ON FINDING OBJECTS IN SPEECH SIGNALS

A J H Simons

Department of Computer Science,
University of Sheffield, Sheffield S10 2TN

ABSTRACT

We present a technique for extracting information from signal traces that takes into account the shape characteristics of events in the signal, as much as the values taken by the signal. The technique derives statistically motivated object-descriptions which are subsequently open to expert human inspection. The approach aims to provide knowledge-based speech decoding strategies with a sound basis; or conversely to provide statistical techniques with more transparent and flexible representation units.

1. INTRODUCTION TO HYBRID RECOGNISERS

Statistical pattern-matching and knowledge-based techniques for acoustic-phonetic decoding lie at opposite ends of a spectrum of possible hybrid approaches, which has largely remained unexplored. Allerhand's work [1] is one hybrid example deserving attention. A present trend is to automate the threshold settings of knowledge-based algorithms [2] or to predicate the variability of stochastic classifiers over models whose internal structure has been influenced to some extent by *a priori* speech knowledge [3]. The large gap remaining between these approaches reflects the considerable difficulty in finding a common ground which is both representationally sound (preserving what a phonetician views as important) and stochastically admissible (preserving the virtue of optimal training procedures).

In outline, our hybrid approach seeks to develop a hierarchical model which preserves stochastic admissibility through all levels of representation. Our starting point is a set of parameter traces. A parameter trace is a time-varying signal, usually computed from a pressure waveform, that makes some aspect of speech explicit, such as its noisiness, or overall energy. Our premise is that salient *acoustic events* or *acoustic objects* may be detected in these traces, that fall into distinct classes governed by the phonology of English. The first stage of our task is to determine the optimal set of acoustic events that adequately describe the behaviour of given parameter traces. The second stage is to relate these events to a syllable model of English phonology. The rest of this paper describes progress made in the first of these two stages.

Our model for extracting information is based on a straightforward physical description of a trace in terms of its magnitude, rate of change and non-stationarity. A finite state machine is used to construct a probabilistic matrix of most likely topological event sequences, from which candidate object templates (isomorphic shapes) may be selected for further investigation. Deterministic algorithms are then devised for extracting instances of each object, in order of greatest significance. Each object is a parameterised description of the magnitude behaviour of the trace over some extended time interval. Various numerical features and distances may be measured in each object, from an object-centred perspective. Objects are clustered on the basis of these features, until an optimal class separation is achieved with respect to a distance metric.

2. DETECTING OBJECTS IN SPEECH PARAMETER TRACES

The human eye and mind are quite expert at organising visual information, such as that presented by a speech signal trace. For example, a good subset of all syllable nuclei are readily estimated from peaks observable in signal amplitude traces or loudness functions [4]; strident fricatives and noise bursts associated with strident plosives are readily detectable from plateaux and spikes visible in high frequency bandpass energy traces.

Intuitively, events such as peaks, dips and shoulders in traces are seen as significant objects because of the Gestalt properties [5] of the geometric forms they describe. A peak or a plateau, a dip or a valley may somehow be viewed as an 'excursion' against a background level which is subsequently 'completed' by the returning trace. Small perturbations are not significant where the overall shape is undisturbed; however there are clearly trading relations between competing interpretations of some shapes. Notions of *similarity*, *familiarity*, *closure*, *belongingness*, *good continuation* and *common fate* are obvious candidate principles from Gestalt psychology to apply to this phenomenon.

The problem of detecting robust classes of object in parameter traces is essentially one of available knowledge. If, on the one hand, you know what the answer categories are, then you can optimise over the training data to distinguish between these classes (cf factor analysis). This is how current pattern-classification approaches to speech recognition work. If, on the other hand, you know what the salient features in your signal are, then you can group these by distance metrics into classes of object (cf cluster analysis). This is one of the approaches adopted by the knowledge-based school. We are starting from a position where we do not wish to prejudge the answer categories or the feature set to be measured.

3. DESCRIBING THE BEHAVIOUR OF PARAMETER TRACES

We adopted a model which views a trace simply as a time-varying magnitude signal. (Different traces measure their ordinates in different scales, such as amplitude, frequency, zero crossing rate per 10ms; we generalise this to a notion of magnitude). We consider that the most important indicators of a trace's behaviour over time are its first and second derivatives. This is in keeping with our current consideration of traces as simple physical phenomena. Also, we compared our derivative-based scheme against several segmentation methods, including piecewise constant and piecewise linear regression [6], hierarchical [1] and curvature-based segment-fitting methods [7]; and an original clustering scheme. We preferred the derivative-based scheme for its greater robustness across scales and greater sensitivity in locating important acoustic boundaries.

