

ROBUST UNIT SELECTION SYSTEM FOR SPEECH SYNTHESIS

Alistair Conkie

AT&T Labs – Research, Shannon Labs, 180 Park Ave.,
Florham Park, NJ 07932 U.S.A.

ABSTRACT

There has been much interest for many years in diphone-based concatenative speech synthesis and, recently, a rapidly increasing interest in unit selection based synthesis (as illustrated by the CHATR [2] system). However, the limitations of both types of system are well known. While intelligibility is generally very high for diphone based systems, the resulting signals do not sound completely natural. This happens for several reasons, amongst them the limited number of phone variants present in a typical system, and the potential artifacts introduced by concatenating at diphone boundaries. For unit selection synthesis, typically phone-based, it is possible to produce sentences that sound surprisingly natural and intelligible from a large database. However, quality is often inconsistent, and the main difficulties appear to be selecting acoustically appropriate units with the correct prosodic characteristics. Also, note that typically no prosody modification is done to achieve the highest possible quality.

In an effort to capture the best features of both systems we have devised a unit-selection and synthesis algorithm that allows finer control than the CHATR system (version 0.8), both by applying selective prosody modification and by exercising finer control over the units that get chosen for synthesis.

We present the algorithm and results of experiments based on our own version of unit selection synthesis.

1. INTRODUCTION

Diphone-based speech synthesis has been researched for many years. Some of the most successful diphone systems have been concatenative [6]. Such systems can produce very intelligible synthetic speech, but tend not to sound completely natural. This lack of naturalness can be attributed, at least in part, to the limited set of units from which speech is chosen (typically ~2000 diphones), coupled with the need to prosodically modify the speech signal of each diphone.

More recently, exploiting lower memory costs and faster processors, systems have appeared that use larger databases and allow units to be chosen that have appropriate spectral and prosodic characteristics. The complexity shifts from choosing units offline to choosing them online. Our unit selection work is based on one such system, the CHATR system [2] from ATR. The main contribution of this work to our system was to provide an extremely flexible framework for general unit selection (compared with older systems that used ad hoc methods).

The AT&T NextGen TTS system [10] follows closely the unit selection algorithm developed at ATR. The algorithm has been combined with the Flextalk [8] front end and incorporated into the Festival [3] system. Synthesis can be either simple waveform concatenation, or PSOLA [6] or has been HNM [9].

The principal goal of the project was to improve the robustness of the unit selection process by introducing techniques from diphone synthesis. The CHATR system succeeds in a number of ways. It provides a very convincing proof of concept, and at times produces very natural sounding synthesis. However in order for a synthesis system to be of practical interest, the synthesis quality must be consistently high. That goal was not completely achieved in the original system. We applied considerable effort towards achieving a high quality database with good coverage. We examined the unit selection algorithm itself and identified a number of possible improvements. The most fundamental change was to choose a different set of units. By selecting half phones instead of phones (the default case in CHATR) we add considerable flexibility to how synthesis is performed without paying too great a penalty in terms of computation time. In particular we allow diphones to be used conveniently in a unit selection context. It is, however, the combination of all the various changes (and painstaking attention to detail and evaluation), rather than one particular thing, which allows the system to perform at a superior level of quality.

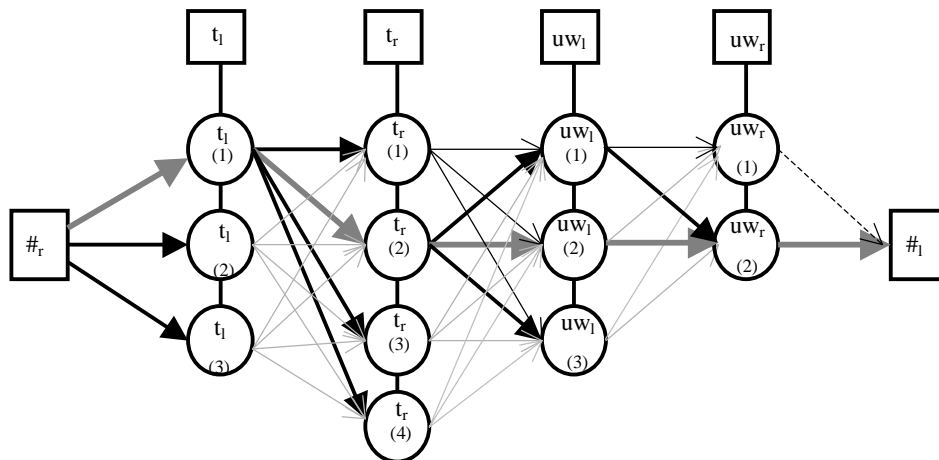


Fig. 1: Viterbi search based on an inventory of multiple instances of each half-phone needed for synthesizing silence -/t-/uw/-silence (the word “two”).

