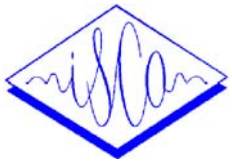

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Interaction of Units in a Unit Selection Database

Mark Beutnagel *Alistair Conkie*

AT&T Labs - Research, Florham Park, NJ, USA
http://www.research.att.com/projects/tts

10.21437/Eurospeech.1999-268

ABSTRACT

The purpose of this paper is to examine some aspects of unit selection for Text to Speech synthesis (TTS). We use Unit Selection as described in [2],[3]. The approach taken was to synthesize a large number of sentences and capture information about the selected units. We used approximately 25 million phonemes resulting from 10,000 files of AP newswire text. Given these statistics we looked at the units selected in order to try to analyse how unit selection works from a statistical point of view. Results of our analysis are presented.

1. Introduction

Unit selection synthesis techniques grew out of a dissatisfaction with older diphone concatenation techniques which allowed for only one example of any particular diphone. Diphone synthesis tended to sound unnatural. Much effort in diphone concatenation synthesis was spent on unit selection, although done off-line rather than the on-line version described here. Units were selected for their ability to join well to neighboring units on average.

This dissatisfaction, coupled with an increasing technological ability to build more memory and CPU-intensive speech synthesis systems motivates our work in general unit selection, which was first implemented in the ATR CHATR system and has come to be used as a paradigm for TTS.

One of the consequences of having a sophisticated unit selection system is an increase in complexity of the system. This can be addressed in various ways, and there are extremely useful visualization techniques available to study unit selection in detail. For example, output speech may be labeled to mark unit boundaries, or units may be displayed in their original contexts for comparison.

As far as we are aware previous analyses have concentrated on methods of visualizing data, focusing on specific examples. In this paper we begin to look at unit selection in the aggregate, collecting a large amount of unit selection data and analysing it statistically.

2. What is Unit Selection?

Automatic unit selection is the search through a large speech corpus at runtime with the aim of finding the best recorded units to render the desired speech. The problem is explained succinctly in [4] and further details of our implementation may be found in [3],[1]. In this section we briefly describe our choice of speech units, the cost functions which determine which units are "best," and how the search is implemented.

Phonetic and prosodic specifications typically come from text processed by prior components of the synthesizer: text normalization, pronunciation, phrasing, etc. In our system, the result is a sequence of half-phones [3], each of which has *target* values for features such as pitch and duration. Each unit in the database also has features that can be compared to the target values.

A *target* cost function estimates the distance between the predicted target values and a particular instance of a half-phone which is a candidate for concatenation. Unit selection attempts to minimize the target costs, choosing units which are as similar to the desired output as possible.

Unit selection also attempts to minimize perceptible acoustic mismatch between pairs of units to be concatenated. This is estimated by a *concatenation* cost function, which depends only on the acoustic properties of the two units, and is independent of their target costs. Both target and concatenation costs are implemented as weighted sums of feature-specific cost functions.

More formally, unit selection is the minimization of the total cost C for a sequence of n units, where the total cost is the sum of the target cost C^t for each unit (and target values t_i) and the concatenation costs C^c for each pair of units.

$$C = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) \qquad [1]$$

Candidate units are selected from the speech database based on unit phone identity, similarity of the unit's phonetic context to the context being synthesized, and the unit's target cost. A Viterbi search is used to find the best available path through the candidates, computing concatenation costs along the way. Figure 1 illustrates the architecture for the half-phones needed to synthesize the word "two." The top row represents the target half-phones; the lattice below the candidate unit instances

from the voice database. In reality, of course, there are many more candidate units than are shown here.

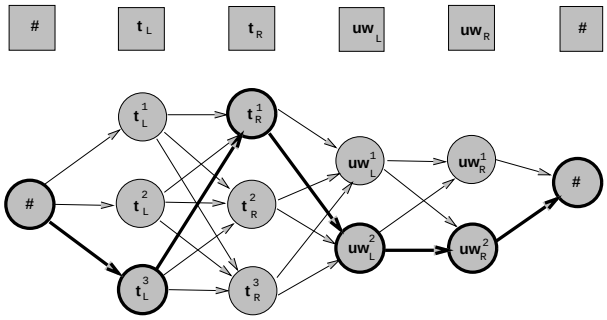

Figure 1: For half-phones in the word "two", a search finds the lowest cost path, selecting one candidate unit from each column for synthesis.