The first derivative is the rate of change of the trace with respect to time. Extrema in the first derivative correspond to points at which the trace is moving most rapidly. The trace was most likely to be segmented at these points by those of our previous techniques which grouped or clustered on the basis of the trace's magnitude. It is cheaper to pick extrema here than to cluster the trace into similar regions in order to find the boundaries.

The second derivative is a measure of the non-stationarity of the trace over time. Extrema in the second derivative correspond to points at which the trace commences, or terminates an excursion. The trace was most likely to be segmented at these points by those of our previous techniques which fitted straight line approximations to the signal, especially those techniques using interpolation error rather than least squares error. The former emphasises the error at breakpoints, whereas the latter emphasises the error over whole segments. It is cheaper to pick extrema here than to fit line segments to the trace in order to find breakpoints.

The first and second derivatives of traces were calculated using a standard method. A good approximation to the derivatives of an empirical signal may be achieved by convolving the signal with the analytic derivatives of the Gaussian. The formulae used were:

$$Deriv1(i) = -G'(s,n) * T(i)$$

$$Deriv2(i) = G''(s,n) * T(i)$$

where * denotes convolution, *G'(s,n) and G''(s,n)* are the first and second derivatives of a Gaussian curve of standard deviation *s* sampled over *n* points, *T(i)* is the *i*th sample of trace *T*. In practice, the size of *n* was determined by the size of *s*, the standard deviation (typically 2 - 4 samples). We adopted a convention of sampling over $-3s$ to $+3s$ (although less would probably have been adequate); ie $n = 6*s + 1$ points. This corresponds to sampling the Gaussian over approximately 99.9% of its area.

The positive and negative extrema of the trace, its first and second derivatives were then extracted using a peak-picking algorithm which discarded a fraction of low-scoring peaks (typically, these were points that failed to score as maxima or minima for more than 25% of the peak-picking window's passage). A label was generated for each extremum, at which point the trace and its derivatives were sampled, producing a data triple. (These *extremum events* were considered key points in the evolution of the behaviour of the trace). The size of the peak-picking window was directly related to the scale of the Gaussian convolution operator.

4. FINITE STATE MACHINES FOR LEARNING OBJECT TOPOLOGY

Our approach describes the commonality of shape shared by different trace events, rather than relying on simple threshold [8], or excursion-based [9] measurements. Furthermore, we select significant classes of shape from an automatic examination of the data. Initially, we model the sequence of extremum events in a finite state network, calculating the *n* most probable paths through this network, then filtering the resulting extrema sequences. A single sequence of extremum labels describes a set of topologically identical objects.

Two networks were constructed, corresponding to bi-gram and tri-gram models of conditioning. The bi-gram model assumed each state in sequence was conditioned by the previous one alone and contained one node for each extremum type from the set:

$$Nodes = \{sigmax, sigmin, dx1max, dx1min, dx2max, dx2min\}$$

The tri-gram model assumed that each state in sequence was conditioned by the previous two states. This approach factored out the contextual occurrences of second derivative extrema, discovered using the first model. Nodes were now drawn from the set:

$$Nodes = \{dx2max\text{-}dx1max, dx2max\text{-}dx2min, \dots sigmax\text{-}dx2min\}$$

Constructed in this way, the network had a theoretical possible maximum $6^2 = 36$ nodes (ie the cross product of all six extremum labels), but in practice only half the nodes from the cross product set actually occurred in the data. This was already good evidence of the strong constraint provided by immediate context.

Each network was built simply by recording transitions from one node to another, as observed during multiple passes through a set of data. Note that, at this point, only the qualitative shape information was being used, not the numerical values associated with each data triple.

The network was then explored using a beam searching technique based on a branch-and-bound algorithm with residual pruning. The algorithm constructed a queue of partial paths through the network, starting (on each pass) from one named node and aiming to finish at another named node. The partial paths were sorted according to the total probability of traversing the path; in this sense, the solution paths were returned in order of maximum likelihood. All paths over a residual likelihood (by default, 1%) were returned, but this limit could be raised dynamically (by 0.5%) in response to imminent memory exhaustion. Partial paths scoring less than the limit were pruned from the queue. Solution paths were maintained on the queue so that they might be further extended.

The results from this processing were extremely satisfactory from the point of view that they both confirmed intuitive models of shape behaviour and revealed aspects of behaviour that were not immediately apparent to a human observer. The results can be summarised by stating that all the variants of expected topologies were eventually generated, but that simpler descriptions of parts of these were found to occur more frequently. The following results are typical of 'positive excursion' traces:

- The greatest number of extrema arise (by definition) from the second derivative signal. The small increase in minima over maxima can be explained by the fact that trace onset behaviour is typically more abrupt (rising in one step) than offset behaviour (descending in multiple steps).
- The most probable event sequences delimited by maxima in the second derivative (ie excursion breakpoints) are therefore peak-offsets. These are portions from the last rise before the signal maximum, then the complete fall encompassing a first derivative minimum up to the arrest of this fall.

