SPEECH SYNTHESIS USING NON-UNIFORM UNITS IN THE VERBMOBIL PROJECT

Simon King[†] *Thomas Portele* *Florian Höfer*

Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn
Poppelsdorfer Allee 47, D-53115 Bonn, Germany http://www.ikp.uni-bonn.de
[†]now at the Centre for Speech Technology Research, University of Edinburgh,
80, South Bridge, Edinburgh EH1 1HN, GB http://www.cstr.ed.ac.uk
email: Simon.King@ed.ac.uk

ABSTRACT

We describe a concatenative speech synthesiser for British English which uses the HADIFIX [8] inventory structure originally developed for German by Portele. An inventory of non-uniform units was investigated with the aim of improving segmental quality compared to diphones. A combination of soft (diphone) and hard concatenation was used, which allowed a dramatic reduction in inventory size. We also present a unit selection algorithm which selects an optimum sequence of units from this inventory for a given phoneme sequence. The work described is part of the concept-to-speech synthesiser for the language and speech project Verbmobil [12] which is funded by the German Ministry of Science (BMBF).

1. INTRODUCTION

Unit concatenation is now the most popular form of speech synthesis. Typically the units are diphones, that is, phone-sized. This means that there will be a join in each and every phone. Since certain phonemes do not join well in this way, we use a mixed inventory system with units typically the size of a demisyllable. This means that fewer units are typically required for a given utterance, but a larger inventory of units is required. We overcome the problem of increased inventory size by allowing some joins at phone boundaries.

2. THEORY

2.1. Hypothesis

We base our concatenation rules, and hence the unit inventory, on the assumption that groups of phonemes have similar co-articulatory effects on neighbouring segments. For example, **t** and **s** have similar effects on neighbouring vowels.

2.2. Motivation

We wish to exploit the hypothesis in order to reduce the number of units in the inventory. To do this we relax the requirement for a complete inventory – one with all possible combinations of demisyllable initial and final vowels / consonant clusters. This means that some units will be concatenated at phone boundaries (so called *hard* concatenation). This is demonstrated by example in section 2.4.

2.3. Evidence

The assumption in section 2.1. is based on phonological knowledge. We can demonstrate the validity of this by examining spectrograms of phones taken from various contexts.

Figure 1 shows spectrograms of examples of the vowel **/A:/** taken from three different left contexts. Clearly, the two examples from the contexts **/b_/**, and **/p_/** are quite similar, whilst the example from the context **/r_/** is very different.

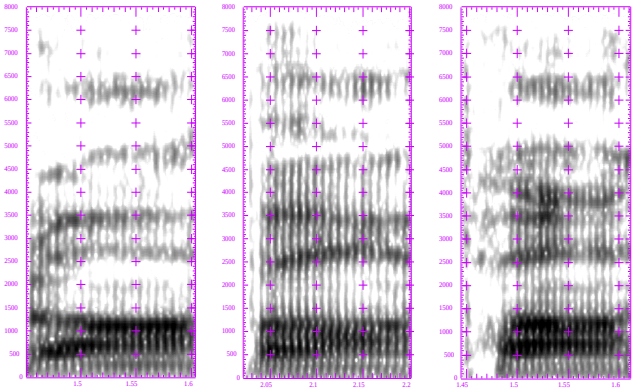

Figure 1. Spectrograms of **/A:/** in contexts **/r_/**, **/p_/**, and **/b_/**. Vertical scale is 0–8kHz

2.4. Example

The following sections will be better understood after a brief example, which demonstrates how a phoneme sequence can be generated without the need for a *complete* inventory. In figure 2 the upper line is the target phoneme sequence, and the lower line is the unit sequence. Phones crossed out are deleted, notation is SAM-PA.

