Six Approaches to Limited Domain Concatenative Speech Synthesis

Robert J. Utama[1], *Lawrence R. Rabiner*[4]

Center of Advanced Information Processing (CAIP)
96 Frelinghuysen Road, Piscataway, NJ 08854
{ru14,lrr}@caip.rutgers.edu

Ann K. Syrdal[2], *Alistair Conkie*[3]

AT&T Labs – Research
180 Park Avenue, Florham Park, NJ 07932
{syrdal, adc}@research.att.com

Abstract

This paper (based on an MS Thesis by Robert Utama in the Electrical and Computer Engineering department at Rutgers University) describes 6 limited-domain Text-to-Speech (TTS) systems that are constrained to the digit string and natural number domains (cardinal numbers only). Each of the 6 unit selection-based concatenative TTS systems were implemented in MATLAB. We evaluate and discuss various factors that influenced the naturalness or overall quality of the synthesized voice. Some of the factors studied were the length and type of the synthesis unit and the extent of co-articulation represented in the recorded speech database. We show that it is possible to create a high quality limited domain TTS system either with maximal or with carefully controlled minimal effects of co-articulation.

Index Terms: speech synthesis, TTS, limited domain TTS.

1. Introduction

In recent years, the unit selection method of speech synthesis, first proposed by Hunt and Black [1], has become the method of choice to perform high quality synthesis. One of the first successful commercial speech synthesizers using this method is described in [2]. Unit selection itself is a concatenative synthesis approach. As such, it is highly dependent on the quality of its training speech database among other factors [3]. In this study we consider different factors that can affect the quality of the synthesized speech, including the extent of co-articulation represented in the database and the type of synthesis unit used in concatenation.

2. Limited Domain Synthesis Systems

2.1. Recording Scripts

Three different recording scripts were created. The first two scripts were used to synthesize digit strings. The third script allowed the pronunciation of sequences as natural numbers, i.e., 10 would be pronounced as "*ten*" instead of "*one zero*" and 111 would be pronounced as "*one hundred and eleven*" instead of "*one one one*."

The first recording script was designed in such a way that all target digits were carefully placed in adjacent phonetic contexts that produced minimal coarticulatory effects on the target digits. This method of limited domain synthesis was shown to be effective in various applications, but had not been formally evaluated previously.

A sentence in the first script was specified (and spoken) in the format:

$$NxN - NxxN$$

where N was a target digit speech unit that was collected into the database and x was a digit or word that provided the neutral/minimal co-articulation context and consequently was not stored in the synthesis database. The first script consisted of 10 7-digit-strings, each divided into two groupings, the first with three digits and the second, with four. The positions of the four target digits (N) within a phrase represented each of four different prosodic contexts commonly used by speakers to designate phrasal groupings of digit strings, such as 10-digit telephone numbers with three prosodic phrase-defined groups indicating 3-digit area code, 3-digit exchange, and final 4-digit line number.

The second digit recording script was designed with the opposite objective of that of the first script. Instead of avoiding co-articulation effects on selected target digits, the second script tried to capture all the co-articulation effects that could possibly occur for each individual digit in each of several prosodic contexts. Since we were building a limited domain TTS system, essentially confined to digit sequences, we could easily meet this condition by making sure that each number was followed by all possible numbers. For example: "*1*" was followed by each of the ten numbers from "*0*" to "*9*." With this method we could capture all possible co-articulation conditions in each prosodic context with a script that contained 100 7-digit phone numbers. We also randomized the script to avoid repetition of numbers (e.g., 010-0101), since this kind of repetition can create undesirable effects such as "tongue-twisters" or unnatural rhythms or prosodic patterns.

The main purpose of the third recording script was to extend the vocabulary of our TTS system from digits to natural numbers. As such the script was not designed to be as inclusive as the second script in terms of the co-articulation transitions between one word to another. We came up with a shortened version of the script in order to record only the necessary combination of co-articulation and prosodic effects. The third script covered the use of decimals, but intentionally left out fractions (e.g., "*one half*").

