SYLLABLE RECONSTRUCTION IN CONCATENATED WAVEFORM SPEECH SYNTHESIS

Mark Tatham*, Katherine Morton*, and Eric Lewis[†]
*University of Essex, UK, [†]University of Bristol, UK

ABSTRACT

In general purpose concatenated waveform synthesis an exhaustive stored waveforms inventory is needed. Our **SPRUCE** system is syllable and word based, but for general purpose work its inventory needs examples of all possible syllables. The high-level synthesis engine used to generate the phonology and prosody of utterances is already general purpose – but its use is constrained by small low-level inventories of re-combinable waveforms. The feasibility study reported here was carried out to determine whether we could take one of the word based limited domain versions of the system, and make it more general by excising syllables from existing polysyllabic words and recombining them into new words. Initially the study treats temporal rather than spectral considerations.

1. PRELIMINARIES

Concatenated waveform synthesis [1] uses an inventory of stored waveforms. This paper reports experiments in enlarging MeteoSPRUCE – a weather forecasting application of our general purpose high-level tts engine SPRUCE [2] to widen its usability without the need for re-recording [3] [4] [5] [6] [7].

Before embarking on the task of excising and recombining we needed to be clear on a number of basic theoretical points:

- Phonological symbolic representations [8] are of limited use for identifying syllables in the waveform. The phonological concept *boundary* carries uneasily through to the waveform.
- Phonetic representations [9] are also symbolic, and although we can identify an allophone string corresponding to a phonological syllable there is still often no clear feature for acoustically delimiting syllables.
- The notion *boundary* as a point for cutting a waveform is misleading. Acoustic syllables often overlap, telescope or merge, and 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 individual times.
- Coarticulation [10] or coproduction [10] [11] responsible for temporal overlap is also responsible for spectral overlap. Even if cuts are made at the 'right' places there is a problem of including spectral boundary effects from both syllables when they are recombined in new but 'wrong' contexts.

2. A SIMPLE EXAMPLE

The 2000-word **MeteoSPRUCE** database includes waveforms of the words *unsettled* and *likely*: let's try using these to create a new word *unlikely* – i.e. to detach the syllable *un* and place it in front of the *like* syllable of *likely*. Phonetic syllable boundaries are marked in the database morphemically if possible or phonologically. Fig.1 shows the database entries.

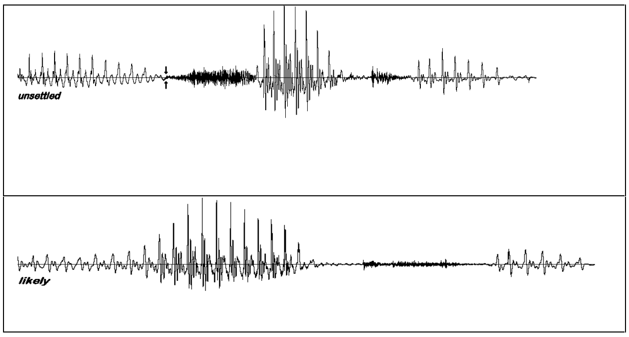

Fig.1 *unsettled* and *likely* in the **MeteoSPRUCE** database.

By cutting *unsettled* at the end of the last pitch period of *un* we can paste the beginning of the file to the start of *likely* to produce a new reconstructed word object **unlikely*. Fig.2 compares the result of conjoining the syllables with a recording of *unlikely* which on this occasion is in the database.

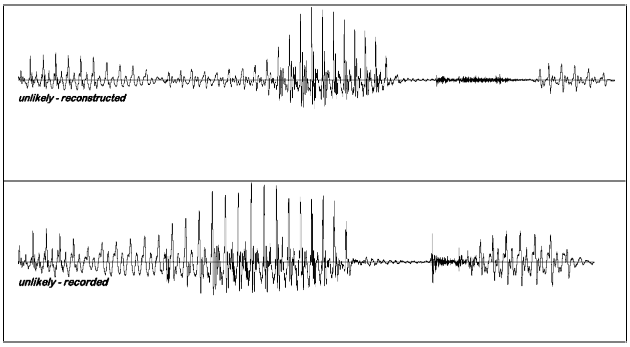

Fig.2 Reconstruction of **unlikely*, and the recorded waveform of *unlikely* in **MeteoSPRUCE.**

The degree of coproduction between syllables is context dependent – we deliberately picked the syllable *un* in *unsettled* because it showed the minimum of 'telescoping' coproduction.

