The British Telecom Research Laboratories Text-to-Speech Synthesis System - 1984-1986

[PART I]

Katherine Morton

Reproduced from Katherine Morton: ***Speech Production and Synthesis*** - unpublished PhD thesis, University of Essex 1987, pp. 142-160. Based on reports to British Telecom Research Laboratories during the period 1984-1986 (part of a series of research contracts over a nine year period). The reports themselves cannot be placed in the public domain for reasons of confidentiality and intellectual property rights.

The BTRL speech synthesis system referred to here is a parallel formant system based on an original design by John Holmes while he was at what was then the Joint Speech Research Unit at Dollis Hill in London. 'Joint' referred to funding collaboration between the British Ministry of Defence and the Post Office - the telephone network and its research subsequently broke away from the Post Office and later became the company ***British Telecom*** or ***BT***. The research unit moved to Martlesham Heath, near Ipswich, in Suffolk. The current (1997) BT system, called ***Laureate***, is a quite different system based on concatenated waveform synthesis.

A. Background
- Linguistic theory: the human speech process
- Text-to-speech synthesis
- The BT Research Laboratories text-to-speech system

B. Synthesis
- Problems for synthesis

C. Allophones
- Coarticulation

D. Transitions
- The JSRU simulation of transitions
- The Essex model of the JSRU simulation
- The Essex program
- Graphical conventions
- Purpose of the Essex transition model

E. Observations on transition handling

F. Conclusion

A. BACKGROUND

The output of the human speech production process is a continuous or quasi-continuous acoustic wave. The aim of synthetic speech systems is to simulate this soundwave. The British Telecom Research Laboratories (henceforth BTRL) simulation is an engineering implementation of the model of speech production developed in the 60s and early 70s. In theory the model will produce a reasonably intelligible speech output, though it cannot in principle be perfect. For example, no account is taken in the system of the variability found in

speech at either the segmental level or the prosodic level. Good practical demonstrations of the near limit of this approach are the DECtalk synthesizer produced by Digital Equipment Corporation and the Prose-2000 system produced by Telesensory (California), both derivatives of the Klatt MITalk. None of these devices however simulates the neurophysiological, anatomical or articulatory properties of speech. They generate control signals analogous to those produced by the phonological processes of human speech, and then move directly to an acoustic model of speech production. [A]

Linguistic theory: the human speech process

The linguistic theory on which such synthesis systems rest assumes speech processing can be divided into two parts: a prosodic aspect, and a segmental aspect.

A prosodic contour is generated to interpret the semantics of a speaker's concept to a limited extent. This contour expresses stress and intonation by varying amplitude, duration and fundamental frequency over the sentence domain. Intonation and word stress are normally dealt with separately in linguistic theory. In effect, the prosodics relates directly the semantics while syntax is seen as largely subordinate. The role of syntax is to get the appropriate words or morphemes from the lexicon and make sure they are correctly ordered within the sentence. In the model of human speech processing, prosodics are generated separately from syntax and phonology.

The segmental phonological representation is specified after syntax and phonological processes. In the model of the human system prosodic and segmental representations are then fitted together to provide the input to the articulatory control system. Coarticulatory processes modify the phonological specification resulting in production of a speech waveform.

The text-to-speech synthesis system is in effect modelling a speaker reading text aloud, though this is a process about which linguistic theory has little to say. A major question in the theory of this aspect of human speech production is to determine what is in fact encoded in ordinary plain text. For example, it seems that a major part of prosodics is not encoded, and that plain text requires a great deal of what might be called active interpretation (Fig. 1) by a speaker before it can be turned into speech.

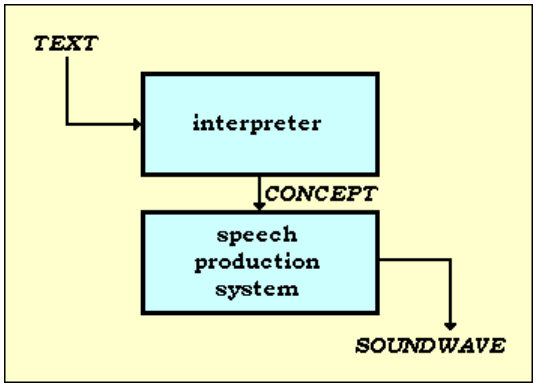

Fig. 1 Human speech production from text.