With the first author serving as the speaker, we recorded speech using each of the three scripts and extracted speech units of various lengths. Word length speech units were extracted from the recordings of the first script, while word, diphone, and phone length speech units were extracted from the recordings of the second and third scripts. These speech units were then used as the acoustic inventory in the unit selection synthesis system that we describe in the next section.

2.2. General Unit Selection Concatenative Synthesis

We used the unit selection method of concatenative speech synthesis. Unit selection provides a very effective method to select the most appropriate pre-recorded segments of speech for a given synthetic utterance. The three factors used to guide the selection process were:

- **Concatenation Cost**
  Concatenation cost is a measure of acoustic mismatch between a pair of speech units when we try to join them

together. We used the acoustic parameters F_0, cepstra, and energy to calculate concatenation mismatch. Speech units that appeared consecutively in the recording script were assigned a concatenation cost of zero. Speech unit concatenation that comes from consecutive units in the database should provide us with the most natural joins and therefore should be utilized whenever possible. The concatenation cost is represented as a weighted sum of the difference between several sub costs as given in Equation (1)

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

where *p* refers to the number of parameters used for the concatenation cost analysis, w_j^c is the weight associated with each parameter, and C_j^c is the cost of the acoustic mismatch at the join of two speech units.

- **Target Cost**
  The target in this context is an approximation of how a normal person would pronounce the utterance that we are trying to synthesize. We used up to 7 parameters to calculate the target cost; including duration, average F_0 over the length of the unit, average energy, previous unit, consecutive unit, unit position and lexical prominence flag for vowel units. The target cost is represented as a weighted sum of the differences between the target and candidate units [2, 4] as given in Equation (2)

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

  where *q* is the number of parameters used for the target cost analysis, w_j^t is the weight associated with each parameter, and C_j^t is the cost of the parameter difference between the target unit and a speech unit in the synthesis inventory.

- **Weight Training**
  The last issue is the problem of determining the optimal weight for the target costs (w_j^t, from Equation (2)). The weights for the target sub-cost calculation were determined using the linear regressive training method described in [1]. The objective of the training was to find a set of weights that minimized the distance between the natural utterance and the synthesized speech signals.

An example of speech synthesis using the unit selection technique is given in Figure 1. Each edge in the graph denotes a cost to concatenate two speech units and each node in the graph denotes a target cost. The output of the unit selection process is the path which gives the lowest total cost and, hopefully, the most natural sounding utterance. In Figure 1, the path that generates the least total cost is denoted by dashed arrows. This path can easily be found using the Viterbi algorithm.

A simple cross-correlation based algorithm was used to mitigate the effect of phase mismatch that occurs when we join two speech units together [5].

3. Perceptual Test

A perceptual test was made up of two separate parts, a digit synthesis section and a natural number section. The digit synthesis test set consisted of 10 unique 10-digit strings with 3-3-4 digit groupings like telephone numbers. Each utterance in the digit synthesis section was 10 digits long with each digit in the sequence picked randomly. The natural number test consisted of 15 unique utterances in the range of 100 to 999 which were picked randomly.

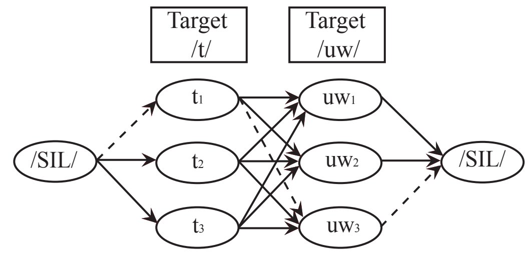

Figure 1: Unit Selection during the synthesis of the word two (/t/uw/), the dash edges represent the path of minimal cost

For the digit synthesis test, the output of six different systems for synthesized speech and one control system (the natural speech recording) were presented to each listener. The six different synthesis methods were then compared.