- Various canonical peaks and dips feature in the middle range of probabilities. Some of these are very similar, all but for the omission, insertion or substitution of one extremum label.

- Unexpected topologies also feature in the middle range; eg the quiescent portion between two positive excursions of the trace, or stepped onsets ('shoulders'). These novel descriptions reflect behavioural characteristics of the type of trace on which the network was trained (eg amplitude).

- Multiple-peaked topologies feature in the low-scoring range. This is to be expected, considering the natural bias against constructing longer paths through the network.

The notion of the *probability*, or *likelihood* of each topological sequence needs elucidating. This does not represent the probability of such an object *occurring* in the data; rather it represents the likelihood of labelling any such sequence of data with a model of that *topological complexity*, given no information about the kinds of shape which are deemed interesting. As we saw, simple models are naturally preferred over more complex ones. The network embodies the Gestalt constraint of *similarity*, in a topological sense.

5. DETERMINISTIC ALGORITHMS FOR DETECTING OBJECTS

It is possible, by virtue of the connections existing in the network, to generate longer sequences which never actually occur in the data. All generated topologies were therefore filtered according to likelihood of occurrence of each event sequence in the data. This reduced topological 'noise' by 50-70%. Canonical, novel and multi-peaked shapes scored highly on this pass. Thus, information was gained about:

a) the most reasonable descriptions of shape;
b) the most likely shapes occurring in the data.

The next stage was to cluster examples of matched shapes in the data. A dynamic programming algorithm clustered similar topologies (at increasing cost); while a euclidean metric based on the magnitude and duration of excursions was used to split the classes so derived. A problem encountered here was the inability to equate the cost of an insertion, deletion or substitution in the DP algorithm with the distance metric of the clusterer.

For this reason, we eventually decided to introduce some stronger assumptions into our model. Given that speech is a sequence of events of some kind, 'complete' events should be determinable from the overall behaviour of the trace's first derivative. However, the second derivative marks the start and end of excursion points. On this basis, we constructed a deterministic algorithm that matched trace onsets against offsets, allowing the most complete wholes (the Gestalt constraint of *closure*) to influence our segmentation of the trace into a hierarchy of events.

The algorithm is a function of two parameters: an absolute constraint ratio linking maxima and minima in the first derivative and a decaying window function, within which a match must be found. The first parameter represents how strong an offset you would consider sufficient to complete the event started by the onset. A figure of 50% (a maximum entropy decision) was found to agree with visual judgements for detecting positive excursion events. The second parameter represents how long you are prepared to wait for an event to be completed. Various simple and complicated functions that decayed in around 300 ms (the longest expected acoustic event) were used successfully. The algorithm may be used to match offsets to onsets in a forward pass, or onsets to offsets in a backward pass. In forward mode, it may be described approximately as follows:

```
Select the next largest dx1max at dx1(i);
    Start := i;
    End := i;
    Onset := dx1(i);
    Limit := - ratio * Onset;
    Offset := Limit;
While (dx1(i) > win(i,Limit)) and (win(i,Limit) < 0) do
    Select the next dx1min at dx1(i+k);
        If dx1(i+k) < dx1(i) then
            End := i+k;
            Offset := dx1(i+k);
        i := i+k;
Record matching onset and offset from Start to End;
    Duration := (End - Start);
    Closure := f(Onset, Offset).
```

The algorithm detects complete peak- or dip-events in traces. It is not confused by small perturbations, which give multiple peaks, where the overall shape is robust (*closure* proves to be a more sensitive constraint than just modelling topology). It degrades gracefully, by returning events in order of greatest prominence. It is hierarchical, in that multiple onsets may eventually be matched against one offset in forward mode, vice-versa in backward mode. A further process detaches peak/dip figures from the background in order of greatest closure, leaving shoulder events, corresponding to stepped onsets or offsets, and quiescent events. This models the Gestalt constraint of *belongingness* (unique association of components to wholes). Excursion-points are used to delimit object boundaries.

6. CLUSTERING ALGORITHMS FOR DETERMINING CLASSES OF OBJECT

Numerical features appropriate to each topological type of object are measured, for complete sets of objects. These are submitted to a standard clustering algorithm. The aim here is to obtain multiple classes of acoustic event, where each class is determined by its internal consistency and differs sufficiently from other classes. Clustering embodies the Gestalt constraint of *similarity* in a quantitative sense, allowing many small factors to contribute to the whole. We have concentrated mainly on clustering peak events in amplitude and zero-crossing traces. Peak features were chosen at random from the set:

PeakFeatures = {duration, area, log_area, max_mag, rel_mag, avg_mag, onset, offset, closure}

and clusters were evaluated with respect to a fine acoustic hand-labelling of events in traces. The most reasonable clusters are obtained using features that characterise the extremes of behaviour, eg {duration, max_mag, onset}, rather than those that tend to average the qualitative shape of peaks.