In general, the orthography of a language like English is a representation at the morphemic level - that is, the letters of the orthography are used to identify morphemes and the way they combine into words, rather than to code their pronunciation. English orthography is not 'phonetic' in nature. It is well recognized therefore that reading aloud involves accessing both a lexicon from which the reader can derive the underlying phonological shape of morphemes represented in the text, and also a full phonological component for processing underlying shapes for speech. [B]

Text itself contains very little information about the appropriate sentence prosodics. Markers like commas, full stops, etc. indicate breaks of one kind or another in the flow of

information, but they do not fully encode prosodics. A text requires full prosodic interpretation by the reader and since a greet deal of prosodics directly depends on the meaning of the text, the reader must understand the text before he can generate the right prosodic contour. Understanding involves a semantic decoding of information and involves some guesswork as to the exact intentions of the writer. It is for this reason that very often alternative prosodic interpretations of a text are equally valid (as for example in the way different actors may have different interpretations of the meaning intended by a playwright).

The process therefore is active - that is, it calls for an interpretation of the original text by involving human intelligence and information not contained in the input. The conclusion is that text standing alone is a slightly defective encoding of speech. Therefore, in text-to-speech synthesis as many defects as possible must be compensated for, as they are when a human being reads aloud.

Text-to-speech synthesis

Most current text-to-speech synthesis systems are put together in the way shown in Fig. 2

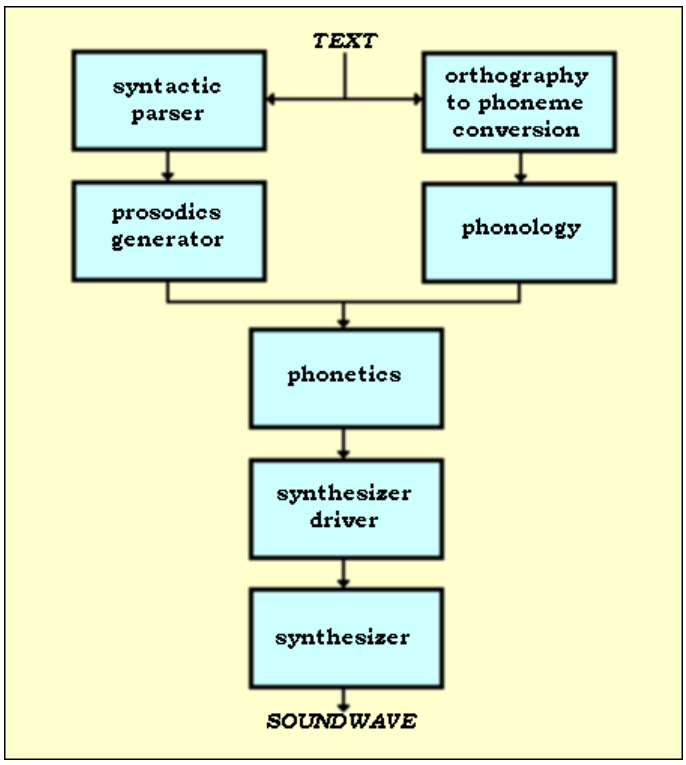

Fig. 2 A typical text-to-speech synthesis system.

The input text , in orthographic symbols, is converted to strings of phonemes. This conversion may be a. by rule, or b. dictionary lookup of whole words and/or morphemes. Phonological rules are then applied to the phonemic strings to produce strings equivalent to those output from a human production phonology. The output strings consist of extrinsic allophones which may be expressed in terms of distinctive features.

At the sane time a rough parsing is performed. In the least sophisticated synthesizer this consists of consulting a dictionary of function words in the incoming text. Those which are not marked as function words are by default labelled content words - the other category in this simple syntactic model. The appropriate sentence stress intonation contour are generated using these labels and whatever punctuation is marked in the text (Mattingly 1966). Word stress is assigned by a. dictionary lookup and b. stress assignment rules.

The description of a human being reading plain text aloud is more complex. When reading aloud, parsing and generating the correct prosodics probably operate at both the semantic and syntactic levels. Not only is syntactic information decoded from the text , but the text itself is 'understood'. Both processes are active in that information not in the text is brought to the task by the human being. In the synthesizer text is treated passively on the basis of the non-semantic information it already contains. In synthesis systems of this kind, this is an obvious area of deficiency: interpreting the text by referring to a semantic component cannot yet be done satisfactorily. [C]

A more accurate simulation of reading text aloud would include modelling the intelligence used in semantic and syntactic parsing. Although current speech synthesis systems can render segments very well, prosodic aspects are often quite poor, reflecting the possibility that segmental processing in the human being is more passive than prosodic processing. That is, as the system passes through various levels the segmental string is rewritten in more detail, whereas deriving prosodics from text requires processing which cannot be a simple transformation of the input. The conclusion is that an artificial intelligence unit to deal with semantics and syntax in text-to-speech synthesis systems will be necessary to improve prosodics.

