Syllable Recovery from Polysyllabic Words

Mark Tatham
Eric Lewis – University of Bristol

Reproduced from Proceedings of the Institute of Acoustics (1998). St Albans: Institute of Acoustics. 279-288

INTRODUCTION – The need for syllable recovery

For general purpose concatenated waveform speech synthesis [1] an exhaustive inventory of stored waveforms for re-arrangement and concatenation is needed. Our **SPRUCE** system [2] is syllable and word based, and to be truly general purpose its inventory needs examples of all possible syllables. The high level synthesis engine (responsible for the segmental phonology and prosody of utterances) is already general purpose – but its use is limited by small lower level inventories of re-combinable waveforms.

The purpose of the feasibility study reported here was to determine to what extent we could take one of the word based limited domain versions of the system, **MeteoSPRUCE**, designed for weather forecasting applications, and extend its usability by excising syllables from polysyllabic words in its inventory and recombining them to form new words – thus widening usability without the need for re-recording [3] [4] [5] [6] [7].

PRELIMINARIES

Before embarking on the task of excising syllable waveforms from longer stretches we needed to be clear on a number of basic theoretical points:

1. The symbolic representations of phonology [8] are often of limited help in identifying syllables in the acoustic signal. For example, the phonological concept of boundary does not easily carry through to the waveform.
2. Representations at the phonetic level [9] are also symbolic, and although we can identify a phoneme or allophone string corresponding to a phonological syllable there is often no clear feature to help us in acoustically delimiting syllables even with this level of representation.
3. The very notion of boundary as a point to make a cut in the waveform can itself be misleading. Acoustic syllables can often be thought of as overlapping, telescoping or merging and, in terms of timing, one syllable may ‘begin’ before the previous one has ‘ended’; that is, the time allocated to a sequenced pair of syllables is not always the sum of the times each would occupy on its own.
4. The coarticulation [10] or coproduction [10] [11] responsible for temporal overlap is also responsible for spectral overlap. Even if cuts are made temporally at the ‘right’ places there is a serious problem of inclusion of spectral boundary effects in both syllables of a separated pair when they are recombined in new but ‘wrong’ contexts.

A SIMPLE EXAMPLE OF WHAT WE HOPE TO ACHIEVE

The limited 2000-word **MeteoSPRUCE** database contains waveforms of the words *unsettled* and *likely*: suppose we would like to use these to create a new word object *unlikely*. The idea is to detach the syllable *un* and place it in front of the *like* syllable of *likely*. Phonetic syllable boundaries are marked in the database by morpheme if possible (as in this case), or phonologically. Fig.1 shows the database entries.

