the large database synthesis systems take a hard decision about the context feature. The problem with hard decisions is that the number of the resulting unit occurrences is too small to cover variability in other acoustic features like pitch and duration. In other words, pitch and duration modification should be performed in order to match the required pitch and duration. This reduces the naturalness of the resulting speech enormously. In the proposed method, no hard decision is taken. Instead, a distance measure between different contexts is used and incorporated in the unit selection cost.

The paper is organized as follows. In section 2, the overall synthesis system will be described. In section 3, the newly introduced context cost in presented. In section 4, results and discussion will be described.

2. OVERALL SYSTEM DESCRIPTION

As mentioned above, the system belongs to the family of concatenative speech synthesis systems that use large speech database. The system is described in more details in [4] and [5]. The system uses phones as a synthesis unit and will be described briefly here. Although the system was originally applied on Arabic language, it can be used for any other language.

2.1 Inventory Construction

The synthesis unit inventory is automatically constructed without any human segmentation or correction. A one hour of speech is recorded from one speaker. A three-state left-to-right continuous-density HMM is constructed for each phone. The phone HMMs are initialized and trained using the recorded database. The resulting models are called context-independent phone models. The context-independent phone models are used to align the recorded database generating the so-called "context-independent alignment". The context-independent alignment is used to grow up a top-down likelihood-based context-clustering tree for each state in the phone model. This operation will be described with some details in section 3. As a result of the context-clustering tree operation, context-dependent phone models are constructed. The resulting models are used to align the recorded database resulting in the so-called "context-dependent alignment". The resulting context-dependent alignment is more accurate than the context-independent one and is suitable for inventory construction.

Using the context-dependent alignment, synthesis unit occurrences are generated from the database. For each phone occurrence, information about phone identity, location, context information, energy information, pitch information, and duration information are stored. This information is used afterward in online unit selection. Detailed information about the whole operation can be found in [4].

2.2 Online Unit Selection

During online synthesis, the target phone sequence and the target pitch and duration information are passed to the synthesis module. A target sequence is constructed for the given target information. The target here represents information that is needed to synthesize a particular phone in the phone sequence. All phone occurrences that belong to a specific target are recalled constructing a phone occurrence trellis shown in Fig 1. A dynamic programming algorithm is preformed to choose the unit occurrence sequence that minimizes the cost function

$$C\left(t_1^n, O_1^n\right) = \sum_{i=1}^{n} C^t\left(t_i, O_i\right) + \sum_{i=2}^{n} C^c\left(O_{i-1}, O_i\right) \qquad (1)$$

where $C^t(t_i,O_i)$ is the target cost between target i and occurrence i along the path O_1^n and $C_c(O_{i+1},O_i)$ is the concatenation cost between two consecutive occurrences O_i and O_{I+1}. The target cost is defined to be the weighted summation of different cost elements

$$C^t\left(O_i, t_i\right) = \sum_{j=1}^{P} w_j^t C_j^t\left(t_i, O_i\right) \qquad (2)$$

It has to be mentioned that all phone occurrences that belongs to the target phone identity are recalled whether or not they belong to the same context. The context measure is then incorporated in the cost function mentioned in equation 2. This incorporation will be explained in details in section 3.

2.3 Speech modification

Although the resulting occurrence sequence is very near to the target, speech modification step should be performed in order to make it exactly matches the target sequence. The modification is done on the pitch and the duration and no vocal tract filter smoothing is performed. This is because the context is modeled in a way that results in smooth concatenative units. The speech is modified using Pitch-Synchronous All-Harmonic PSAH model [4]. This technique belongs to the sinusoidal modeling family. It can be interpreted as the analysis-by-synthesis overlap-add synthesis but working in pitch synchronous rate. This makes it easy to perform time-scale modification [4]. It can be also interpreted as the Harmonic + Noise model but using harmonics to model the whole speech spectrum. This eliminates the need of detecting the maximum voicing frequency needed by the latter model. A detailed description of the analysis, modification, and synthesis of this model can be found in [4].

As will be shown later, the pitch and time scale modification factors are reduced due to the borrowing mechanism used in the proposed context modeling method. This enables the usage of less complex time domain method such as TD-PSOLA without any loss of output speech quality.

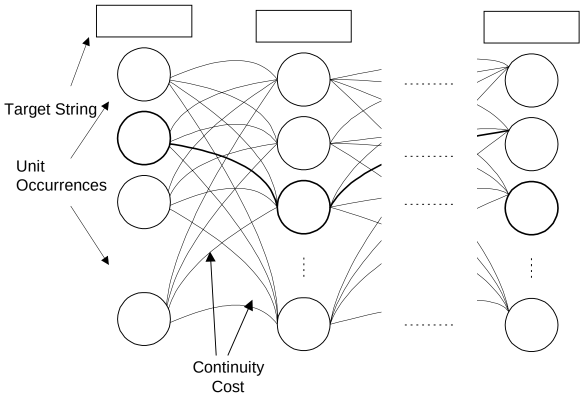

Fig. 1. Online unit selection

3. THE NEWLY INTRODUCED CONTEXT COST

3.1 Introduction and motivation

Figure 2 shows a close look to the calculation of the target cost function for one phone in the target phone sequence and the corresponding occurrences. The target function is defined as the weighted summation of target cost elements (see equation 2).

