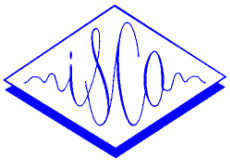

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

Diphone Synthesis using Unit Selection

Mark Beutnagel *Alistair Conkie* *Ann K. Syrdal*

AT&T Labs - Research, Florham Park, NJ, USA

ABSTRACT

This paper describes an experimental AT&T concatenative synthesis system using unit selection, for which the basic synthesis units are diphones. The synthesizer may use any of the data from a large database of utterances. Since there are in general multiple instances of each concatenative unit, the system performs dynamic unit selection. Selection among candidates is done dynamically at synthesis, in a manner that is based on and extends unit selection implemented in the CHATR synthesis system [1][4]. Selected units may be either phones or diphones, and they can be synthesized by a variety of methods, including PSOLA [5], HNM [11], and simple unit concatenation. The AT&T system, with CHATR unit selection, was implemented within the framework of the Festival Speech Synthesis System [2]. The voice database amounted to approximately one and one-half hours of speech and was constructed from read text taken from three sources. The first source was a portion of the 1989 Wall Street Journal material from the Penn Treebank Project, so that the most frequent diphones were well represented. Complete diphone converage was assured by the second text, which was designed for diphone databases [12]. A third set of data consisted of recorded prompts for telephone service applications. Subjective formal listening tests were conducted to compare speech quality for several options that exist in the AT&T synthesizer, including synthesis methods and choices of fundamental units. These tests showed that unit selection techniques can be successfully applied to diphone synthesis.

1. Introduction

Concatenative speech synthesis systems have grown in popularity in recent years. As memory costs have dropped, it has become possible to increase the size of the acoustic inventory that can be used in such a system. The first successful concatenative systems were diphone based [7], with only one diphone unit representing each combination of consecutive phones. An important issue for these systems was how to select, offline, the single best unit of each diphone for inclusion in the acoustic inventory [12] [6]. More recently there has been interest in automation of the process of creating databases and in allowing multiple instances of particular phones or groups of phones in the database, with selection decided at run time. A new, but related problem has emerged: that of dynamically choosing the most adequate unit for any particular synthesized utterance. The most well known such system, CHATR [4], uses phones as the basic concatenative units. In this paper we describe a modified version of the CHATR system where the basic synthesis units are diphones rather than phones. We examine the consequences for unit selection of such a choice, and present results of subjective listening tests comparing size of units, several inventory pre-processing techniques, and synthesis methods.

2. Diphone Selection

Using diphones instead of phones has several consequences. For American English there are something like 50 phones (composed of phonemes and distinctive allophones) compared to perhaps 2000 diphones. For a given database there are many more diphones than phones, and many fewer instances of each diphone than of each phone. Statistical techniques that are valid for dealing with a phone-based system may be no longer valid for a diphone-based system. We describe in this paper how it is possible to approximate the calculations necessary to estimate costs so as to deal with the sparsity problem. Another aspect of using diphones is that there has been extensive work in this area in recent years and the benefits and limitations of diphones are well understood.

2.1. The CHATR Unit selection mechanism

The CHATR unit selection mechanism is based on having a fundamental unit that is a phone. In a typical database (of perhaps 30-60 minutes of speech) there are many instances of the most common phones in various contexts, and even the least common phones are generally adequately represented. For a given synthesis specification there are generally many instances of the requested phones that can be used to make up the sequence that will be heard. A selection is carried out using a Viterbi search. A set of costs is assigned. A "target cost" is associated with how well each candidate unit

matches an ideal target unit, with a low cost indicating that the unit is appropriate. A "concatenation cost" is assigned to a pair of phonemes to estimate how well perceptually they can join together. A low cost represents a smooth join. Once costs are assigned to the phonemes and pairs of phonemes it is possible to build a network and find the lowest cost of traversing the network. The sequence of phonemes in the lowest cost path are then used for synthesis.

To produce synthesis reasonably quickly a preselection phase is carried out, and units which seem to be wildly different from the profile required (e.g. in terms of context) are not considered for the search phase.

Partly as a consequence of the fact that a typical CHATR database is prepared automatically and so has only approximately accurate phoneme boundaries marked, join points in CHATR are not normally phoneme boundaries. Instead a search mechanism is employed to find "optimal" join points close to the notional phoneme boundaries. It is not clear how necessary this procedure is if the boundaries are accurately marked, nor is it evident that the procedure always improves quality.