Units can join in two ways : *soft* and *hard*. In *soft* concatenation, the join is made within the phone and in *hard* concatenation the join is made at a phone boundary. Because we allow *hard* concatenation, a phone can be taken from a unit in which it occurs

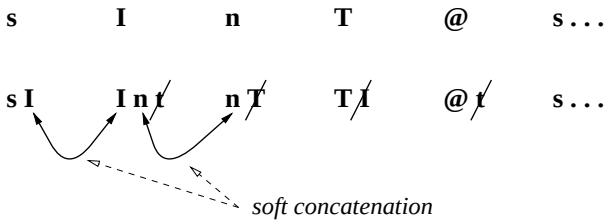

Figure 2. Mixed concatenation methods

in a similar, but *not identical*, context as the target phoneme. In the example, the **/@/** is taken from the unit **/@ t/**, because the **/t/** closely matches the target context of **/s/**. This is shown in the table in section 2.6.2.

Some phonemes, such as the **/T/** above, never have *soft* joins because they were found to be too prominent. Also note that the inventory contains no special units containing silence for phrase initial/final positions. Despite these limitations, high quality synthesis is possible by choosing an appropriate sequence of units.

2.5. Implementation

Restating the hypothesis from section 2.1. : if a desired target phone is not available from the exact context required, we can take it from a *similar* context.

We define *similar* in one direction only, that is, if the context **/X_/** can be used instead of **/Y_/**, then this does not imply that **/Y_/** can be used instead of **/X_/** (although this may frequently be the case). However, a limitation of the system as it stands, is that the same lists are used for both left and right effects – that is, if the context **/X_/** can be used instead of **/Y_/**, then the context **/_X/** can be used instead of **/_Y/**. This is clearly an approximation, and is discussed further in section 6.2.

So, we can test our hypothesis by making lists of alternative contexts for each phoneme in our set. The main factor in grouping phonemes is place and manner of articulation.

For our proposed inventory scheme, every phoneme must have such a list, and we will rank these lists in order to choose between various alternative contexts from the inventory.

2.6. Context equivalences

In the following tables, the column labelled 'phoneme' is the desired context (left or right) of a target phoneme, and the other column gives alternative contexts, listed best–first.

2.6.1. Stops

phoneme	alternatives	phoneme	alternatives
p	f	b	d
t	s S T tS	d	b dZ
k	g	g	k

2.6.2. Fricatives and affricates

phoneme	alternatives	phoneme	alternatives
f	p s	v	b
T	t tS	D	d
s†	S T t† f	z	Z d
S	T s t	Z	z s S
h	*none*	tS	t d p S s
dZ	d t		

† see example in section 2.4.

2.6.3. Nasals

phoneme	alternatives	phoneme	alternatives
m	n	n	N m
N	n m		

2.6.4. Liquids and glides

There were thought to be no suitable alternative contexts for these phonemes : r, l, w, j.

2.6.5. Short vowels and shwa

phoneme	alternatives	phoneme	alternatives
I	@ i: eI	e	@ eI
{	@ A: V	Q	@ O:
V	U @ I@ {	U	V u: @

phoneme	alternatives
@	Q I @U A: e { V U u: 3: O: aU eI

2.6.6. Long vowels

phoneme	alternatives	phoneme	alternatives
i:	I u:	eI	I
aI	I	OI	I
u:	U	@U	U
aU	U	3:	e
A:	@	O:	@
I@	@	e@	@
U@	@		

2.6.7. Syllabic consonants

phoneme	alternatives
=l	3: @ l
=n	3: @ n

3. INVENTORY

Following Portele's work on German synthesis [8], we used the same principles to design an inventory for British English. The units in the inventory fall into five categories : demisyllables (initial and final), clusters (vowel and consonant) and suffixes.

The size of the inventory was dramatically reduced by using the assumption from section 2.1. Units containing phone sequences which can be constructed from two or more other units can be eliminated.