Two variants of the clustering algorithm were tested. The first, standard version, normalised distances in each dimension with respect to the total sample variance for that dimension; and used normalised euclidean distance as its cumulative cost-function. The second version normalised distances with respect to a pooled estimate of the variances of the two classes about to be merged. The increase in actual variance over the estimate was used as an indicator of the increase in entropy caused by merging the classes.

We are currently considering two methods for halting the clustering: one uses heuristics to pick the first sharp increase in the cost-function [1] and the other evaluates the cluster-tree against the fine acoustic labels to obtain the best global class-to-label correspondence.

7. EFFECTIVENESS OF THE METHOD

We have described a method for detecting objects in parameter traces, using weak assumptions. In an example run, the processes described here extracted separate statistical class models of peak-events corresponding to syllable-initial /s/, aspirated /t/, strongly-aspirated /t/, syllable-final /sh/, utterance-final /sh/, weak-fricative and voice-peak from a simple zero-crossing trace. In a test of our ability to replicate human perception, two subjects were asked to choose five events in order of 'apparent significance' from the same trace. The subjects and the object-detection algorithm chose the same five events; in one case the ordering was different for the algorithm. The subjects weighted their choice by the area under the trace, whereas the algorithm weighted its choice by abruptness of onset.

8. REFERENCES

[1] M H Allerhand (1986), 'A knowledge-based approach to speech pattern recognition', PhD Dissertation, Darwin College, University of Cambridge.

[2] V Zue, J Glass, M Phillips and S Seneff (1989), 'Acoustic segmentation and phonetic classification in the SUMMIT system', Proc. IEEE ICASSP, Glasgow, S8.1, 389-392.

[3] K F Lee, H W Hon, M Y Hwang, S Mahajan and R Reddy (1989), 'The SPHINX speech recognition system', Proc. IEEE ICASSP, Glasgow, S9.3, 445-448.

[4] P Mermelstein (1975), 'Automatic segmentation of speech into syllabic units', JASA, 58 (4), 880-883.

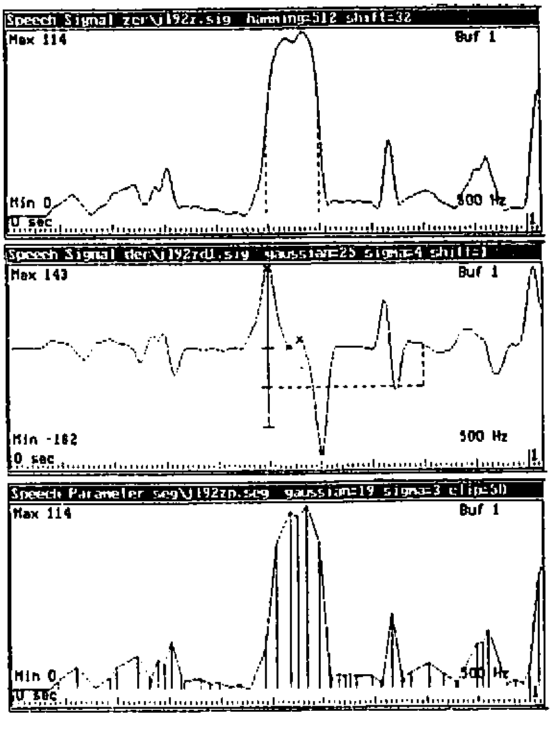
[5] K Koffka (1936), *Principles of Gestalt Psychology*, Kegan Paul, London.

[6] J S Bridle and N C Sedgewick (1977), 'A method for segmenting acoustic patterns, with applications to automatic speech recognition', Proc. IEEE ICASSP, Hartford Conneticut.

[7] P Green, M Cooke, H Lafferty and A Simons (1987), 'A speech recognition strategy based on making acoustic evidence and phonetic knowledge explicit', Proc. European Conf. on Speech Tech., Vol. 1, 373-376.

[8] B Williams, S Hiller, F McInnes and J Dalby (1989), 'A knowledge-based nasal classifier for use in continuous speech recognition', Proc. European Conf. on Speech Comm. and Tech., Paris, Vol. 2, 252-255.

[9] P Green, M Cooke, M Crawford and A Simons (1988), 'Acoustic-phonetic reasoning with the speech sketch: a progress report', Proc. 7th Symp. Fed. Acoust. Soc. Europe, eds. W A Ainsworth and J N Holmes, Institute of Acoustics, 353-360.



Figures: parameter trace; object detection using a rectangular window; boundary segmentation at excursion points.