2.2. The Diphone Unit Selection Mechanism

Diphone unit selection parallels much of the above description of CHATR. There is a target cost associated with each diphone rather than to each phone. This target cost is essentially a sum of the costs that would be assigned to the pair of phones individually. In this way we avoid the sparsity problem mentioned previously. Concatenation cost is calculated in the same way as before. The Viterbi search process is also as before. Currently we find that it is not necessary to include a search for optimal coupling points, probably because we are making the joins at stable points much more often than in the case of phones. The parallels and lack of changes necessary are a tribute to the generality of the original CHATR unit selection mechanism.

2.3. Training

It is in training that diphone selection differs somewhat from unit selection based on phones. If we count categories based on diphones, then there are too many, and they are too sparsely populated, for the statistics that are employed by CHATR. We resolve this problem by first calculating weights based on a more phone-oriented scheme and then calculating combinations.

3. Voice Database

The following sections describe the creation of a female voice.

3.1. Female Voice

A single female speaker was selected based on results of a formal listening test [13]. The two and one-half hour listening test had 41 subjects rate the natural and synthesized voices of six female professional speakers over a variety of conditions. A total of 38,376 ratings were collected. Listeners were asked to rate on independent 5 point scales the intelligibility, naturalness and pleasantness of test sentences spoken by or synthesized from each of the speakers. Speaker selection was based on the outcome of this test.

Recordings were made of the selected speaker reading a variety of texts in a specially designed acoustically "dead" room. This was done using a high quality head-mounted microphone to a DAT tape recorder set to record at a 48 kHz sampling rate. The resulting recordings were transferred directly to computer files, downsampled to 16 kHz and verified for quality.

The texts used in the experiment described here were of three types. The first type was a portion of the Wall Street Journal (WSJ) Corpus. Approximately one-half hour of WSJ recordings were used in the present experiment. Paragraphs of WSJ articles were read in their entirety. We sought with the WSJ corpus to achieve a large sample of phones, syllables, and accents, whose distributions are representative of read American business news, and to capture discourse-level prosodic variations.

The second type of text was a set of individual sentences that was specifically designed to include all diphones in American English. This set of sentences ensured that we had coverage of phoneme pairs that are rare, something we would not be able to guarantee with just the WSJ corpus, given that some phone sequences are possible but extremely rare. These recordings constituted about one-half hour of speech.

A third type of recordings consisted of prompts for telephone network services. The prompts consisted of verbal interactions not well represented in the other corpora, including greetings and other expressions, questions, assertions, and directions. Prompts varied in length from single sentences to paragraphs. The prompt recordings totaled about one-half hour of speech.

The recordings were edited into sentence or paragraph length files and then labeled with information pertinent to the unit selection synthesis process. Phonetic labels were added, based on the so-called ARPABET phoneset. This was done in two stages. First the Entropic Aligner program was used to automatically segment and label the signals. A second manual pass was done to correct and verify the segmentation and labels. Aligner-generated word labels were automatically corrected to match the corrected phone labels, and syllable labels were also added. A set of prosodic labels [10], tones and breaks, were added to the database to complete the annotation thereof. These labels were again calculated

automatically, and manually corrected. The final form of the database consisted of the speech files together with several sets of labels (with timing information), the most pertinent being phones, syllables, words, tones and breaks. Other information about the signals (f0, energy, etc.) was generated from the signals when needed.

4. System Description

We describe the major components of the system.

4.1. The Festival Framework

Festival, together with Edinburgh Speech Tools, is a publicly available speech synthesis system distributed by the University of Edinburgh. It is based on a very general architecture implemented in C++. A Scheme [9] interpreter is embedded in Festival and provides high-level flow control. Use of an interpreted language makes experimentation faster and allows flexible scripting.

Among the tools in the Speech Tools Library are classification and regression trees (CART). CARTs are created by the *wagon* utility. CART components are widely used within Festival, e.g. for phoneme durations, post-lexical processing, and accent prediction.

The experimental AT&T system we describe in this paper used the standard version of Festival with some modifications. The standard version of Festival was used for text normalization, phonetization (letter-to-sound mapping), and prosody generation. For the experiments described here we confined changes to the synthesis module, post-lexical processing and the voice. The modifications to post-lexical processing were confined to training our own version from our own data, either in a form that is phoneme specific or syllable specific.

4.2. Dynamic Unit Selection

The CHATR unit selection mechanism was used as the synthesis method for this experiment. The mechanism is general, and it was possible to use the same module configured in different ways to provide both diphone and phoneme bases selection.