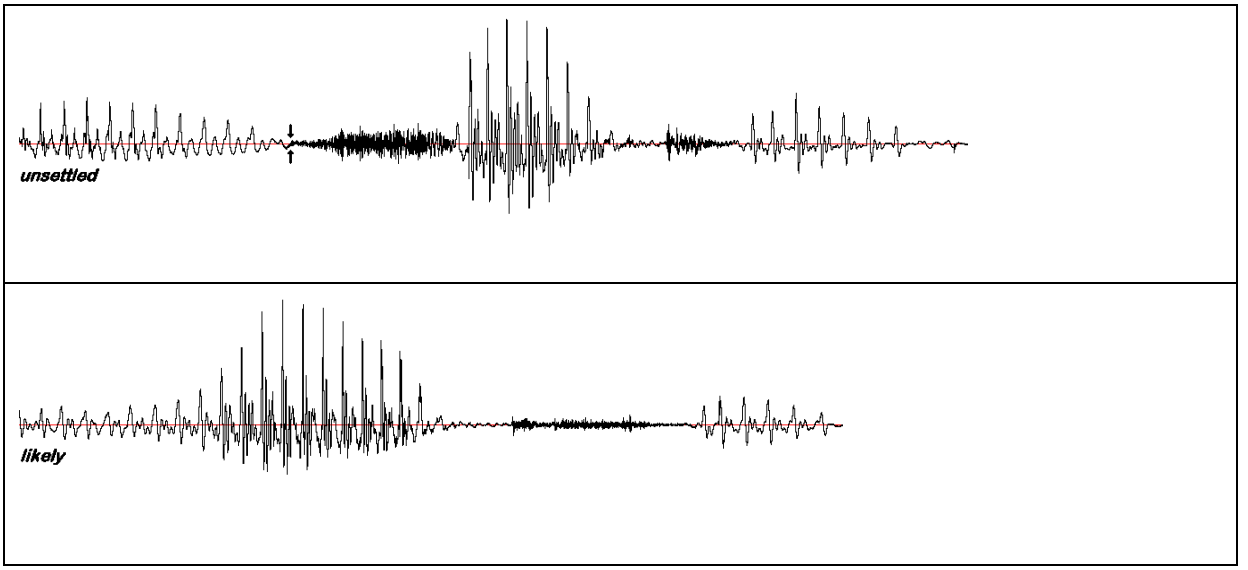

Fig.1 Waveforms of *unsettled* and *likely* as they appear in the **MeteoSPRUCE** database. The arrows indicate the point at which *unsettled* is marked for the syllable boundary.

By cutting *unsettled* at the end of the last pitch period associated with *un* we can paste the beginning of the file to the start of *likely* to produce a new reconstructed word object **unlikely*. In our overall model we refer to such stretches of actual waveform as phonetic syllables. Fig.2 compares the result of conjoining the phonetic syllables with an actual recording of *unlikely* which *does* happen to be in the database.

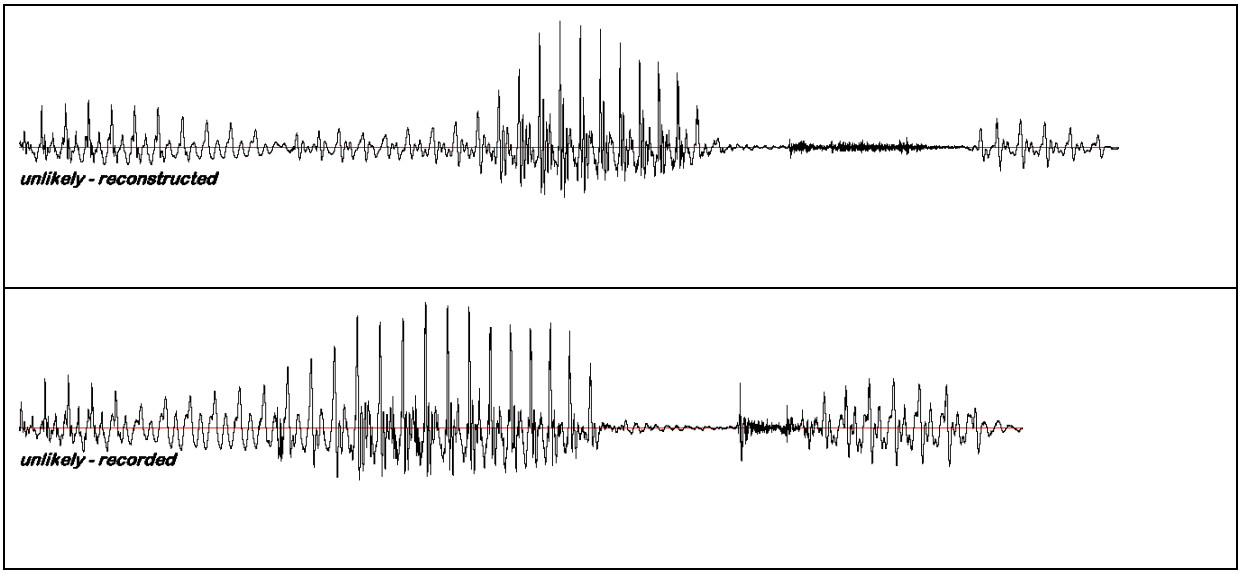

Fig.2 Reconstruction of **unlikely*, and the waveform of *unlikely* as it occurs independently in **MeteoSPRUCE**.