The British Telecom Research Laboratories text-to-speech system

As in the generic text-to-speech system outlined above, the BTRL system is based essentially on the Joint Speech Research Unit's synthesis-by-rule system (Holmes *et al.* 1964 and subsequent papers from JSRU). Input consists of a string of text characters, rather than the abstract concept used by a human being when not reading text aloud (Fig. 3). As explained, although text is a suitable input for generating segmental aspects of speech, it is not suitable for generating prosodic aspects. The BTRL system therefore shares with almost all synthesis systems that its segmental rendering is adequate but that its prosodic rendering is unsatisfactory. For such a synthesis system to generate a satisfactory output, it would be necessary to provide a concept extractor as in the human being, to interpret the meaning of text.

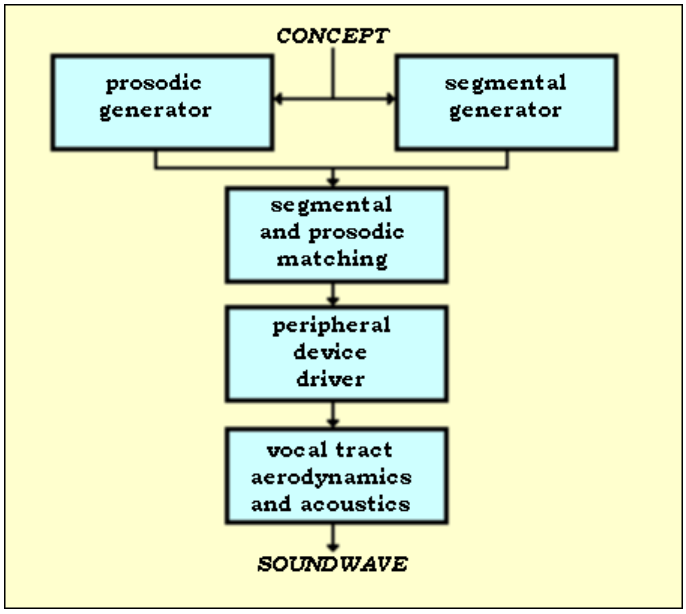

Fig. 3 Human speech production from concept.

B. SYNTHESIS

The approach for synthesizing speech has been to simulate human speech production by implementing the model just described. This assumes

- that individual words can be represented by a concatenation of phonemes or allophones,
- that over the sentence domain an approximate specification of the prosodics without artificial intelligence is possible,
- that it is possible to specify a driving signal for the peripheral sound generator (i.e. the synthesizer),
- that a set of rules can be formulated relating phonemes to the driver parameters for speech output. [D]

Synthesis systems incorporate a lookup table which specifies numerical values for allophones. These values, and rules operating to combine then sequentially, form the input to a synthesizer driver.

Problems for synthesis

There are four major problem areas in setting up synthesis systems of this type:

- establishing a suitable number of phonetic objects - allophones - to provide for contextual variants;
- providing a lookup file of numerical values that adequately specify these allophones for driving the synthesizer;*

  [**footnote:* The synthesizer is usually thought of as a separate device from the set of rules which derive its input - the two together being termed 'the synthesis system'.]
- specifying the intonation contour in such a way that segmental speech can be fitted to it;
- specifying word stress and rules for interacting with the segmental allophones.

C. ALLOPHONES

An important part of the synthesis-by-rule (henceforth SbR) program is concerned with deriving allophones from the phonemic string assigned to the input orthographic string. As has been discussed in previous reports, this can be done in two stages:

- deriving extrinsic allophones by phonological rules, which in the human being make up the lowest level representation of the voluntary aspects of speech, and
- deriving intrinsic allophones by phonetic rule, which are a further representation incorporating involuntary transformations to extrinsic allophones, i.e. coarticulation.

Coarticulation

In the human being coarticulatory phenomena produce intrinsic allophones. Coarticulation occurs when, for two conjoined extrinsic allophones, a particular articulatory parameter required in both segments is differently specified in each. For example, segment A may require a high front tongue position and segment B may require a low back tongue position. Such an abrupt discontinuity is impossible at the temporal boundary between the two segments, so the surface effect is one of progressively deforming the specification of segment A as the boundary approaches, and, from a deformed specification, progressively moving towards the 'target' specification of segment B. (E] The phenomenon is shown in stylized form in Fig. 4.