3.1. Construction

For each class of unit (initial demisyllable, etc.), a table of all possible units was constructed. This was

achieved by making lists of unit initial and final clusters/vowels. A fragment of the table for initial demisyllables is shown below :

	3:	@	@U	A:	aI
b	b3:	b@	b@U	bA:	baI
d	d3:	d@	d@U	dA:	daI
g	g3:	g@	g@U	gA:	gaI
p	p3:	p@	p@U	pA:	paI
sp	sp3:	sp@	sp@U	spA:	spaI
t	t3:	t@	t@U	tA:	taI
⋮	⋮				

Table 1. Part of the full table of initial demisyllables

3.2. Elimination

The elimination of units was done systematically, starting with the table of all possible units. The size of this table depends on the number of consonant clusters (e.g. `pl`, `kw`, `str`, `nd`, ...) used in the demisyllable units. We then eliminated units; for example:

eliminate unit **InT** $\longrightarrow$ *use* **Int** *and* **nT**

After this process, we were left with around 700 of the original 3100+ units.

3.3. Recording

The inventory was recorded with a suspended microphone in an anechoic chamber. No laryngograph was used. The initial inventory was recorded in a single day, with additional units being recorded in a subsequent session. Conditions were carefully controlled to minimise mismatch between the two sessions.

The sampling rate of the inventory is 32kHz, rather than the more common 16kHz, to maximise perceived quality and intelligibility.

3.3.1. Labelling

Pitch marking was carried out automatically using PMARK [9], and an initial automatic segmentation was carried out using the HTK toolkit. Both pitch marks and segmentation labels were hand checked and corrected.

3.4. Trial

The first inventory of 700 units was used to synthesise some test utterances. After this trial, we decided to add further units, including a small number of vowel-consonant-vowel and consonant-vowel-consonant units.

3.5. Final inventory

We arrived at a final unit inventory consisting of around 850 units of 7 types : initial demisyllable, final demisyllable, vowel-vowel, consonant-consonant, suffix, vowel-consonant-vowel and consonant-vowel-consonant. The number of units in the last two categories is quite small.

4. SELECTION ALGORITHM

Since the unit sequence for a target phoneme sequence is not unique, we must define an *'optimum'* sequence and write an algorithm to find one for every possible input.

The *optimum* sequence is defined *locally*, and in terms of the context equivalence lists of section 2. For each phoneme in the target utterance, the unit is selected which contains it in the *closest* context to the target context. The definition of *closest* is given by a tree structured set of rules which use the context equivalence lists of section 2.

The rules give a score to each unit being considered as a source for a particular target phoneme, and the highest scoring unit in the inventory is chosen. The choice of a scoring system means that a unit sequence is found for *any* target phoneme sequence *whatever* the inventory size[1].

4.1. Rules

Scoring of units from the inventory is done with a set of tree-structured rules. These rules were constructed by hand, and the scores assigned by each one were hand picked to achieve the desired result. Both left and right context is considered by the rules, with greater importance attached to right context since coarticulation is largely planned [13].

Context of the target phoneme is examined up to three phonemes left and right. Exact phonetic matches between the target sequence and the unit under consideration score more highly than contextually equivalent matches, and context positions closer to the target phoneme score more highly than ones further away.

As well as using the assumption from section 2.1., the rules are designed to prefer initial demisyllables in utterance-initial positions, and final demi-syllables or suffixes in utterance-final positions.

4.2. Concatenation

When units are concatenated with *soft* joins – that is, within a phone – a cut point must be defined. No attempt is made to find an optimal point based on spectral or other measures, such as in [3], mainly because the actual signal processing (PSOLA) is handled by a separate module [7].

The within-phone concatenation points are chosen on a phoneme by phoneme basis, for example, **/w/** will always be joined 30% into the segment. Clearly, there is scope for improvement here, even without access to the acoustic signal – either optimising the cut-points for different contexts, per unit, or per unit pair.

5. IMPLEMENTATION

The work described here forms part of a complete concept-to-speech[2] synthesiser for British English.

[1] provided there is at least one example of every phoneme!

[2] input is marked up text in *EI* (Erweiterte Informationen, [6]) format

The unit selection and prosody generation modules were implemented in an architecture similar to that in [2], partly because the software implementation uses [11]. Parts of the word accent algorithm are taken from [4]. The assignment of pitch accents is similar to that in [1]; realisation of pitch accents using a *tilt* representation is taken from work by Taylor [10].