3.1. Synthesizers Compared

- *No Co-art*: synthesis using word-length speech units from the first script. The only criterion used for unit selection was the unit's position in the utterance.
- *Forward*: synthesis using word-length speech units from the second script. The synthesis criteria used were unit position and the identity of the preceding unit (i.e., appropriate co-articulation with the preceding word).
- *Backward*: synthesis using word-length speech units from the second script. The synthesis criteria used were unit position and the identity of the following unit (i.e., appropriate co-articulation with the subsequent word).
- *F&B*: synthesis using word-length speech units from the second script. The synthesis criteria used were unit position and the identity of both the preceding and following units (i.e., appropriate co-articulation with both preceding and subsequent words).
- *Diphone*: synthesis using diphone-length speech units from the second script. The synthesis criteria used were: unit position, the three concatenation costs, and identity and position of the preceding and subsequent unit.
- *Phone*: synthesis using phone-length speech units from the second script. All the concatenation and target cost criteria were used in this particular system.

For natural numbers, we compared only the phone-length unit selection synthesis system (using both second and third scripts) and natural speech recordings.

The perceptual test was administered using a website, and each test subject accessed the perceptual test using their own computer and listening equipment. The 30 adult volunteer listeners were composed of 16 native and 24 non-native English speakers. Listeners controlled the presentation of each test utterance with the click of a mouse, and they could listen to a stimulus as many times as they wished. In order to familiarize listeners to the task and range of stimuli represented in each section of the test, listeners first rated a short practice set that was not scored. Each test subject rated each utterance on a 5 point scale, with 5 being the best quality (essentially natural) and 1 being the worst quality (very unnatural). The order of test stimuli within each of the two parts of the test was randomized between listeners. When listeners finished the test, their responses were automatically logged.

3.2. Results

3.2.1. Digit Synthesis Results

A repeated measures analysis of variance (ANOVA) was performed on the ratings data of the digit synthesis test. The ANOVA design for the digit synthesis results included the following within-subject factors: Sentences(10) + Systems(7) + Sentences x System (70). There were significant main effects for both Systems (F(6,234) = 245.824, p<0.0001) and Sentences (F(9,351) = 4.632, p<0.0001) factors. In addition, the interaction of System x Sentence was also significant (F(54,2106) = 5.382, p<0.0001) .

As illustrated below, pairwise comparisons (p<0.05) among the seven systems tested indicated the following:

- Recorded speech (*Record*) was rated significantly higher (mean = 4.778) than all synthesized speech systems.
- *F&B* (mean = 3.488) and *No Co-art* (mean = 3.403) systems ratings were statistically equivalent to each other but the F&B system ratings were significantly higher than those of the other four systems
- *No Co-art*, *Backward*, *Phone* and *Forward* ratings did not differ significantly from each other.
- *Diphone* systems (mean = 1.440) had significantly lower ratings than all other systems tested.

Record F&B No Co-art Forward Backward Phone Diphone

Systems underlined by a common line did not differ from each other; systems not underlined by a common line did differ.

The main effect for Sentence simply indicated that some sentences were more difficult than others, and the System x Sentence interaction indicated that the some sentences were more problematic for some systems than for others.

Mean ratings for the seven digit synthesis systems tested are shown in the form of box plots in Figure 2. The whiskers in Figure 2 represents the 95% confidence intervals.

3.2.2. Natural Number Test

A second repeated measures ANOVA was conducted for the natural numbers test. The test design for within-subject factors was: Systems (2) + Sentences (15) + Systems x Sentences (30). There were significant main effects for Systems (F(1,39) = 279.582, p<0.0001) and Sentences (F(14,546) = 2.454, p<0.002), but no significant interaction between Systems and Sentences.