Even allowing for the fact that we would not expect any two versions of *unlikely* to be acoustically identical, we can see that there are a number of things wrong with the reconstructed version. In particular the transition between the syllables *un* and *like* appears protracted and awkwardly joined (this is true both auditorily and visually). An improved reconstruction is obtained by means of a normalising procedure which deals with syllable overlap. The procedure involves setting up a **synthetic syllable**, derived in the normalisation process from the **phonetic syllable** (see below).

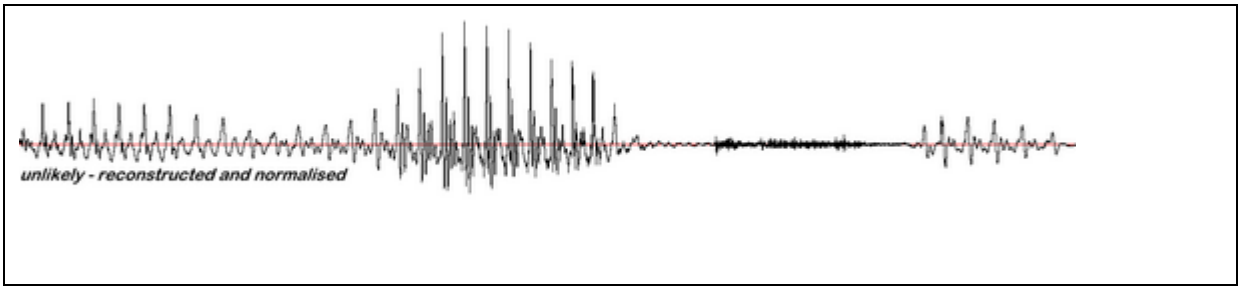

Fig.3 Reconstruction of **unlikely* using the derived synthetic syllable *un* and the recorded word *likely* (also normalised at the beginning of the word to form the synthetic syllable *like*).

Degree of coproduction between syllables appears to be context dependent, and we discuss later how the phenomenon varies – we deliberately picked the syllable *un* in *unsettled* because it showed the minimum of 'telescoping' coproduction. For the moment though we have identified three main stages in the reconstruction procedure: a. phonetic syllable excision, b. normalisation, c. synthetic syllable conjoining.

DESCRIBING AND IDENTIFYING SYLLABLES

We must clarify our concept of **recovery:** For the reasons given above the desired syllable waveform is often not there to be recovered. A stretch of waveform of approximately the right length may well be excised from a suitable word, but because of coproduction effects is almost certainly not directly re-usable in other than a similar coproduction context. Recovery means *more* then excision therefore: it means also **reconstruction**. The excised stretch of waveform – the **phonetic syllable** – is going to be used as the basis for reconstructing the desired waveform – the **synthetic syllable**.

The procedure we have developed for syllable recovery calls for syllable models defined on three different levels, each serving a different purpose. The relationship between the three definitions needs to be explicit. The three levels are used to characterise the **phonological syllable**, the **phonetic syllable** and the **synthetic syllable**.

- **Phonological syllables** – In linguistics this syllable defines a unit higher than individual sound segments [12]. It was introduced for two purposes: to form a framework for characterising the sequencing of simple segments, and to provide the primary unit for modelling a language's prosodic behaviour. Considerations of phonetic detail are irrelevant at this level: what is important is the way in which individual segments are organised hierarchically or non-linearly into a syllabic unit. We characterise phonological syllables using the normal tools and methodology of linguistics [13].

The phonological syllable is important in our model because it enables us to refer directly to a listener's perception of speech sound sequencing – the phonological syllable characterises for us the result of successful perception. Since our synthesis philosophy revolves around satisfying a listener's perceptual abilities we need a level specifically designed to capture just that.