New (non-standard) label information, or features, were added to the voice database, making them available for consideration in the target and concatenation cost computations. These features include lexical stress, syllable boundaries, word boundaries, and whether a segment occurs in a content word. These features, which are accurately marked in the label files, allow the possibility of selecting units based on these perceptually relevant categories (as well as on acoustic parameters such as pitch and duration).

Additionally, in support of these new features, associated cost functions were added. The database training process was then repeated in order to find appropriate weights for the newly added features.

4.3. Synthesis Methods

CHATR initially supported a simple synthesis method, that concatenates the waveforms of speech units directly. Because of the rich database and the use of prosodic targets in unit selection, the results are often acceptable. For the purpose of this paper, we denote this method as WAV.

Two synthesis options were added in the experimental AT&T system. The new modules implement Harmonic plus Noise Model (HNM) [11] and PSOLA synthesis, and may be selected at run time. Both optionally support prosody modification. HNM also performs spectral smoothing at concatenation points. It does this by attenuating mismatched formants near the concatenation boundary, effecting a fade-out/fade-in formant transition.

For the tests, five synthesis variants were used: HNM, PSOLA, and WAV, each with no prosody modification, and HNM and PSOLA with prosody modification. The WAV synthesis method, because of its simple nature, is not amenable to prosodic modification.

4.4. Post-Lexical Processing

Post-Lexical Processing (PLP) is needed in order to increase the possibility of finding a good sequence of phonemes in the database to represent the specified sequence for synthesis. The lexical entries used by the system are in general not a good match for the contents of the database. A good set of Post-Lexical rules can increase the match and consequently the synthesis quality. In this experiment we tried two different sets of rules, constructed in different ways. One is a very general set, essentially syllable based, designed to work across several databases. The second, specific to our database, is phoneme-based and attempts to take account of vowel reduction, reduction to flaps and other phenomena. It represents a first attempt to maximize the matching of synthesis specifications with a database.

4.5. Pruning

Since the database consists of many units, some very similar to others and some very different from anything else, there are units which may never get used. Since the database is very large it makes sense to try to eliminate some of the units. This can be done by the process of pruning, which is implemented in the CHATR system. Since pruning may also remove some useful units, database quality may decrease. For the experiments described here, we had one case where there was no pruning and a second with pruning, carried out in two steps, with approximately 10% of data pruned each time. Thus the pruned database was approximately 20% smaller than the unpruned database.

5. Experiment

A formal listening test was conducted to evaluate combinations of variations in synthesis units, database pruning, post-lexical processing, and synthesizers. The following sections describe test materials, test conditions, test procedures, listeners, and the statistical treatment of the results.

5.1. Test Materials

Eight utterances were selected to serve as test stimuli in the formal listening test: four were chosen from a corpus of flight information prompts, and four others from the set of revised Harvard phonetically balanced sentences [3] [8]. Utterances ranged from 5-20 words in length; one flight prompt was composed of two sentences, and the remaining test utterances were individual sentences. Test utterances were selected prior to their being synthesized under any conditions tested in the experiment. All test utterances had a sampling rate of 16 kHz and were energy normalized 16-bit linear files.

5.2. Test Conditions

Each of the 8 test utterances was synthesized under each of the 40 combinations of the following synthesis conditions, resulting in a total of 320 test items.

- Units (2): Phones or Diphones
- Synthesis Methods (5): No Prosodic Modification (HNM, PSOLA, or WAV) or Prosodic Modification (HNM or PSOLA)
- Post-Lexical Processing (2): General PLP or Specific PLP
- Database Pruning (2): No Pruning or Pruning

Prosodic modification or no prosodic modification gives rise to different prosodic realizations, so it is interesting to know by just how much prosody will be modified. This will have a bearing on synthesis quality. The unit selection attempts to match the (same) specified prosody in either case, but will only approximately achieve the specification. If prosodic modification is then applied, durations and F0 values will be modified.

An example of the prosodic variation found between the prosodically modified and unmodified test sentences is illustrated in Figure 1.

5.3. Test Procedure

Test utterances were 40 - 6500 Hz bandpass filtered by a Wavetek Brickwall Filter System 716 and presented to listeners over Sennheiser HD 250 Linear II calibrated headphones.