The System main effect reflects the unsurprising fact that recorded speech (mean = 4.765) was rated significantly higher than the *Phone Usel* (phone-length unit selection) system (mean = 3.562). The mean rating and 95% confidence intervals of the natural number test can be seen in Figure 3.

The mean rating of the phone unit-selection system for natural number synthesis was 3.562, which lies above the 95% confidence interval's upper bound (3.446) of the phone digit unit-selection system. Therefore the system performed slightly better in synthesizing natural numbers than digit strings.

3.2.3. Effects of Listener Native Language

An additional ANOVA analysis was conducted that included the between-subjects factor of native English versus non-native

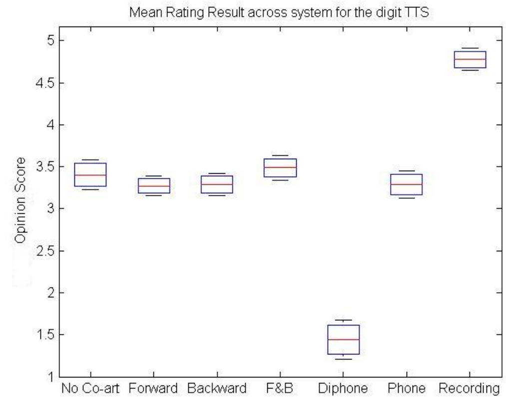

Figure 2: Mean Ratings and 95% Confidence Intervals for the digit synthesis systems

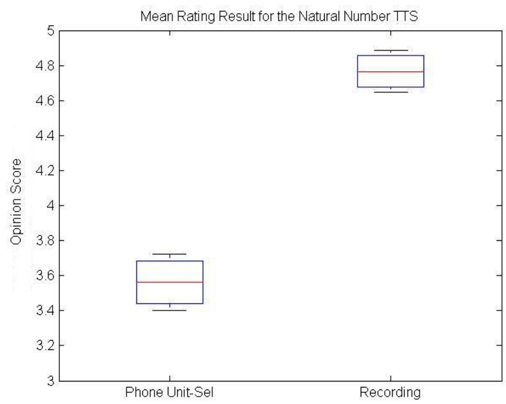

Figure 3: Natural number mean ratings and 95% confidence intervals

English language status. For the test of digit string synthesis, there was a significant effect of Language Status (native vs. non-native speaker) (F(1,38)=6.32, p<0.016) and significant interactions between Language Status and Systems (F(6,228)=2.272, p<0.038) and between Language Status and Sentence (F(9,342)=2.476,p<0.010). There was also a significant 3-way System x Sentences x Language interaction (F(54,2052)=2.489, p<0.0001).

In Figure 4, the mean opinion scores for the native and non-native listener groups are plotted. The mean ratings given by native listeners are usually significantly higher than ratings by the non-native listeners. The only exception to this pattern is that mean ratings for the *Phone* and *Record* systems were almost identical for both native and non-native speaking test subjects.

In the case of the natural number test, there was no significant difference between scores for the native and non-native English speaking listeners.

3.2.4. Effects of Listening Apparatus

The last ANOVA analysis that was conducted for this study was to determine the effect of listener equipment. Out of 40 test subjects that we used, 28 of them used headphones for the listening test whereas the other 12 used loudspeakers. ANOVA test results revealed that there were no significant main effect of

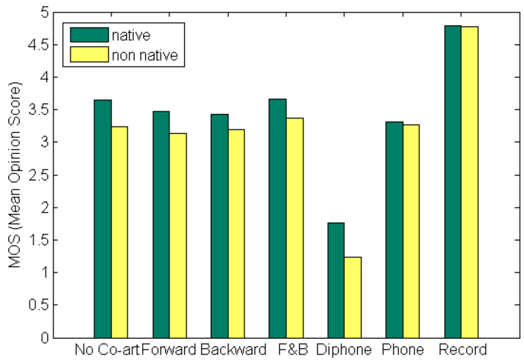

Figure 4: Mean ratings of digit synthesis by native and non-native listeners

the equipment used for listening. Only the System x Equipment interaction was significant ($F(6,228) = 3.763$, $p<0.001$). The mean ratings that describe the effect of the listening apparatus can be seen in Figure (5). For most systems, with the exception of *phone* and *record*, loudspeaker users gave higher MOS ratings. We believe that the headphone users gave lower ratings because they were better able to discriminate problems in the synthesis.