2. CHATR UNIT SELECTION

The goal of unit selection is to provide a mechanism whereby segments of prerecorded speech are selected for synthesis. These segments are provided from a database. They may be parameterized in some form (LPC, HNM) or other or may be digitized speech samples.

Unit selection synthesis, such as used in the CHATR system requires a set of speech units that

can be classified into a limited number of categories such that sufficient examples of each unit are available to make statistical selection techniques possible. The original CHATR system uses phonemes as units. Selection of units from the database is done for each new synthesis request. To avoid problems of concatenation at phoneme boundaries a flexible join technique is used allowing movable boundaries. Prosodic modification is optional rather than required since there are a variety of different prosodic contexts represented in the database. A typical unit selection database is much larger in size than a diphone database.

The CHATR unit selection mechanism can be described as follows (based on [4]). A synthesis specification is provided, giving a list of phonemes, durations, and target points on a desired f_0 curve. The first stage involves, for each required (or target) phoneme in turn, looking at all phonemes with the same name that are in the database. Each of these database phonemes is assigned a Target Cost based on the similarity to the requested unit (lower cost equals more similar). A second type of cost, Concatenation Cost, is calculated based on how well two units join together (lower cost equals smoother join).

A network can be constructed and costs as described above assigned to each unit and to links between units. The task is then to find the lowest cost path through the network. In practice due to computational complexity, not all paths can be searched. In this case we limit the search beam width.

3. MODIFIED UNIT SELECTION

We opted to use half phonemes (or equivalently half diphones) as opposed to phonemes. This allows combinations of units to imitate diphones or phonemes or more complicated structures as appropriate. Additionally there are the same number of instances of each half phoneme as of phonemes so the statistical analysis of units will continue to be possible. It would be much more difficult to collect sufficient statistics using diphones directly.

With half phones as the lowest common denominators we can always build larger units to match any given scheme by manipulating Concatenation Costs. For CHATR, if we have half phones instead of phones there will be instances where we can profit from a diphone-type join as opposed to a phone-phone type join. Hence synthesis quality is (potentially) better. The experimental results seem to bear this out in practice.

The (at least potentially) larger number of joins caused by choosing smaller units is clearly a concern but it appears that if the various costs are related in some plausible way to perception we can find paths through the network which have something approaching minimum overall cost. They are reasonably close to optimal, and better than using a fixed (e.g. diphone-only scheme).

If this is true, then in general, for any concatenative unit scheme (phonemes, diphones, demi-syllables) and database, half-phoneme units can perform at least as well and potentially better than larger alternative units.

3.1. DATABASE

In contrast to the CHATR approach, which placed the emphasis on a new paradigm for synthesis, our approach was motivated by a desire to create a text-to-speech system in a similar style, but with the emphasis on robustness. We feel that this requires making the database quality, both segmentation and labeling, an important element of our strategy to achieve Robust Unit Selection. There were two ways in which this was done. Firstly, the composition of the database was carefully considered so that there were (a) sufficient examples of phoneme-phoneme pairs, or in other words diphone coverage; (b) a wide variety of prosodic styles and contexts including material from the newspaper text and some interactive prompts-style utterances. Secondly, the material recorded, after being automatically labeled as a first pass, was then manually corrected. The result of this is a high quality annotated database representing several speaking styles.

There are currently 100k units (50k phonemes) in the database, corresponding to approximately 80 minutes of speech.

3.2. ALGORITHM

The unit selection algorithm itself does not differ greatly between CHATR and our NextGen system. Concatenation costs are calculated as before. Target costs are calculated per half-phoneme. If anything this allows finer control than CHATR unit selection. The Viterbi search (Figure 1) is more complicated, given that there are now (approximately) twice as many calculations to make. This does not however lead to exorbitant computational costs.

Some minor adjustments are made to the way potential units are pre-selected for the Viterbi search, to reflect the more predictable context that a half-phoneme will have compared with a phoneme. A larger number of adjacent units are examined in the half phone case when considering context. The half phoneme is a suitable basis unit for both the phoneme and the diphone, and the system can easily mimic either unit type. Concatenation Cost can be modified so that it is very costly to join half-phonemes together at a phoneme boundary (other than units that naturally belong together). This effectively leads to diphone synthesis given a sufficiently high penalty, with the additional attraction that synthesis can still take place even where there are missing diphones. A second possibility is to make phoneme-internal joins very expensive, except for naturally co-occurring units. This case is close to the original CHATR case. In practice, global weights chosen such that there are more diphone joins than phoneme joins (approximately in the ratio 2 to 1) seem to give the best results.

Training the weights is a non-trivial process. There are two things that help: first, processing power continues to increase, certainly compared with [4] so that the process is much less onerous than previously. Training a database takes several hours on an SGI R10000. Secondly, although we increase the number of units in the database by a factor of two, the processing time increases by a factor of 2 also, and not by the larger factor that a larger database would give, where [4] says it is quadratic.

Target costs are done in two stages since the process is computationally expensive. A cheaper preselection phase weeds out unlikely candidates, and a second pass calculates more detailed weights.