It is shown in the figure that the target here is defined as a combination of some acoustic features. Closer look at the context information in the figure shows that the context is defined as a set of leaf identification in the corresponding HMM-state context-clustering tree. In our case, it is defined with three numbers that correspond to the three leaf identifications in the 3-state HMM context-clustering trees. In other synthesis methods, only occurrences that match the target context exactly are recalled. In that case, the number of occurrences will be too small to cover other acoustic features of the target e.g. pitch and duration. It can be seen in the figure that O_i and O_{i+1} will not be recalled and occurrences like O_1 will only be recalled. The reader should note the pitch contour in both cases. One quick solution to this problem is to have a symbolic function that counts the leaf identification mismatch between the occurrence and the target. This sort of hard decision will result in the same output for differently context occurrences. For instance, O_i and O_{i+1} will have the same context distance.

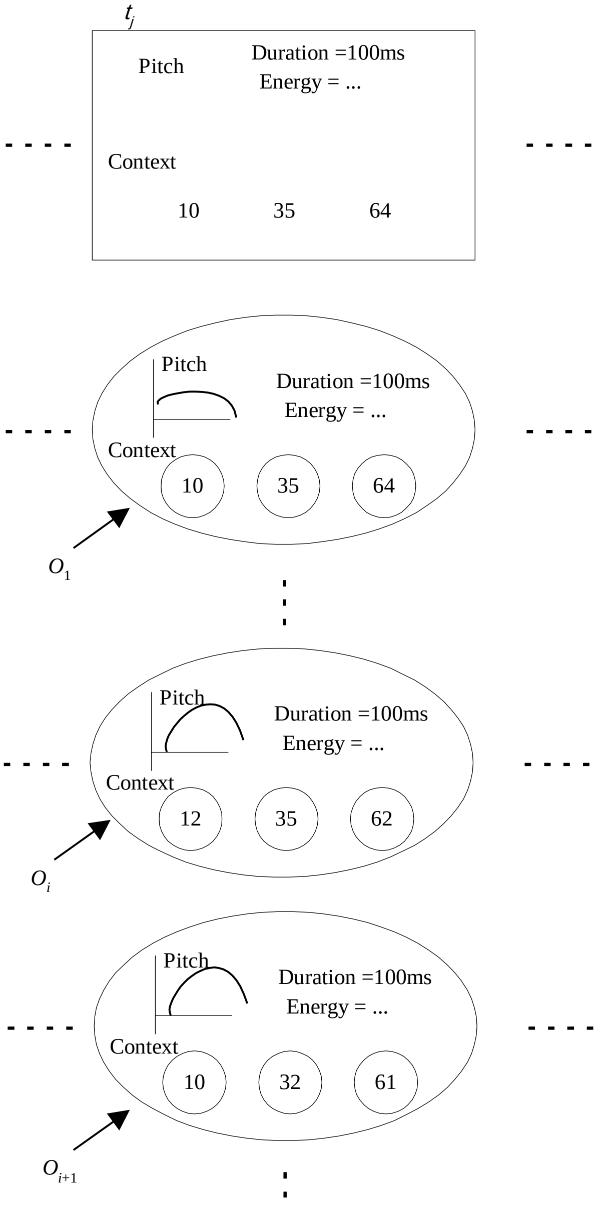

Fig. 2. A close look to on target in the target sequence and the corresponding occurrences.

The proposed method uses quantitative distance measure between context leafs and eliminates all the mentioned problems. The quantitative distance is incorporated in the target cost function in equation 2 and is assigned a weight exactly the same as other target cost elements. In the next subsection, the calculation of this distance will be explained with the overall context-tree growing algorithm.

3.2 Context Clustering trees

Modeling context is a very well known problem in the field of speech recognition and speech synthesis. Top-down likelihood-based decision trees are one of the famous techniques that are commonly used. Detailed description of such technique and using it in the field of speech recognition can be found in [6]. The main idea behind this algorithm that it uses linguistic information about neighboring context and clusters data for similar contexts together. The assessment of the resulting clusters is determined in a likelihood-based criterion. In our case, a tree is grown for each state of the phone model. The algorithm starts with all the data that is aligned to the HMM state in one cluster and split the data according to a yes-no phonetic question about the neighboring context phones. The splitting process results in an increase in the likelihood. This increase in the likelihood can be calculated as in [4]

$$\begin{aligned} \Delta L &= L_y + L_n - L_p \\ \Delta L &= \frac{N_y}{2}\ln\left|\Sigma_y\right| + \frac{N_n}{2}\ln\left|\Sigma_n\right| - \frac{N_p}{2}\ln\left|\Sigma_p\right| \end{aligned} \qquad (3)$$

where L_y, L_n and L_p are the likelihood of the yes-answer, no-answer, and parent clusters respectively, and Σ_y, Σ_n, and Σ_p are the covariance matrices of the Gaussian distributions that represent yes-answer, no-answer and parent clusters respectively. N_x denotes the number of data vectors in each cluster. The question that results in the highest increase in the likelihood is chosen and assigned to the parent cluster and two clusters containing the data are created. This splitting operation is performed until the likelihood increase is below a pre-determined threshold. At the end of the tree-growing algorithm, context-clustering tree is created. The parent nodes contain phonetic questions and the leaf nodes contain model parameters and represent specific context. Each leaf is given a number and this number is the identification number mentioned in the target cost described above.