3. The Synthesis Data – The Experiment

The data for the analyses presented here was generated by presenting a large set of text files to the speech synthesis engine and recording all the units selected for synthesis. There are some 25,000,000 phonemes in the test data.

Clearly this data is collected with one particular version (snapshot) of the system with a specific set of parameters and optimizations. However we believe that the data presented here is generally useful in that we do *not* think it is unique to our system and configuration, but rather is typical of the kind of result that would, for example, be obtained if a different front end were used, or a different set of parameters. This, however, remains to be confirmed.

4. Unit Selection Corpus

The speech corpus consists of a substantial number of recordings of one female US English speaker. The total speech used in this study is approximately 1.5 hours. The recordings are of three broad types and there are approximately equal amounts of each type. All are read text, and the recordings adhere closely to the text. The text for the first type of recording is a set of sentences that were designed to be diphone-rich. The sentences themselves are meaningful English sentences. The second text set can be classified as newspaper style. The third is of interactive prompt-style utterances.

5. Some Introductory Statistics

The speech inventory database contains 42,185 manually segmented Arpabet phones, divided into 84,370 half-phone units. Phone names are distinguished by appending a "1" for the first (left) half or a "2" for the second (right) half. E.g. `aa` becomes `aa1 aa2`.

The synthesized unit selection data contains 50,000,000 half-phone units. These selected units cover 85% of all units in the inventory database.

Figure 2 shows the relative frequency of occurrence of units in the inventory database compared with occurrence in the synthesized unit database. A value of 1 signifies the relative frequency was equal in the inventory database and the experimental data. A high value means that there were more such units in the experiment than in the database. These values seem to be as we would expect. The values for the relatively rare phones are less than 1 reflecting the way that diphones were intentionally introduced into the inventory database. The high value for "pau" (pause) reflects the way that synthesis is carried out – there is a convention to add final pauses to the speech.

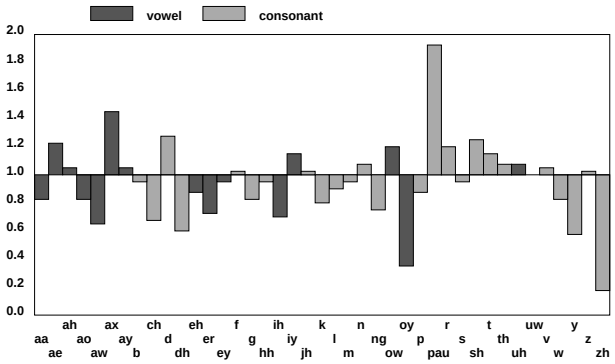

Figure 2: The relative frequency of occurrence of units in the database compared with occurrence in the unit database.

Table 1 shows how many of the inventory units in each phone class where actually selected while synthesizing the test corpus. Half-phone classes are sorted by frequency.

Type	Units used(%)	Type	Units used(%)
g2	100	...	...
ch2	98	zh1	80
g1	98	ih1	79
hh2	98	y2	79
v2	97	zh1	79
ch1	97	y1	76
hh1	97	uh1	63
...	...	uh2	61

Table 1: A partial list of the percentage of half phone inventory units actually used for synthesis.

If we look in detail at why the "uh" values are low (average 62%), we find that in the database "uh" occurs most frequently (45% of the time) associated with a following "l" whereas in the specifications generated for synthesis this is not the case (only 2% of the cases have a following "l"). This result is due to allophonic mismatch, and is a good illustration of the importance of matching a database – whenever feasible – to the material to be synthesized.