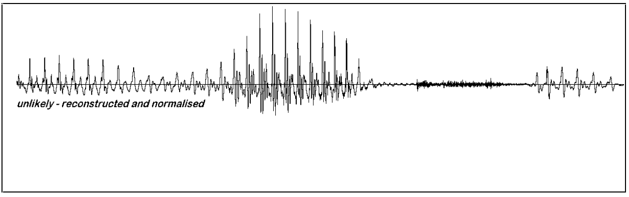

Fig.3 Reconstruction of **unlikely* using the derived synthetic syllable *un* and the recorded word *likely* (also normalised at the beginning of the word to form the synthetic syllable *like*).

So far, we have identified three stages in the reconstruction procedure: a. phonetic syllable excision, b. normalisation, c. synthetic syllable conjoining. There are errors in the reconstruction, and the transition between the syllables *un* and *like* appears protracted and awkwardly joined. An improvement (Fig.3) is obtained by a normalising procedure dealing with syllable overlap. The procedure involves setting up a **synthetic syllable**, derived in the normalisation process from the **phonetic syllable.**

3. IDENTIFYING AND DESCRIBING SYLLABLES

To clarify the concept of **recovery:** it may be possible to excise a stretch of waveform of the right length from a suitable word, but because of coproduction effects it is unlikely to be directly reusable except in a similar context. Recovery means excision *and* **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.

Phonological syllable – a unit higher than the 'sound' segment [12]. Introduced to form a framework for characterising the sequencing of simple segments, it provides the primary unit for modelling prosody. Phonetic detail is irrelevant at this level: nonlinear organisation into syllabic units is important. We characterise phonological syllables as in linguistics [13].

In our model the phonological syllable figures prominently because it enables direct reference to a listener's perception of '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 this.

So, listeners identify a unit at the beginning of *unsettled*, pronounce it in isolation and tell us that it is the *same* 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 different acoustically. The goal of the reconstruction procedure is to use a portion of the waveform of *unsettled* to change *likely* into a correctly *perceived* new word *unlikely*.

Phonetic syllable – a descriptive unit characterising part of a *human* acoustic signal prompting a listener to identify a *phonological* syllable. This is where distinguishing acoustic features are identified, as well as other acoustic features. The model describes the waveform as in acoustic phonetics [14].

What 'sounds' are sequenced in a phonetic syllable is a phonological rather than phonetic matter in our reconstruction procedure. The phonetic syllable is the waveform which triggers the phonological syllable – *and* its phonetic description.

There has been a lot of discussion concerning the relationship between phonetic and phonological characterisations of the same stretch of speech [15]. The phonetic syllable models the acoustic signal and the phonological syllable models a cognitive *response* to the signal. The models are linked since 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 syllable – a model of an acoustic stretch which can be manipulated to trigger in the listener a response of the right phonological syllable. The synthetic syllable may or may not be the same as the phonetic syllable from which it is derived.

In **SPRUCE** a waveform in the database can be a phonetic syllable (modelling the human syllable, e.g. *snow*), but it is also there as a synthetic syllable – a model for concatenation to produce a new word, e.g. *snowing*. The synthetic syllable derives 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 is both context and *type* sensitive.

4. SYLLABLE TYPES AND CONTEXTS

We classify syllable types by their phonological start (onset) and end (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, taking full account of all acoustic effects of quality change resulting from coproduction all combinations would need to be considered. For this initial study we reduced the problem to a working model of *temporal* syllable combining. Defocusing phonetic quality at syllable boundaries, we refocused on temporal properties of onset and offset.

Examination of all words in the database revealed that our working model might need deal only in initial and final segment *types*, rather than all possible occurring individual segments. We established segment types according to the usual phonetic parameters [4]. So, all syllables include a vowel segment preceded by up to three phonetic consonants and followed by up to four:

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

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

- vowels (including initial semivowels and [h]) – *all (you, how), me* [initial, final]
- diphthongs – *air, dry* [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:

1. We chose monosyllabic words here because the recording and normalisation procedures eliminate initial and final coarticulatory effects in words. Syllables not in the database as monosyllabic words are excised from words which *are* in the database: here coarticulatory effects *are* present.
2. Vowels and diphthongs are entered as different types because diphthongs appear to be more resistant to coproduction trimming or truncation.
3. Plosives are separated into initial and final – the *stop phase*

is important in initial position, and the *burst phase* in final position. But in finals we did not find it necessary to distinguish between voiced and voiceless plosives – for the speaker who made the recordings (author MT, Southeast England accent) there was no big difference in the bursts.

4. All nasals appear to behave similarly and likewise all liquids.

5. TAKING THE *UN* EXAMPLE FURTHER

Table I [at the end of the paper] shows data from examples in the database with and without *un*. Of the five combination types, the one with least overlapping is the + *voiceless fricative* type. [s] has a similar duration whether prefixed or not. Initial voiceless plosive are halved in duration, and initial voiced plosives truncated even more. Initial nasals are unaffected, retaining full duration (observation based on other words beginning with [n]); while initial liquids are nearly halved. Truncation is an acoustic effect of coproduction – probably more an overlapping or telescoping effect. Careful listening to the prefixed examples enables detection of co-articulatory phenomena – e.g. frication can be detected during the last two pitch cycles of [n] in *uncertain*.