The target cost weights, which determine the relative importance to be placed on each feature, are calculated using regression. We found that this seemed to work best when the features used to describe each unit are as independent as possible. Calculating appropriate Target Costs is still an area of much ongoing research, e.g. [5].

Calculating the Concatenation Costs is computationally the most expensive part of the algorithm and hence the costs are Vector Quantized for processing efficiency. This part remains the same as for CHATR, except that we make no attempt to do optimal coupling at join points.

4. RESULTS

We carried out formal listening tests to compare the new unit selection scheme in two different configurations. The system was configured to produce in one case only phonemes and in the second case only diphones, by adjusting the costs associated with each type of join. The results are shown in Figure 2. These are the two extreme cases possible. We can also adjust parameters to favor one type of join over the other. Our informal listening tests suggest that optimal performance is achieved when the ratio of diphone joins to phone joins is approximately 2 to 1.

A second formal test examined the case where, for a given database a diphone set was selected and compared with the complete database used in unit selection mode. A significant preference for the signals made using the whole database was found.

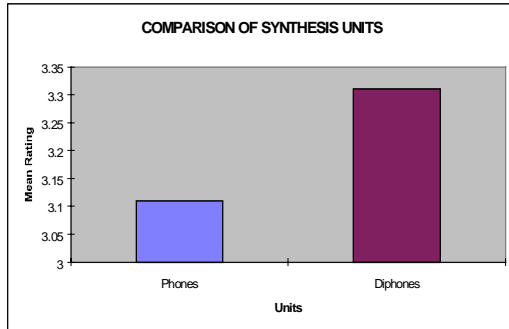


Fig. 2: Comparison of synthesis units.

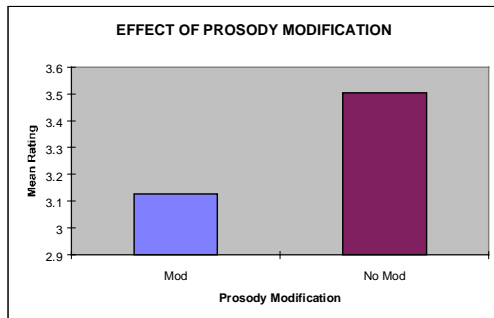


Fig. 3: Effect of prosody modification

In a third test, synthetic sentences were constructed using unit selection under two different conditions (a) without prosodic modifications to the units and (b) with prosodic modifications to the units. Our results indicate that the technique with no prosodic modification was preferred. This preference seems to be true for system-generated prosody and for prosody copied from natural sentences (Figure 3.).

Some limited prosodic modifications do appear to be beneficial. Quality improves when we allow limited smoothing at diphone boundaries. Also limited duration modification and amplitude smoothing give improvements in informal testing.

Database size is an important variable in determining quality. Each time we have increased the number of units in the database we have seen a significant increase in quality.

5. DISCUSSION

The combination of a high quality speaker, high-quality recordings, a high-quality database and robust unit selection has allowed us to improve our synthesis quality significantly. This is particularly true with respect to which we believe is directly due to the large speech database and the ability to mimic prosody without extensive signal processing. We are fortunate, too, to have available, in Flextalk, a high quality front end to our synthesis system.

Unit selection contributes a general mechanism where previously there were only ad hoc methods. For the foreseeable future this mechanism will enable many interesting experiments to be carried out. Computationally the technique is expensive, but not so expensive as to render it impractical for non-research applications. Unit selection does not obviate the need for signal processing modification. It does provide an alternative method of synthesis, which can preserve important aspects of the original recorded speech signal. It can be used effectively in conjunction with signal processing techniques.

6. REFERENCES

- [1] M. Beutnagel, A. Conkie, and A. Syrdal (1998). "Diphone synthesis using unit selection." In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper F.2 (R52).
- [2] A. Black, "CHATR, Version 0.8, a generic speech synthesizer", System documentation, ATR-Interpreting Telecommunications Laboratories, Kyoto, Japan, March 1996.
- [3] A. Black and P. Taylor "The Festival Speech Synthesis System: system documentation" Technical Report HCRC/TR-83 Human Computer Research Centre, University of Edinburgh, Scotland, Jan. 1997.
- [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] Michael W. Macon, Andrew E. Cronk and Johan Wouters "Generalization and discrimination in tree-structured unit selection" proc. 3rd Synthesis Workshop, Nov. 1998.
- [6] 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.
- [7] 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.
- [8] R. Sproat, J. Hirschberg and D. Yarowsky "A corpus-based synthesizer" ICSLP Oct. 1992.
- [9] Y. Stylianou, J. Laroche and E. Moulines "High Quality Speech Modification based on a Harmonic + Noise Model. Proc. EUROSPEECH, 1995.
- [10] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal "The AT&T Next-Gen TTS System", 137th Acoustical Society of America meeting, Berlin 1999.