6. Diphone and Phone Boundaries

Two kinds of joins are possible: mid-phone (diphone) joins and phone boundary joins. The boundary between two units in the test set that are adjacent in the inventory database is not considered a join. The system as

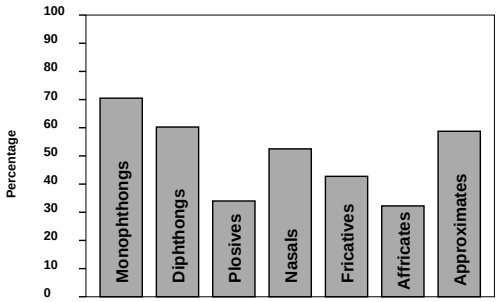

Figure 3: Percentage of synthesized phones without an internal (diphone) boundary (by class)

configured has an intentional bias towards diphone joins as opposed to phone boundary joins. This is achieved by a global weighting factor for each type of join. With the data at hand we can look further to see, for a particular type of phoneme, what kind of join is preferred and to what degree. This will be partly influenced by the variety of contexts available for any given synthesis configuration specified. It will also be influenced by how well any two units can join together. We might expect that for vowels, for example, a diphone join is a reasonable thing to have. On the other hand, for a stop consonant it might be reasonable to join at the phoneme boundary, especially if it is the closure boundary. Figure 3 shows how likely we are to have complete phonemes appearing in the synthetic output. This is equivalent to expressing how often diphone joins are avoided (e.g. by the choice of longer units).

Table 2 represents the percentage of diphones (e.g. `k2-ae1`) in the test set which were rendered as contiguous database units (e.g. units $\mathrm{k2}_{138}$-$\mathrm{ae1}_{139}$) rather than as a concatenation point (e.g. units $\mathrm{k2}_{138}$-$\mathrm{ae1}_{957}$). This data is analogous to the previous plot for phones. Like that plot, this data is collapsed into 7 phone classes and excludes silences. Because two phones, and hence two phone classes, are involved, this data is a 2-dimensional table. The left column represents the left-hand units, and the column headers represent the right-hand units. Percentages of "whole" diphones ranged from a low of 53% for affricate-to-affricate diphones (e.g. jh-ch) to a high of 99% for approximate-to-nasal transitions (e.g. l2-n1). Counts varied greatly, from 542 occurrences of affricate-to-affricate diphones up to 2.1 million for plosive-to-monophthong diphones.

%'s	**Mon**	**Dip**	**Plo**	**Nas**	**Fri**	**Aff**	**App**
Mon	91	97	95	85	95	96	88
Dip	93	91	97	93	96	93	94
Plo	63	77	96	91	93	92	71
Nas	92	91	88	98	96	95	96
Fri	61	76	88	90	97	94	73
Aff	77	74	82	96	87	53	86
App	85	86	97	99	96	96	95

Table 2: Percentage of diphone selections, summarized by class pairs, which are taken intact from the speech database. **Mon**ophthongs, **Dip**hthongs, **Plo**sives, **Nas**als, **Fri**catives, **Aff**ricates, **App**roximates.

7. Analysis Using the Specification

Another interesting topic we explored is that of finding out how well units chosen for synthesis have duration and average F0 values that correspond to the values specified. First of all, for F0s there are several factors that come into play in explaining the result of plotting chosen values of F0 versus specified values. In Figure 4 we see that the specified values seem to fall into several well-defined vertical bands. This we interpret as being the result of having an automatically-generated F0 contour. Such contours generally only have a limited set of curves or of target values that can be chosen and our system is no exception. Secondly, the vertical bands are elongated because in our system the chosen unit average F0 values can be reduced in proportion to the portion of the half phone that is considered voiced. Thirdly, even though the F0 specifications are not accurately adhered to the results of synthesis are nevertheless generally very intelligible and natural-sounding.