Preliminaries to a New Text-to-Speech Synthesis System

Marcel Tatham

University of Essex, Colchester

Introduction

This paper describes some of the thinking behind the design of a new text-to-speech synthesis system, and informally discusses decisions based on pilot studies for the system. The system concerned is the focus of the recently begun SPRUCE[1] Project, the outcome of which is to be a demonstrator of text-to-speech synthesis whose output is more natural sounding than anything so far achieved on a consistent basis. The scheme incorporates several innovations, some of which are mentioned below, but also careful reworkings of earlier ideas.[2]

Various pilot studies of aspects of the design of the system have been carried out over the past two or three years in the Advanced Speech Technology Laboratory at Essex University. The task facing SPRUCE over the next few years is to put together existing pieces of the experimental scheme, and integrate them into a robust whole. Here I shall be discussing some of the aims of the Project, and how these might satisfy the needs that have arisen as the demand for voice output devices grows. A major point will be the basic philosophy of the system, and how some of the shortcomings of existing text-to-speech synthesisers will be met.

Current text-to-speech synthesis

Current text-to-speech synthesis systems are nearing the theoretical limits of their ability to produce high quality speech output. A strong commonality of feeling exists among researchers which is expressed in the frustration they share that text-to-speech synthesis is almost good enough to satisfy the demands repeatedly made over the past two decades or so for acceptable voice output devices, but not quite good enough to result in any significant uptake by potential users. We have here a classical case of a miss being as good as a mile – almost good enough is not good enough. For the researcher the results are scalar and each small improvement is seen as significant, but for the user or the marketplace the situation is binary: text-to-speech synthesis either works or it does not. The sad story is that it does not. For SPRUCE to be a worthwhile project and not just a small refinement of some existing system there had to be serious consideration of why results so far had not been entirely successful.

Available text-to-speech synthesis systems have a number of features which contribute toward determining the characteristic quality of the final output. Interestingly, all sound different, yet at

[1] The SPRUCE academic collaborators are Essex University (Marcel Tatham and Katherine Morton), Bristol University (Eric Lewis and Rodney Sampson), Liverpool University (Colin Goodyear and Barrie Cheetham).

[2] For a discussion of the architypal text-to-speech system see Holmes 1988, Chapter 6.

the same time are all recognisably similar: you know when you are listening to a text-to-speech system, as opposed, say, to a resynthesis system or a compressed/coded recording.

0. General philosophy

The general idea of text-to-speech system is to produce a loose working model of a human being reading text out aloud. The only part of the human process that is (usually) bypassed is the visual input: the text is normally either input from a keyboard or from some text file. If the model is serious it cannot help but make claims about the human process, some of which are hard to justify and others of which are quite simply wrong.[3]

In the model which supports synthesis human speech is assumed to be the result of the concatenation of idealised segments of speech, each with its own intrinsic duration, which, on demand from the brain, are strung together in the peripheral vocal mechanism. Because the system is time-governed and because the neuro-muscular mechanism is being pushed to its temporal limits the individual sounds are assumed to be degraded as they are conjoined (the process known in phonetic theory as **coarticulation**).

The underlying model represents each of the segments using a number of articulatory parameters each of which initially has independent control. This independent control is then constrained at the physical level by what combinations of parameter values are possible and what are not. In addition at a higher level it is also constrained as to what combinations are cognitively possible in any one particular language or dialect. Since, in general, text-to-speech systems are acoustically based the articulatory specification of segments is translated into an acoustic parametric specification.

Additionally, whereas in the phonetic model the constraints on parameter value combinations-- are made explicit, in acoustically based synthesis they are not: it is simply noted that for a given segment a particular combination of parameter values is the required specification to encode the sound which results from the articulatory combination.

Above the vocal mechanism the model does not deal with physical systems at all, but with cognitive units and the processes in which they play a role. Thus it is assumed that the individual physical sounds have direct abstract cognitive correlates. This abstract level in the model corresponds to phonology in linguistic descriptions of speech production. In linguistics the phonological model on which text-to-speech systems are based takes as its primitives features (parameters) which are combined to specify individual segments of speech. Decisions regarding modifications of these segments take the form of production rules operating on the features of segments.

[3] An example of the first type of claim is that speech is composed of small segments strung together; an example of the second type is that the physical parameters of speaking (articulatory or acoustic) are independently controllable.

Prosodic features such as stress and intonation are assigned at this abstract level by rule. They depend to a large extent on the categories of syntactic elements (words) and the overall syntactic and phonological structure of sentences.

1. Input

All systems accept plain text as input, although there is sometimes the possibility of adding markers to assist later stages of the system. An example of this is the addition of symbols indicating the selection of basic intonation contours, or the position of nuclear stress in the input sentence. The markers usually assist the generation of prosodic information either by relieving the system of a computational stage, or by flagging exceptions or additions to the default procedures.