Subjective ratings of each test utterance were collected from each listener, resulting in a total of 14,080 observations over the course of the experiment. For each test utterance, listeners were presented a 5-point Mean Opinion Score (MOS) rating scale from which to select their judgments using a touch sensitive screen, with 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad. The order of presentation of test utterances was independently randomized for each group tested.

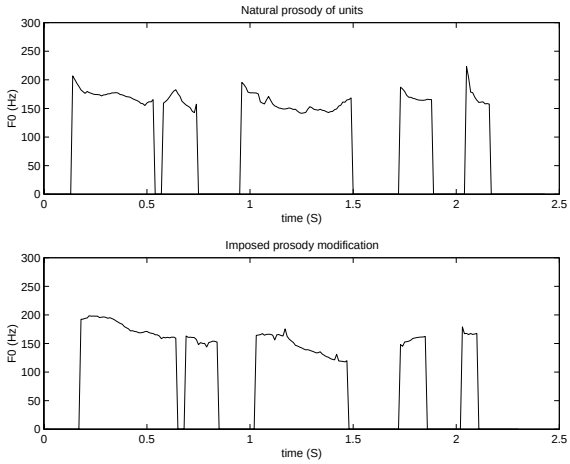

Figure 1: Example of F0 variations for one of the test sentences

A brief familiarization session preceded testing, during which listeners were presented samples representing a wide range of the variation among the test utterances presented during the experiment, and they were given practice in using the rating scale and touch screens. The listening test lasted approximately one hour, including the initial instructions and practice session.

5.4. Listeners

A total of 44 listeners (41 females and 3 males) were tested in four groups of 11 listeners each. Listeners were between 18 and 60 years of age, with a median age of 45.5 years and a mean age of 44.9 years. All were native speakers of American English with no known hearing loss. Listeners were experienced with previous voice quality listening tests, but they were unfamiliar with text-to-speech synthesis.

5.5. Statistical Analysis

A repeated measures Analysis of Variance (ANOVA) with the following five factors was performed: Unit (2), Synthesizer (5), PLP (2), Pruning (2), and Sentence (8). Statistical significance of main effects and interactions was determined from the resulting F statistics, with $p < 0.05$.

6. Results

There was a significant main effect of Unit ($F=53.082$; $df=1,43$; $p<0.0001$), reflecting the fact that diphone units (with a mean rating of 3.20) were rated higher than CHATR-style phone units (whose mean rating was 3.10).

There was a main effect of Synthesizer ($F=45.438$; $df=4,172$; $p<0.0001$), indicating that there were significant differences in ratings among the five synthesis methods represented. The three synthesis methods that used no prosodic modification were each significantly more highly rated than either of the two synthesis methods that used prosodic modification. In the case of no prosodic modification, HNM's mean rating of 3.29 was slightly higher than PSOLA's mean rating of 3.25, and both were higher than the mean rating of 3.16 for WAV. For the two synthesizers that used prosodic modification, PSOLA had higher mean ratings (3.09) than HNM (2.96).

There was a significant main effect of PLP ($F=33.215$; $df=1,43$; $p<0.0001$), due to the more general version of PLP receiving a higher rating (3.21) than the more specific version (3.09).

There was also a main effect for Pruning ($F=37.65$; $df=1,43$; $p<0.0001$), reflecting the fact that the no database pruning condition was rated higher on the average (3.18) than the pruned condition (3.12).

There were numerous significant interactions among factors. This means that the differences observed among interacting test conditions were not attributable to the main effects alone. The interactions are described and discussed below.

There was a significant Unit by Synthesizer interaction ($F=11.838$; $df=4,172$; $p<0.0001$), which reflected differences in synthesizer ratings depending on whether the units were phones or diphones. For diphone units, the ratings for HNM without prosody modification (mean=3.38) was significantly higher than PSOLA without prosody modification (mean=3.28) and WAV (mean=3.21). These three were significantly higher than PSOLA (mean=3.08) and HNM (mean=3.02) with prosody modification. For phone units, the ratings for PSOLA without prosody modification (mean=3.22) was statistically equivalent to HNM without prosody modification (mean=3.20). Both of these were higher than WAV (mean=3.11) and PSOLA (mean=3.09), which were equivalent to each other and higher than HNM with prosodic modification (mean=2.89). This interaction is likewise attributable to differences in ratings for the various synthesizers between diphone and phone units. Mean ratings for HNM with no prosodic modification, for example, declined 0.18 points from diphone to phone units, whereas ratings for PSOLA with no prosodic modification declined only 0.06 points, and those for WAV declined 0.10 points. Ratings for PSOLA with prosodic modification were equivalent whether diphone or phone units were used, but ratings for HNM with prosodic modification declined 0.13 points from diphone to phone units. In general, HNM synthesis methods improved much more by using diphone units than did PSOLA synthesis methods.