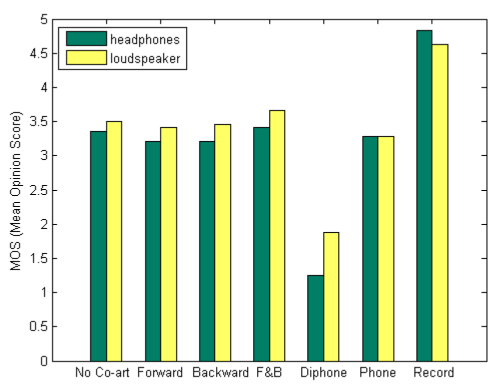

Figure 5: Mean ratings by headphone and speaker for digit synthesis

4. Summary and Conclusions

We compared the subjectively rated quality of six different limited domain speech concatenation techniques. The six different systems used different methods to handle co-articulation effects as well as the effects of the type of speech units used for concatenation. From the results of the MOS quality test we conclude the following:

- Of the six synthesis systems compared, the two systems that had the highest MOS ratings were the word length synthesis system that strictly minimized co-articulation in target units and the word length synthesis system that used co-articulation constraints from both preceding and following contexts. Judging from this result we believe that listeners are sensitive to errors introduced by including inappropriate co-articulation effects in an utterance. However, the inclusion of co-articulation effects was not essential for high quality synthesis. These results seem to suggest an "all or nothing" effect of co-articulation on synthesis quality.
- Although being able to operate on sub-word length speech units makes the TTS system more flexible, synthesis quality may significantly decline unless the synthesis is done properly. There were more concatenation points in the diphone- and phone-based systems than in the word-based systems, yet only the diphone system performed relatively poorly. The diphone-based system did not employ any form of prosody prediction, which probably accounts for its poor quality. The synthesis quality might be improved if we provided the TTS system with more descriptive prosody information.

  A prosody look-up table was implemented for the phone length system. The look-up table stored the average prosody information, such as F_0, energy and unit duration, for a given speech unit at a given location. Even though the phone based unit selection TTS system employed a simple prosody prediction algorithm, it had a much higher MOS rating of 3.285, representing a great deal of improvement when compared to the diphone based TTS system.
- The MOS rating of phone-based natural number synthesis was higher than its digit synthesis rating even though the same synthesis method was used for both. This might be explained by the fact that the weights were trained only for natural numbers. Initially it was thought that training the system only for natural numbers would be sufficient, since the digit vocabulary is a subset of the natural number vocabulary. However, in practice this turned out not to be the case.
- The use of headphones for a listening test is desirable. The MOS ratings suggested that headphones enable a user to better discriminate problems in the synthetic voice. Hence, headphone users gave lower ratings than users of speakers.

5. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *ICASSP*, vol. 1, pp. 373–376, 1996.

[2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," tech. rep., Joint Meeting of ASA, EAA, and DAGA, Berlin,Germany, 1997.

[3] A. W. Black, "Perfect synthesis for all of the people all of the time," in *IEEE TTS Workshop 2002*, (Santa Monica), IEEE, 2002.

[4] A. Black and P. Taylor, "CHATR: a generic speech synthesis systems," tech. rep., COLING 1994, 15th International Conference on Computational Linguistics, Kyoto, Japan, 1994.

[5] T. Dutoit and M. Cernak, "TTSBOX: A MATLAB toolbox for teaching text-to-speech synthesis," (Philadelphia), ICASSP'05.