So, for example, listeners can identify a unit at the beginning of the word *unsettled*, pronounce this unit in isolation and tell us that it is the *same* (phonologically) as a unit identified at the start of the word *unlikely*. This cognitive similarity is not the same as acoustic similarity – coarticulatory phenomena constrain the two *un*s to be systematically acoustically different. The goal of the reconstruction procedure will be to use a portion of the waveform of *unsettled* to change *likely* into a correctly *perceived* new word *unlikely*.

- **Phonetic syllables** – At this level the syllable is a descriptive unit characterising part of an acoustic signal which prompts a listener to identify a *phonological* syllable. We confine our model to the acoustic domain, so this is the place where distinguishing acoustic features are identified, as well as other acoustic features which may contribute little or nothing to the perceptual process. The model describes the waveform using the

normal parameters and descriptive tools of acoustic phonetics [14].

The phonetic syllable model is a complete stretch of waveform. What 'sounds' are sequenced in that waveform is a phonological matter rather than a phonetic one for the purposes of our reconstruction procedure. So in the illustration above the phonetic syllable is the waveform identified as triggering the perceived phonological syllable *un – and* its phonetic description.

There has been a lot of discussion concerning the relationship between phonetic and phonological models of the same stretch of speech [15]. For us, the phonetic syllable models the acoustic signal and the phonological syllable models a listener's cognitive *response* to the signal. The two models are formally linked in as much as they each deal with the same signal. Notice that we are using the term to refer to both a stretch of waveform *and* its acoustic model.

- **Synthetic syllables** –This is an acoustic model of a stretch of waveform which can be manipulated to trigger in the listener a response corresponding to the appropriate phonological syllable. The synthetic syllable may or may not be the same as the phonetic syllable from which it is derived.

In **SPRUCE** a syllable waveform in the database can exist as a phonetic syllable (this is the model of the original human waveform, say, of a monosyllabic word like *snow*), but it also exists as a synthetic syllable – a model able to be concatenated with another to produce a new word like *snowing*. The synthetic syllable is derived from a phonetic entry in the database by a normalisation procedure which varies in complexity depending on syllable type and the environment from which it is to be excised – that is, the normalisation process for deriving synthetic syllables from phonetic syllables is sensitive to the *type* of syllable as well its original context.

SYLLABLE TYPES AND CONTEXTS

We classify syllable types in terms of their phonological segmental beginning (onset) and ending (coda). Initially we were concerned about coarticulatory effects between phonetic syllables, i.e. that reconstructed words should have the correct temporal *and* spectral phonetic properties at new syllable boundaries. However, to take full account of all acoustic effects of quality change resulting from coproduction all combinatorial possibilities would need to be considered. For the current feasibility study we scaled the problem down to make a reasonable start and to set up a working model of how in the first instance syllables combine *temporally*. Thus we defocused considerations of phonetic quality at syllable boundaries in favour of the temporal properties of syllable onset and offset under coproduction or overlap conditions.

A careful examination of all word objects in the **MeteoSPRUCE** database led us to believe that our first working model might deal only in terms of initial and final segment *types*, rather than take account of the differences between all possibly occurring different *segments*. We chose to establish types of segment according to the usual parameters of phonetic classification [4]. Thus, all syllables include a vowel segment preceded by up to three phonetic consonants and followed by up to four, thus:

- $C_0^3 + V + C_0^4$

There are constraints on the consonantal sequences which fortunately cut the number of possible syllables down to one which can be managed – perhaps around 8000, though variations dependent on stress and timing greatly enlarge this number. However, by taking only initial and final zero or one consonant *types*, we narrow down the combinatorial possibilities considerably. So, syllables may begin and end as:

- vowels (including initial semivowels and [h]) – *a*ll *(*y*ou, h*ow*), m*e [initial, final]
- diphthongs – *ai*r, *dr*y [initial, final]