There was a significant three-way Synthesizer by PLP by Pruning interaction ($F=6.299$; $df=4,172$; $p<0.0001$), which basically indicated that pruning made less difference in synthesizer ratings with the specific PLP than with the general PLP.

There were a set of related interactions involving the factors Units, PLP, and Pruning. The Unit by PLP interaction ($F=95.736$; $df=1,43$; $p<0.0001$) reflected the fact that the superiority of diphone units over phone units was only observed for the more general version of PLP; the mean rating for diphones was 3.31, while the mean for phones was 3.11. For the more specific version of PLP, there was no difference in ratings between phone (mean=3.10) and diphone (mean = 3.08) units, and their ratings were equivalent to the phone units with the general version of PLP. Similarly, the PLP by Pruning interaction ($F=11.799$; $df=1,43$; $p<0.001$) was due to the superiority of no pruning over pruning only for the general PLP condition; no pruning received a mean rating of 3.26, and pruning, a mean of 3.16. For the more specific PLP condition, the mean for no pruning was 3.10, and the mean for pruning was 3.08. The three-way interaction of Unit by PLP by Pruning was also significant ($F=16.283$; $df=1,43$; $p<0.0001$), basically because the rating for diphones with general PLP and no pruning (3.38) was higher than all others, and the rating for diphones with general PLP and pruning (3.24) was higher than the other remaining conditions.

7. Conclusions

Using an experimental TTS system, we have compared some interesting variations in synthesis units, database size, pruning methods and synthesizers. We evaluated the various combinations of synthesis conditions in terms of synthesis quality via formal perceptual experiments.

The major results of the experiment are as follows:

- Listeners rated synthetic speech that had not been prosodically modified higher than prosodically modified synthetic speech.
- Synthesis with diphone units was more highly rated overall than synthesis with phone units, but diphones were superior only for the more general version of PLP, not for the more specific version.
- The more general version of PLP was rated higher than the more specific version.
- Diphone units were rated relatively much higher than phones for HNM synthesizers, both with and without prosody modification, than they were for PSOLA synthesizers, for which diphone and phone ratings were more equivalent.
- Synthesis from an unpruned database was rated higher overall than that from a pruned database, but the unpruned database was only superior for the general version of PLP, not for the specific version.

Apart from the specific results, the experiment basically demonstrates that the synthesis approach we describe is a viable way of blending diphone and unit selection techniques.

8. REFERENCES

1. A. Black. *CHATR, Version 0.8, a generic speech synthesis.* System documentation. ATR - Interpreting Telecommunications Laboratories, Kyoto, Japan, March 1996.
2. A. Black and P. Taylor. *The Festival Speech Synthesis System: system documentation.* Technical Report HCRC/TR-83. Human Communications Research Centre, University of Edinburgh, Scotland, UK, January 1997.
3. J. P. Egan. Articulation testing methods. *Laryngoscope*, 58:955–991, 1948.
4. A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP*, 1:373–376, 1996.
5. E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 (5/6):453–467, 1990.
6. J. Olive, J. van Santen, B. Moebius, and C. Shih. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pages 191–228. Kluwer Academic Publishers, Norwell, Massachusetts, 1998.
7. J. P. Olive. Rule synthesis of speech from diadic units. *ICASSP*, pages 568–570, 1977.
8. IEEE Subcommittee on Subjective Measurements. IEEE recommended practive for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17:227–246, 1969.
9. Revised^4 report on the algorithmic language scheme., Nov. 1991. Available at http://www.cs.indiana.edu/scheme-repository/doc.standards.html.
10. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling english prosody. *ICSLP*, 2:867–870, 1992.
11. Y. Stylianou, T. Dutoit, and J. Schroeter. Diphones concatenation using a harmonic plus noise model of speech. *Proc. EUROSPEECH*, Sept. 1997.
12. A. Syrdal. Development of a female voice for a concatenative text-to-speech synthesis system. *Current Topics in Acoust. Res.*, 1:169–181, 1994.
13. A. K. Syrdal, A. Conkie, Y. Stylianou, J. Schroeter, L. F. Garrison, and D. L. Dutton. Voice selection for speech synthesis. *J. Acoust. Soc. Am.*, 102(5), November 1997.