It is clear from the discussion that merging any two leaf-nodes will result in a decrease in the likelihood. We have chosen this decrease to be the distance measure between different leafs. The likelihood decrease can be calculated the same way like the increase mentioned in equation 2. Although we chose this distance because it is simply calculated during the tree-growing algorithm, one could choose any other distance measure. For instance, Kullback- Lablier distance between the Gaussian probability density functions representing leafs could be used and gives similar results [4].

3.3 Usage with other systems

Although this distance measure is very suitable for synthesis system that uses context-clustering trees during the construction of the synthesis unit inventory, it can be used with other systems that use other construction techniques. For instance, it can be used with synthesis systems that uses hand-constructed diphone inventory as follows. For each diphone, all occurrences that belong to that diphone are recalled and fed to the context-clustering tree-growing algorithm. This will result in similar distance measure that is described above. It should be mentioned that for long units such as phones and diphones, the unit should be divided into smaller units either equally or by using an HMM alignment. This will results in better modeling accuracy. The developed tree-growing algorithm requires only an alignment data for the given unit whether or not it results from HMM alignment.

3.4 Target cost weight Adjustment

In our synthesis system target cost weights are adjusted automatically using linear regression [3,4]. Incorporating the context cost as a target cost element makes it very critical to use the trained weights. Some context mismatches appears when using such weights. In order to eliminate this problem, adjust-listen operations have been performed starting from the automatically trained weights until satisfactory results are obtained. The need of such hand adjustment may be because the objective function used in weight training is not perceptually suitable. The existence of such suitable function could eliminate the hand adjustment

4. DISCUSSION AND RESULTS

As mentioned in the introductory section, this method has been used with an Arabic synthesis system that uses phone as synthesis unit [4,5]. A subjective listening test is performed using a set of 28 person. Each person is asked to listen to a set of 10 sentences that are generated from the system and to give a score to each sentence. This test is similar to the Mean Opinion Score (MOS) that is popularly used in speech coding. Pitch contour, energy contour, and segmental duration of the synthesized sentences are extracted from naturally spoken sentences that are - for sure - not a part of the database. Figure 3 shows the results versus the listeners. It is clear form the figure that the scores are very near to the final score. Comparing these sentences with the naturally spoken sentences, some listener reported that they can not tell the difference between the natural and synthesized sentences.

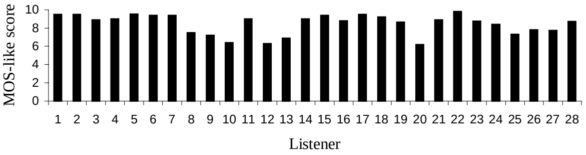

Fig. 3. Mean opinion score test results

The proposed method makes it easy to measure context mismatch for longer synthesis units such as syllables and words. This can be done in the same unit selection framework with minor implementation changes. The authors would like to highlight that although the proposed method reduces the amount of speech modifications required, larger speech database has to be used in order to eliminate the modification operation completely.

5. CONCLUSION

We have developed an objective context-distance measure to be used in online unit selection in concatenative speech synthesis. The context cost is introduced as a part of the cost function used in unit selection process. The major advantage of such distance is that it reduces significantly the modification required in prosodic parameters. These modifications are necessary when using hard decision about context and may result in severe degradation of the speech quality. Results on an Arabic synthesis system validate the importance of the introduced technique.

6. REFERENCES

[1] Beutnagel M., et al " The AT&T Next-Gen TTS system", 137th Acoustical Society of Amirica meeting, Berlin 1999.

[2] Donovan R. E., "Trainable Speech Synthesis", Ph.D. Dissertation, Cambridge University, 1996.

[3] Hunt Andrew J., Black Alan W., "Unit Selection in a Concatenative Speech Synthesis System Using Large Speech Database", International Conference of Acoustics, Speech, and Signal Processing, ICASSP 1996.

[4] Hamza, W., "Arabic Speech Synthesis Using Large Speech Database", Ph.D. Thesis, Cairo University, 2000.

[5] Hamza, W. and Rashwan, M., "Arabic Speech Synthesis Using Large Speech Database", International Conference of Spoken Language, ICSLP 2000.

[6] Bahl L., deSouza P., Gopalakrishnan P., Nahamoo D., Picheny M. "Decision trees for phonological rules in continuous speech", International Conference on Acoustics, Speech, and Signal Processing, 1991.