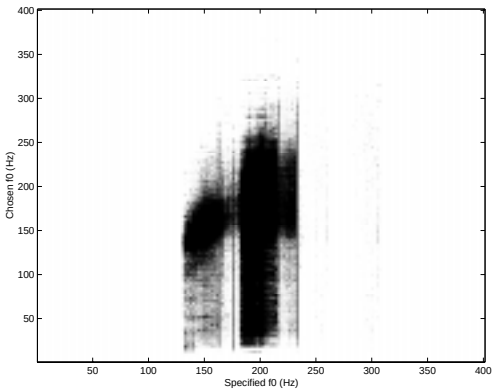

Figure 4: The F0 of chosen unit versus F0 value specified for joins that were chosen

Turning to the case of durations, illustrated in Figure 5 we see an even more marked effect from the specification. Duration specifications are quantized in units of 5mS which is reflected in the vertical striations which appear. The variations possible are even more marked than in the case of F0s. This reflects the fact that duration is generally speaking a secondary factor to F0 in terms of importance when we choose units. Again, the remark above about not adhering to specifications and the impact on quality applies.

8. F0 Mismatch at Boundaries

Figure 6 displays a plot of average F0 frequency versus average F0 frequency for pairs of units that are adjacent at synthesis time. The plot is of the entire 50,000,000 database. The further off the diagonal a point appears the greater the mismatch of F0. The values that appear in two bands for the high frequency region towards either axis are an artifact of the way that F0 values are calculated for partially voiced phonemes as mentioned above. This plot can be contrasted with Figure 4.

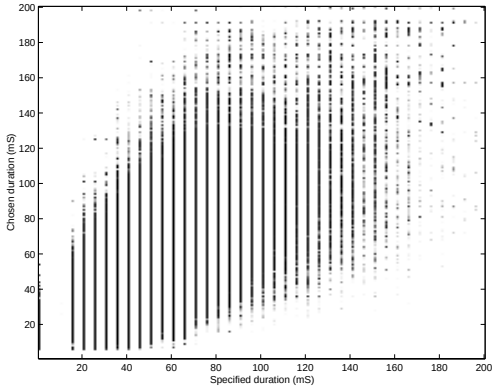

Figure 5: The duration of unit chosen versus duration value specified for joins that were chosen

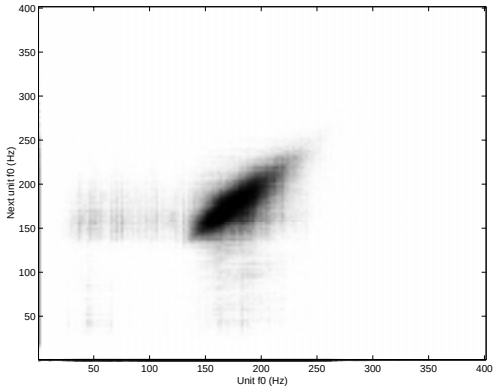

Figure 6: The F0 value of a unit on one side of a synthesis join versus the F0 value of the unit on the other side

9. Analysis of Rare Diphones

By rare we mean diphones that occur infrequently in the synthesis database. For this experiment we restricted ourselves to looking at the 300+ diphones that occur only once in the database. Most of these diphones were included in the database by embedding them in carefully constructed sentences. In a half-phoneme system such as ours, we can also construct diphones from pairs of half phonemes. This may be attractive since by this mechanism we can construct a large variety of pseudo-diphones with different F0s, durations etc, compared with single "natural" instances with fixed F0 and duration. Table 3 shows that there is a very clear preference for the "natural" diphones despite their fixed character. We conclude from this evidence that if a database is designed to include such diphones (as ours is) they will be preferred over combinations of half phones and synthesis quality will likely be higher as a result.

Class	Diphs(%)	Class	Diphs(%)
C-C	89	V-C	81
C-V	81	V-V	82

Table 3: A breakdown of the proportions of "natural" rare diphones used for synthesis, by class.

10. Conclusions

There are many aspects of how unit selection works and what constitutes useful material for a unit selection speech database. The work presented here gives an overview of some of the basic questions that can be asked about unit selection. Much work clearly remains to be done.

11. REFERENCES

1. Mark Beutnagel, Mehryar Mohri, and Michael Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. In *Eurospeech*, Budapest, 1999.
2. A. Black. *CHATR, Version 0.8, a generic speech synthesis.* System documentation. ATR - Interpreting Telecommunications Laboratories, Kyoto, Japan, March 1996.
3. Alistair Conkie. Robust unit selection for speech synthesis. *Proc. 137th Meeting of the ASA*, 1999.
4. A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP*, 1:373–376, 1996.