Despite limited data we thought it worth seeing if we could use Table I results to predict the behaviour of other segments in like environments. We do *not* present this as a valid generalisation, but to illustrate *procedure*. Table II shows the results. In one example of each type we could predict changes from prefixing *un* to words with initial voiceless fricative, voiceless plosive or liquid. The result for an initial voiced plosive was disappointing, but could be explained by a segmentation measurement problem: it was difficult to differentiate between the ending of the nasal in the prefix and some vocal cord vibration possibly 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 usually 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.

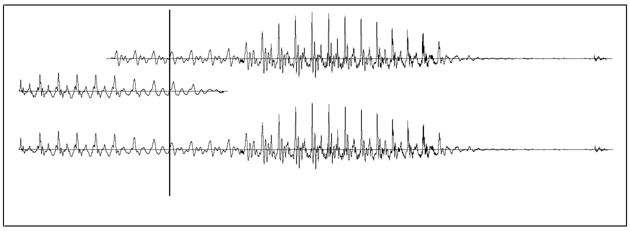

Fig. 4 – Example of the excision and reconstruction procedure.

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; trimmed syllables conjoin to form the new word *unlike*.

There are several negative points to note in this procedure:

- *un* in *unsettled* has slight coproduction ([s]-derived signal is clearly visible toward the end of the excised waveform);
- the reconstituted word has *conjoining* not coproduction – 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 – not a problem in the normalised **MeteoSPRUCE**;
- the stress pattern of the new word must match the original stress values of the excised syllables – e.g. the secondary stressed *un* may not be satisfactory for reconstituting the word *under* beginning with a primary stressed *un*.
- if there is a change of *f0* at the boundary it needs neutralising. *f0* mismatches are normalised in the intonation algorithm applied when the word is used in a synthesised utterance.

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

6. RE-COMBINING RULES

We now begin determining how the different syllable types pattern when linearly combined. Again this is a first approximation – our objective is still to determine how far we get in triggering the acceptance of appropriate phonological syllable *combinations* in the listener with the simplest model. It might later 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 database were examined to locate linearly sequenced boundary effects. Still ignoring qualitative coarticulatory effercts, we set out to examine the temporal effects of coproduction or overlap.

We determined that small effects occurred with:

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

rule 1: There are no adjustments to be made where the boundary is preceded by {vowel, plosive, nasal, liquid}-final types, and followed by {fricative}-initial type. If the plosive and fricative are homorganic the burst is trimmed away.

rule 2: Where the boundary is preceded by {fricative}-final type and followed {fricative}-initial type trim both fricative durations back from the boundary by 25%, [*North+sea*].

rule 3: 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 is one of the simpler ones. The boundary is normalised for amplitude mismatch in the conjoining procedure. Amplitude normalising comes into play when elements from the database or reconstituted elements are concatenated if the earlier database normalisation process appears inadequate.

Rule 3 is quite specific, applying only to this data. The normalised database is fairly uniform with respect to *f0* – the mean *f0* varies minimally. Three cycles of each syllable is a workable overlap for coproduction of these types. But this is a compromise, since in waveform concatenation it is essential that conjoining should occur at like points – so it is not feasible to trim to a temporal fineness less than one period; this varies with pitch. Three pitch cycles is just a useful working value with no special claims.

7. CONCLUSION

The **SPRUCE** family of tts systems shares a high-level general purpose engine, but has restricted domain individual low-level inventories of waveform samples. We have used one application, **MeteoSPRUCE**, to investigate the feasibility of enlarging the database by recovering syllables from polysyllabic words and recombining them to form new ones. We identify the need for three levels of syllable model – phonological, phonetic, synthetic. The phonological syllable models the perceptual response to a waveform, the phonetic syllable a stretch of waveform spoken by a human speaker and triggers the corresponding phonological syllable in a listener, and the synthetic syllable a waveform derived from a phonetic syllable, capable of manipulation by rule to trigger a similar and correct cognitive response. We believe that our listener-oriented approach can be developed to provide a useful enhancement of synthesiser capabilities.

REFERENCES

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

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

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

[4] Boeffard, O., Miclet, I. and White, S. 1992. Automatic generation of optimized unit dictionaries for text-to-speech synthesis. *Proc. of the Int. Conf. on Spoken Language Processing*, Banff, 1211-1214

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

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

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

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

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

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

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

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

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

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

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

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.

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).