A more detailed description of the full system can be found in [5].

6. ASSESSMENT

Only an informal assessment of quality was performed due to time constraints and the difficulty of finding native speakers of English in Germany. The trial inventory proved very useful in assessing the quality of the first (700 unit) inventory, and it was found that the addition of another 150 units improved quality noticeably.

6.1. Compromise

Unlike diphone synthesis, where a complete inventory is necessary, a non-uniform unit inventory allows a tradeoff between size and quality. An inventory size reduction not only saves storage space and processing time, but reduces the amount of expensive labelling work involved in generating a new inventory.

6.2. Limitations

The implementation had several limitations. Firstly, the context equivalence lists were used for both left and right effects. This method assumes the gross position of articulators is the major factor, rather than their dynamic behaviour, and does work reasonably well. However, the assumption is clearly not true, particularly for diphthongs, and some improvement could be expected from using separate lists for left and right.

Secondly, no special units containing silence were used. This was alleviated by the use of rules to prefer initial demi-syllables in utterance-initial positions, and final demi-syllables or suffixes in utterance-final positions.

Thirdly, the use of fixed within-phone cut-points was suboptimal, and there is considerable scope for improvement through either the use of context-sensitive cut-points, or by reference to the acoustic signal.

Finally, the method for eliminating units from the tables of all possible units was fairly time-consuming, but is only required once per language.

7. CONCLUSION

We have shown that the use of non-uniform units can overcome some of the problems of diphone synthesis without the inventory size becoming impractically large. A method of reducing the number of units in the inventory, and an algorithm for determining the unit sequence for an arbitrary target phoneme sequence have been described.

ACKNOWLEDGEMENTS

Many thanks to Paul Taylor for much helpful advice.

NOTES

The inventory, list of units and source code are available from the author. Verbmobil reports are available at http://www.dfki.uni-sb.de/verbmobil/ .

REFERENCES

[1] Alan Black and Paul Taylor. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *Proc. ICSLP '94*, Yokohama, Japan, 1994.

[2] Alan Black and Paul Taylor. CHATR : a Generic Speech Synthesis System. In *Proc. COLING94*, Kyoto, Japan, 1994.

[3] Alistair Conkie and Stephen Isard. Optimal coupling of diphones. In Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 293–304. Springer Verlag, New York, 1996.

[4] Julia Hirschberg. Using discourse content to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, editors, *Talking Machines*, pages 367–376. North-Holland, 1992.

[5] Simon King. Final report for Verbmobil 1 Teilprojekt 4.4. Technical report, IKP, Universität Bonn, October 1996. Verbmobil-Report 195.

[6] Stefan Merten. Erweiterte Informationen in Sprachsynthesesystemen. Technical Report Verbmobil-Memo 112, DFKI Kaiserslautern, September 1996.

[7] Univ. of Dresden. Zeitsyn, 1996.

[8] Thomas Portele, Florian Höfer, and Wolfgang Hess. A mixed inventory structure for German concatenative synthesis. In Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 203–227. Springer Verlag, New York, 1996. Also available as Verbmobil-Report 149.

[9] Ansgar Rinscheid. Automatische Bestimmung von Periodenmarken mit dem emark-Algorithmus. In *Fortschritte der Akustik, DAGA '93*, pages 1048–1051, Frankfurt, 1993.

[10] Paul Taylor and Alan Black. Synthesizing conversational intonation from a linguistically rich input. In *Proc. ESCA Workshop on Speech Synthesis*, Mohowk, N.Y., 1994. ESCA.

[11] Paul Taylor, Simon King, and Alan Black. CSTR Speech Tools, 1996/7. email {pault, simonk, awb}@cstr.ed.ac.uk.

[12] W. Wahlster. Verbmobil - Translation of Face-To-Face Dialogs. In *Proc. Eurospeech 93*, pages 29–38, 1993.

[13] D. H. Whalen. Coarticulation is largely planned. *Journal of Phonetics*, 18:3–35, 1990.