All systems currently respond to mis-spelled input text; that is, no system detects and corrects spelling errors. This omission is often used to improve pronunciation by deliberately mis-spelling words, thus fooling the system into what for it would be an incorrect pronunciation but which to the listener sounds better than the default effort. The reason for this need often lies in a system's inability to deal adequately with the text in the conversion from orthographic representation to a phonological (phonemic) or phonetic representation.

2. Orthography to phoneme conversion

Acting on the idea in the underlying model that an initial stage of the reading aloud process consists of converting strings of orthographic symbols into strings of phonemic symbols, all text-to-speech systems incorporate orthography to phoneme conversion almost always achieved by the application of spelling rules. These recognise orthographic symbols or particular groupings of symbols and interpret them within their immediate context to output a corresponding string of phonemic symbols. Orthography to phoneme conversion rules are more or less successful for a large percentage of words in English, but a quite significant number of words cannot be handled by rule and need to be treated as exceptions held in a list. This is an unrealistic way of modelling the human reading process.

Systems vary as to how many difficult words are assembled in the exceptions dictionary. The number can be a low as 100 or as high as a few thousand: it depends on the sophistication of the rule set. The strategy for employing the combined dictionary and rule system is to proceed by searching firstly for a word in the incoming text in the dictionary: if it is there its phonemic representation is retrieved; if it is not there then the word is passed to the orthography to phoneme conversion rules to generate an appropriate phonemic representation.

3. Prosodics

The underlying model deals with prosodics within phonology. Stress is handled, for example, by assigning a stress pattern to individual words and then by assigning sentence stress. Intonation is also treated largely as a phonological phenomenon. Our ideas about stress (Fudge 1984) and intonation (Pierrehumbert 1981) have changed considerably during the past twenty years or so: the recent views are only just beginning to filter into speech technology. In current text-to-speech suprasegmental effects vary considerably in their success. This is undoubtedly one of the key areas where the usual systems fail – poor prosodic rendering is perceptually very noticeable.

Difficulties arise from the fact that prosodic information is not encoded in the orthographic representation, other than very crudely as punctuation marking sentence and phrase boundaries. It needs to be generated within the system itself, with little in the input to work on. The best systems perform a syntactic parse on the text as the basis for a set of rules to generate appropriate stress and intonation contours. The worst systems attempt to assign stress using only phonological information (how the phonemes have been strung together), and only the punctuation markers to select from a small set of stored canonical contours representing the intonation patterns. Because the supporting information is minimal there is plenty of scope for error.

Similarly other aspects of the prosodics are handled more or less successfully. Thus, for example, durational variations of individual segments can be derived by rules which examine its phonological context and the general intonation contour of the sentence. Additionally, overall rate of utterance can be adjusted for the entire text, but to achieve varying rates for portions of the text usually means resorting to markers entered by hand in the text itself. For English, rhythm is usually determined by the assertion in the model that stressed syllables occur isochronically.

4. Segments

The linguistic model on which synthesis is based takes the phonemic string and converts it by rule to a phonetic string. This is done by examining the segmental context of each element and determining whether it should be modified. In synthesis systems the modification takes the form of a rewrite of the entire phonemic symbol as an entire phonetic symbol. Unlike linguistics synthesis systems do not normally recognise that the segments are more effectively described in terms of their component features. The reason for this departure from the model is partly historical (early synthesis systems did not use feature-oriented phonology as their basis), and partly to avoid the double conversion from phoneme symbol to phonological features and then to acoustic parameters. A linguist would argue that important information is thus lost because there is no demonstrable one-to-one relationship between phonemes and acoustic parameters. This is potentially a source of serious error in the system. When errors do occur they are fudged by spurious increase in the number of entries in the allophones inventory. An example of this is the need to have separate allophonic representations for plosives with and without aspiration or with and without a release phase.

The string of phonetic or allophonic segments (essentially an abstract representation of how the text is to be spoken) is converted into numerical representations for the purposes of driving the synthesiser engine. Each sound is individually represented according to its parametric specification (where the parameters are those dictated by the synthesiser) and where a sound corresponds to a segment of the underling linguistic model. I am deliberately avoiding discussion of diphone based systems (see Holmes 1988, p. 77) since this notion has no foundation in linguistics.The segment specification is accompanied by its so-called intrinsic duration – its canonical duration, subsequently modified by rules which adjust it according to segmental and prosodic context.

A characteristic of this model is the need to provide smooth transitions between the segments. The advantage is that it is only necessary to specify as many segments as are found to be needed for a rough allophonic representation of the output (generally between 60 and 150). The disadvantage of the model is that the calculated transitions do not always sound natural.