- voiced fricatives (including final voiced affricates) – *those, breeze (merge)* [initial, final]
- voiceless fricatives (including final voiceless affricates) – *said, once (French)* [initial, final]
- initial voiced plosive *stop phases* (including voiced affricates) – *go (join)* [initial]
- initial voiceless plosive *stop phases* (including voiceless affricates) – *too (chart)* [initial]
- final plosive *burst phases – flood, right* [final]
- nasals – *melt, mean* [initial, final]
- liquids – *right, like, more* (not allowed in Southern English **MeteoSPRUCE**), *full* [initial, final]

Notes:

5. The examples given of initial and final types are from the **MeteoSPRUCE** database. We choose monosyllabic words for the examples because the recording and normalisation procedures eliminate initial and final coarticulatory effects. Syllables which are not in the database as monosyllabic words have to be excised from words which *are* in the database: in such cases coarticulatory effects *are* present and ignored in this first working model.
6. Vowels and diphthongs are entered as different types when a syllable does not end or begin with a consonant because diphthongs appear to be more resistant to coproduction trimming or truncation.
7. Plosives are separated into word initial and word final – it is the *stop phase* which is important for conjoining in initial position, and the *burst phase* which is important in final position. However in final position we did not find it necessary to distinguish between voiced and voiceless examples – for the speaker who made the recordings for the database (author MT) there was no detectable difference in the burst phases (Southeast England accent).
8. All nasals appear to behave similarly (*ng* in final position only).
9. All liquids appear to behave similarly.

TAKING THE *UN* EXAMPLE FURTHER

Let us extend our earlier simple example of what we hope to achieve in syllable recovery by examining in more detail how the syllable-length prefix *un* combines temporally with different following syllable types.

combination type	***free-standing***	***prefixed by* un-**
+ *initial voiceless fricative*	[s] in *certain* – 94ms	[s] in *uncertain* – 98ms
+ *initial voiceless plosive*	[p_{stop}] in *pleasant* – 80ms	[p_{stop}] in *unpleasant* – 41ms
+ *initial voiced plosive*	[b_{stop}] in *broken* – 65ms	[b_{stop}] in *unbroken* – 11ms
+ *initial nasal*	[*known* not in database]	[n_2] in *unknown* – 88ms (7 pitch cycles)
+ *initial liquid*	[l] in *likely* – 89ms (7 pitch cycles)	[l] in *unlikely* – 52ms (4 pitch cycles)

Table I – Examples of initial segment durations of five types*: voiceless fricative, voiceless plosive,*

voiced plosive, nasal, liquid. Note that when prefixed by *un* the initial segments (other than the voiceless fricative) appear truncated by coproduction.

Table I shows data measured from examples with and without *un* in the database. Of the five combination types shown, the one showing the least coproduction overlapping is the + *voiceless fricative* type. Here we find that [s] has a similar temporal value whether prefixed or not. An initial voiceless plosive appears halved in duration, with an initial voiced plosive truncated even more. An initial nasal seems unaffected, retaining its full duration (observation based on other words beginning with [n]); while an initial liquid is almost halved in duration. Apparent truncation is one of the acoustic phonetic effects of articulatory coproduction – this is probably an overlapping or telescoping effect, rather than strictly truncation. Careful listening to the individual sections of the prefixed examples enables detection of coarticulatory phenomena – frication can easily be seen and heard, for example, during the last two pitch cycles of [n] in *uncertain*.

combination type	***free-standing [measured]***	***prefixed by* un *[predicted]***	***prefixed by* un *[measured]***
+ *initial voiceless fricative*	[f] in favourable – 89ms	89ms	80ms
+ *initial voiceless plosive*	[p_{stop}] in pleasantly – 78ms	39ms	41ms
+ *initial voiced plosive*	[d_{stop}] in does – 55ms	9ms	17ms
+ *initial nasal*			
+ *initial liquid*	[l] in like – 7 pitch cycles	4 pitch cycles	5 pitch cycles

Table II – Comparison of predicted durations of segments following *un* (based on measurements made on words without *un*) with measured durations. No suitable data was available for initial nasals. Initial voiceless fricative and plosive give good results as does the initial liquid, but the result for the initial voiced plosive is disappointing (see the text for a possible explanation).