5. Segment conjoining

Segment conjoining is accomplished by means of a set of transition rules. These vary from system to system, but no system is entirely satisfactory. One problem is that the computational load is increased markedly if the different parameters are treated by individual rules sets (which is necessary if more natural sounding speech is required). Experiments have shown that conjoining of the formant amplitude parameters is more critical perceptually than that of the formant frequency parameters. In the model of human speech describes coarticulation of individual articulatory segments; one-to-one correlation between coarticulation and conjoining of acoustic parameters has not been shown, though it may well be a fair approximation given the above comments about allophone inventories.

6. Segmental and supra-segmental fitting

Merging the segmental representation with the prosodic contour comes at different points in different systems. All are based on an underlying model which assumes that segmental and supra-segmental systems are essentially separate, and all therefore require the prosodic contour to be fitted to the segment string. The level of difficulty here correlates with the degree of sophistication of the prosodic contour itself, which in turn depends on the syntactic complexity of the sentence in question, and whether or not the system has any syntactic information available.

The SPRUCE Proposals

SPRUCE addresses each of the above characteristics of current text-to-speech systems on the assumption that the each is a potential source of error leading to loss of naturalness.

0. General philosophy

It is important to have a clear understanding of exactly what it is that we are trying to simulate in text-to-speech systems – in particular what constitutes the naturalness that we can so clearly, as human beings, detect to be lacking in contemporary synthesis. But quite apart from the fact that our model of human speech may not in fact have been appropriate, a description of just exactly what constitutes naturalness still eludes us. In SPRUCE the problem is sidestepped to some extent by incorporating a subsystem that is effectively a resynthesis of actual human speech.

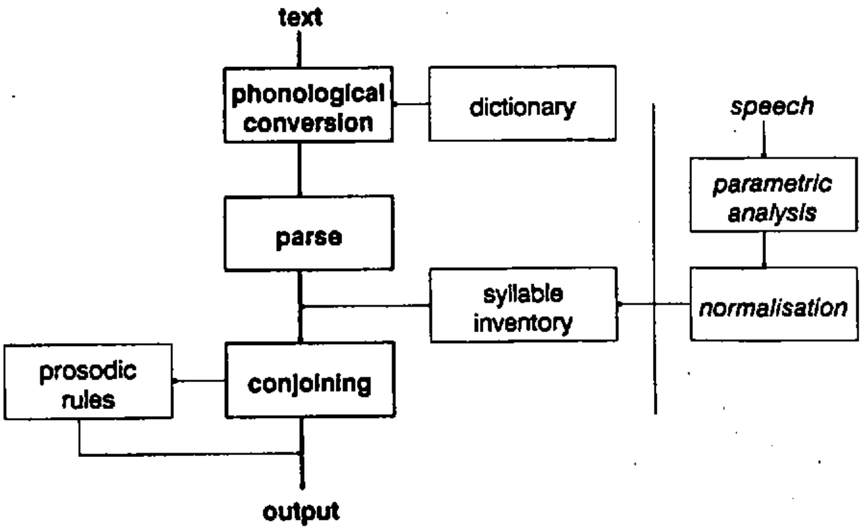
Probably the most important departure from accepted synthesis philosophy found in SPRUCE is that speech is no longer regarded as a string of allophone sized segments coarticulated together. The basis for building the output is shifted to the syllable. Parameters are no longer regarded as independently controllable. Instead they are grouped, and relatively simple control signals command groups which respond as a unit to such commands. The group is an object with its own built in reactions to control signals; that is, the signal itself does not specify the procedures of reaction. Groups are classified according to how they react to control signals.

After fairly extensive pilot work it was found that the best unit to form the basic building blocks for synthesis was the syllable. There are a number of reasons for this, but they fall simply into two types: those concerned with the preservation of the elusive naturalness in real human speech, and those which take into account the complexity of the corresponding conjoining algorithm.

The obvious units to examine are those determined linguistically: sentence or phrase, word, morpheme, syllable, allophone. The correlation with naturalness is obvious – whole sentences are the units which preserve the most naturalness, as shown by any number of demonstrations of resynthesised speech; allophone sized units as used in almost all systems preserve the least naturalness. The problem of course is that generality of the system inversely correlates: the most general systems are based on allophones, the least general on whole sentences.

The choice for SPRUCE lay in the middle since sentences were too prescribed ,and abstractly represented allophones too unnatural and too difficult to conjoin convincingly. We saw little point in going for morphemes rather than words – why split words according to complicated rules when whole words would do just as well? In the end the syllable was chosen as the basic unit on the grounds of versatility. The tradeoff lay in the conjoining rules: the larger the linguistic unit the less susceptible it is to error in conjoining; listeners seem to be more tolerant of joins in synthetic speech the larger the unit. This is certainly true of resynthesised speech. However, it was found in pilot studies that we could satisfactorily conjoin syllables without perceived loss of naturalness, though at the price of more elaborate rules than would have been needed for words. Figure 1 shows a block diagram of the system.

Fig1 Block diagram of SPRUCE



1. Input

In its simplest version SPRUCE accepts text input in the same way as existing systems. It also accepts marked text for special effects, though these are not to prevent failure of lower level parts of the system as they are in current synthesis. Marked text in SPRUCE is reserved for adding pragmatic markers (see below).

2. Orthography to phoneme conver sion

Orthography to phoneme conversion is avoided in SPRUCE, th ough it is there for use in special circumstances. Under normal conditions all input words are searched for in a lexicon in which, for each entry, are found various specifications and markers associated with word-items. The projected lexicon is very large. Each lexical entry indicates a word's syntactic category, certain prosodic features (such as word stress), and the syllabic make-up of individual words. The representations here are completely abstract, though they obviously relate to normal orthography and to the representations used lower in the system for the physical specification of syllables.

3. Prosodics

The principal prosodic feature to be assigned is intonation. In SPRUCE, after lexical filtering, an input text has all its words marked according to their syntactic category. This enables a syntactic parse of the sentences and phrases to be performed without much difficulty. The parse output is used in assigning abstract markers of intonation to the text. These markers are later interpreted in terms of actual time-governed f0 values. The model is based on Pierrehumbert (1981). Provision is made for varying rate of delivery within a sentence, once again dependent on the syntactic parse.

Human speakers, when reading a text aloud, need to understand what they are saying in order to assign prosodic features unambiguously. SPRUCE does not attempt a semantic parse of the text (except when pragmatic markers are available), and is therefore unable to have any understanding, except that based on syntax, of the text. The prosodic unit however has been designed to estimate the probability of error for any one solution (and some sentences may have more than one solution) and to equate this score with a score associated with the probability of the unit's preferred solution occurring in the language. This procedure minimises the likelihood of unusual contours which might sound rediculous to the listener. When SPRUCE's prosodic module fails it usually fails gracefully.

4. Segments

The objects from which the speech output is built are syllable sized. The lexicon assigns abstract syllabic representations to the input text. These are later interpreted in terms of more physical representations. The system holds a large inventory of files of normalised parametric analyses of samples of real human speech. It should be stressed that these are not parameterised recordings: they are normalisations which have been derived from parametric analyses. Each entry in the inventory has markers attached which are associated with the way in which the syllable behaves in certain contexts: this assists the process of conjoining. Although these syllable-sized objects, because of the normalisation, are abstract they are unlike the usual segment representations in other systems. In the latter it is usual to use a single column of parameter values to represent each allophone – a highly abstract representation, since in no way does it convey any sense of time. Despite the usual presence of a time marker for each allophone, such a representation could not convey changes which occur during an allophone. In SPRUCE tiny variations in parameter values during a syllable are preserved in the normalisation procedure.

5. Segment conjoining
The conjoining process is surprisingly simple compared with some of the elaborate systems devised for joining allophones. A small number of conjoining templates, together with the markers found attached to each syllable determine the way in which syllables are conjoined. Rhythm and other factors determine local variants on the templates themselves as they are applied.

6. Segmental and supra-segmental fitting
Segmental and supra-segmental fitting are accomplished using an algorithm based on Silverman (1987), designed to complement the earlier assignment of an abstract intonation contour. This stage of intonation assignment is straightforward, the difficulties arising in the earlier stage rather than here. This is the point where anomalies of rhythm are tidied up.

The system leaves markers and hooks for altering some pragmatically determined changes such as the conveying of attitude or emotion. The pragmatic information is contained in the only markers allowed in the input text. For a discussion of the incorporation of effects dependent on pragmatics into synthetic speech see Morton (forthcoming).

Results

Results of pilot studies for SPRUCE indicate promise. Greater naturalness has been achieved with consistency than is managed by current text-to-speech synthesis systems. It remains to be seen whether the Project can carry over the promise of the pilot work into a fully versatile yet robust final system. The areas of greatest difficulty in putting together the system have been highlighted as the normalisation of the parametrically analysed syllable units, efficient searching of the lexicon and inventory of syllables, unit conjoining, and the method of indexing the final f0 contour for subsequent processing under pragmatic constraint.

References

E Fudge. *English Word Stress.* Hemel Hempstead: George Allen and Unwin (1984)

J N Holmes. *Speech Synthesis and Recognition.* Wokingham: Van Nostrand Reinhold (1988)

K Morton. 'Pragmatic phonetics', in W A Ainsworth, Advances in Speech, *Hearing and Language Processing.* London: JAI Press *(forthcoming)*

J Pierrehumbert. 'Synthesizing intonation', *J. Acoust. Soc. Am.* 70, pp 985-995 (1981)

K E A Silverman. *The Structure and Processing of Fundamental Frequency.* PhD Thesis, University of Cambridge