Although we are dealing only with a tiny amount of data we thought it would be worth seeing whether we could use the results in Table I to predict the behaviour of other segments in similar type environments. We do not present this as a valid generalisation, but as an illustration of *procedure*. Table II shows the results. In just one example of each type we were able satisfactorily to predict the changes brought about by prefixing *un* to words with an initial voiceless fricative, voiceless plosive and liquid. The result for an initial voiced plosive was disappointing, but could be explained by a segmentation measurement problem: we found it difficult to differentiate between the ending of the nasal in the prefix and some possible vocal cord vibration associated with the following plosive – we tended to label the entire duration of vocal cord vibration as nasal, whereas it might have been nasal + plosive voicing. It is well known that voicing trails off during the closure phase of a voiced plosive, but it is difficult to say whether we are dealing with nasal 'intrusion' into the stop (which we assumed) or vocal cord vibration meriting the label *non*-nasal. In practice this fine linguistic point need not bother us.

The most interesting case in our data set is the coproduction overlap when *un* is followed by a liquid. Fig.4 illustrates the overlap process. In this example

- the sentence to be synthesised calls for the word *unlike* which we assume is not in the database;
- the database is searched for the syllables *un* and *like*;

- *un* is found in *unsettled*, indexed that it has been only minimally coproduced;
- *like* is found in the word entry *like* (with no coproduction);
- according to our overlap rule (see **Section 5** below), if syllable one ends in a nasal and syllable two begins with a liquid then syllable one is trimmed by three pitch periods at its end and syllable two is trimmed by three pitch periods at its start;
- the trimmed syllables are conjoined to form the new word *unlike*.

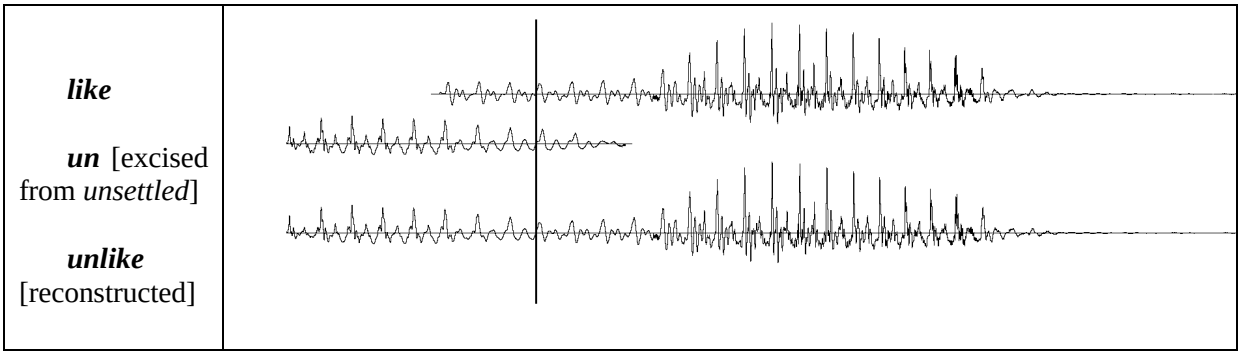

Fig. 4 – Example of the excision and reconstruction procedure. a. The word *like* is selected; b. the prefix *un* is excised from a word (*unsettled* in this case) in which it has been minimally coproduced; c. after trimming to simulate coproduction the two files are conjoined. Any mismatch in pitch period frequency is normalised in the subsequent intonation algorithm applied later when the word is used in a synthesised utterance.

There are several negative points to note in this procedure:

- *un* in *unsettled* does have slight coproduction (high frequency [s] derived signal is clearly visible toward the end of the excised waveform) – it is assumed that the trimming process will remove any residual qualitative coarticulatory effects;
- the reconstituted word exhibits *conjoining* not coproduction – that is there is no forward or backward coarticulation consistent with a genuine *unlike* as produced by a human being;
- 'utterance rate' must be consistent throughout the database – it would be no good if some syllables were recorded faster than others – this is not a problem in the temporally normalised **MeteoSPRUCE**;
- the stress pattern of the new word has to match the original stress values of the excised syllables – for example, the secondary stressed *un* we are using may well not be satisfactory for reconstituting, say, the word *under* – which begins with a primary stressed *un*.
- there may well be a change of fundamental frequency at the boundary – any such change has to be neutralised.

For us the most difficult and theoretically unsound of these points is that the synthetic syllables have been conjoined and not properly coproduced – they are thus not phonetic syllables. The only question to be asked here is *does it matter?* Our initial answer to this question is: *Yes, sometimes, but certainly not always.* In the terminology of our model the real question is: *Can the synthesised combination of syllables trigger the correct phonological response in the listener?* And subsequently, *Can the synthetic word be perceived as having no errors?* And the tentative answers here are: *Yes, almost always*, and *Usually*, respectively. We shall be seeking firmer answers to these questions by more formal systematic testing.

RE-COMBINING RULES

Once our model has established a means of classifying types of syllable we can proceed to determine how they pattern when linearly combined. Any such patterns are expressed as rules. Again this is a first approximation – our objective is to determine how far we get in triggering the acceptance of appropriate phonological syllable combinations in the listener with the simplest model. Perhaps it will be necessary to adopt a more elaborate approach involving recognising that, as with our internal syllable model, a non-linear model may be more useful in characterising how syllables concatenate.

The polysyllabic words in the **MeteoSPRUCE** database were examined with a view to determining boundary effects when syllables are linearly sequenced. Having no regard to qualitative coarticulatory effects on the acoustic signal we set out to examine temporal effects of coproduction or overlap.

We were able to determine that minimal temporal effects occurred at the following boundaries:

- any syllable followed by a pause;
- vowel, plosive, nasal, liquid offset + fricative onset (e.g. *a + fraid, ad + vance , un + certain, al + so*).

This gave us a basis for modelling some basic *synthetic* syllables – phonetic syllables temporally unaffected by boundaries, or one in complete isolation (e.g. a monosyllabic word). Thus: *a, fraid, ad, vance, un, al, so*. It also enabled our first re-combining rules:

1. ***rule:*** There are no temporal boundary adjustments to be made where the boundary is preceded by {vowel, plosive, nasal, liquid}–final types, and followed by the {fricative}–initial type. Note that if the plosive and fricative are homorganic the burst phase of the plosive is trimmed away. [There are no polysyllabic word examples of this in the database, but the process is akin to what happens in a word like *effects*, in which the [t] is not released in the example in the database.]
2. ***rule:*** Where the boundary is preceded by a {fricative}–final type and followed a {fricative}–initial type trim both fricative durations back from the boundary by 25%. [*North+sea*]
3. ***rule:*** If the first syllable is a {vowel, nasal, liquid}–final type and the second syllable is a {vowel, liquid}–initial type then trim each by three pitch cycles. [*easi+er* or *influ+ence, un+like, al+ready*] (Diphthongs are an exception here and there is no boundary trimming of either syllable in a diphthong + vowel or liquid sequence. [*dri+er, Ire+land*])

Rule 2 turns out to be one of the simpler rules. The new boundary created by the application of such a rule is amplitude normalised as part of the conjoining procedure. Amplitude normalising comes into play when any elements from the database or reconstituted elements are concatenated if the earlier normalisation process involved in building the database appears inadequate and amplitude conjunction is disjoint.

Rule 3 is very specific and perhaps applies only to this database. The normalised database is fairly uniform with respect to fundamental frequency – that is, the mean fundamental frequency varies minimally between items in the database. Three cycles in each syllable represents an appropriate overlap time for coproduction of these types. This is necessarily a compromise, since in waveform concatenation it is essential that conjoining should occur at like points on two joined waveforms – it is therefore not feasible to trim to a temporal fineness less than one period in duration. At a higher pitch four cycles may be more appropriate, whereas at a lower pitch two might be better. The accuracy is further compromised in **MeteoSPRUCE** by strategies for preserving micro-intonation effects. Three pitch cycles is therefore no more than a useful working value and makes no special theoretical claim.

CONCLUSION

Enlarging the waveform database of a concatenated waveform speech synthesis system is difficult, and can in the worst case involve re-recording the entire inventory. The **SPRUCE** family of systems share a high level general purpose synthesis engine, but have restricted domain individual low level inventories of waveform samples. We are using one application, **MeteoSPRUCE**, to investigate the feasibility of enlarging the database by recovering syllables from polysyllabic words and recombining them to form new words. We have identified the need for three levels of syllable model – phonological, phonetic and synthetic. The phonological syllable characterises a listener's perceptual response to a heard waveform, the phonetic syllable characterises a stretch of waveform in an utterance spoken by a human being and which triggers the corresponding phonological syllable in a listener, and the synthetic syllable characterises a waveform derived from a phonetic syllable and which is capable of manipulation by rule to trigger a similar and correct cognitive response in the listener. We believe that notwithstanding the traditional view that satisfactory recombination of segmented syllables is not possible because of coproduction and other effects, the approach we have adopted is capable of making a start toward providing in many cases a useful enlargement of synthesiser capabilities of sufficiently natural quality to make the results worthwhile and acceptable to listeners.

REFERENCES

[1] T. Dutoit (1997) *An Introduction to Text-to-Speech Synthesis*. Dordrecht:Kluwer Academic Publishers

[2] E. Lewis and M.A.A. Tatham (1991) **SPRUCE** - a new text-to-speech synthesis system. *Proceedings of Eurospeech '91*. Genova: ESCA

[3] J.P. van Hemert (1991) Automatic segmentation of speech. *IEEE Transactions on Speech Processing* 39:4, pp 1008-1012

[4] O. Boeffard, I. Miclet and S. White (1992) Automatic generation of optimized unit dictionaries for text-to-speech synthesis. *Proceedings fo the International Conference on Spoken Language Processing*, Banff, pp. 1211-1214

[5] S. Nakajima (1994) Automatic synthesis unit generation for English speech synthesis based on multi-layered context oriented clustering. *Speech Communication* 14, pp 313-324

[6] N. Campbell and A. Black (1995) Prosody and the selection of source units for concatenative synthesis. In *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive and J. Hirshberg. eds. New York: Springer Verlag

[7] A.J. Hunt and A. Black (1996) Unit selection in a concatenative speech synthesis system using a large speech dtabase. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Atlanta

[8] C. Gussenhoven and H. Jacobs (1998) *Understanding Phonology*. London: Arnold

[9] A.C. Gimson (1989) *An Introduction to the Pronunciation of English*. London:Arnold

[10] M. Tatham (1995) The supervision of speech production. In C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds.) *Levels in Speech Communication – Relations and Interactions*. Amsterdam: Elsevier, pp. 115–125

[11] C.A. Fowler, P. Rubin, R.E. Remez and M.T. Turvey (1980) Implications for speech production of a general theory of action. In B. Butterworth (ed.) *Language Production*. New York, NY: Academic Press, pp. 373-420

[12] J.A. Goldsmith (1989) *Autosegmental and metrical phonology: a New Synthesis.* Oxford: Blackwell

[13] J.A. Goldsmith (1995) *The handbook of phonological theory*. Cambridge MA: Blackwell

[14] D. O'Shaughnessy (1987) *Speech Communication – Human and Machine*. Reading, Mass.:Addison-Wesley

[15] C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds.) *Levels in Speech Communication – Relations and Interactions.* Amsterdam: Elsevier [various contributors]