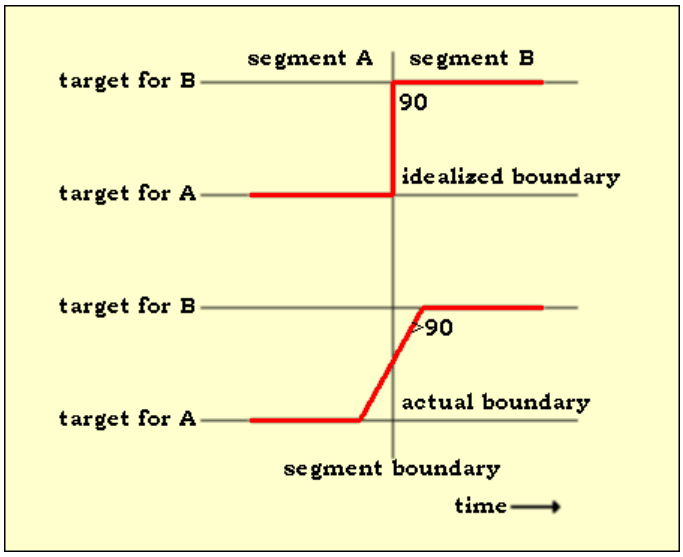

Fig. 4 Comparison of idealized and real boundary transitions.

The phonetics literature details underlying mechanisms and effects responsible for this observation. The phenomenon is caused by mechanical inertia, general properties of the neuro-muscular system and by properties of the overall articulatory control system.

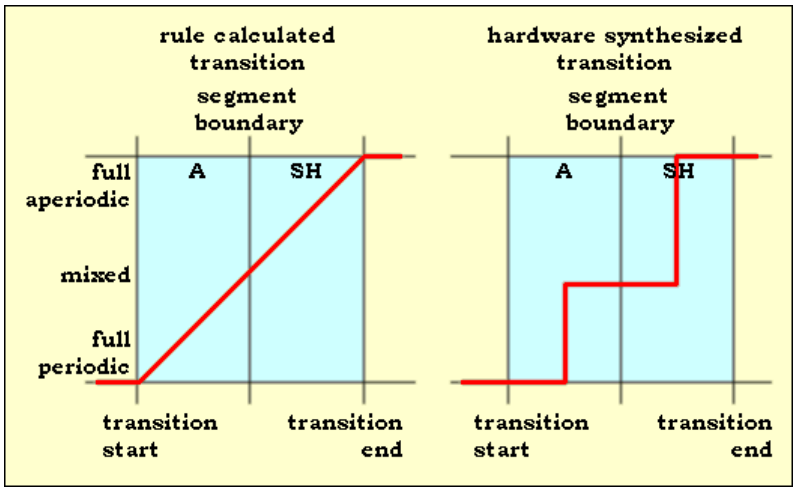

Fig. 5 Voicing parameter transition, showing hardware distortion (on the right) of the calculated transition between full periodic excitation and full aperiodic excitation.

The BTRL software calculates a similar transition, but the hardware at the moment cannot simulate the continuously variable transition shape. Since the hardware is designed with only three possibilities: full periodic, mixed, and full aperiodic, the segment transition is rendered as discontinuous, rather than as continuous.

Transition simulations which produce speech output that is intelligible or unnatural can be dealt with in several ways:

- refine the rule system,
- increase the number of allophones in the segmental lookup table,
- insure that the hardware and software are compatible and that the hardware can adequately deal with the output of the software system.

Initially it will be useful for the BTRL system to increase the number of allophones included by JSRU the segmental lookup table. For example, at the moment the entry for /B/ (JSRU transcription; see Holmes *et al.* 1964] a compromise between three possible

specifications; the three allophones /B/ in a. word initial, b. intervocalic, and c. word final positions.

These allophones have different acoustic characteristics in conjunction with other segments. Adding to the allophone table means that provision must also be made in the SbR program for calling these new entries in particular contextual environments or they will never be accessed.

D. TRANSITIONS

The JSRU simulation of transitions.

The JSRU simulation of transition effects between conjoined segments determines boundary values for parameters, and manipulates these values according to rules set out elsewhere in this report. [F] From data supplied by entries in the lookup table, the rules calculate boundary values, the slope of the transition and the character of the transition (currently a straight line). The angles are subsequently smoothed by low-pass filtering the synthesizer control signal, except in the case of abrupt changes, such as a plosive burst, where smoothing is not applied. The JSRU hardware handles this quite well.

However, by changing the slope of the transitions, shapes approximating transitions in natural speech can be achieved. Slopes can be altered by adding new entries to the allophone set, specifying different values for FC (fixed contribution), ID (internal duration) and ED (external duration), and by changing the rank value assigned to these new segments.

Essex model of the JSRU simulation

We have copied the lookup tables specifying parameter values of the segments onto our own laboratory computer, and programmed rules for calculating parameter transitions across boundaries as they appear in the RTL2 software [RTL2 is the real-time computer programming language used for JSRU software]. The output is presented as parameter traces on a VDU or graph plotter. We were particularly interested in having access to this kind of objective presentation of the conjoining rules, since anomalies and distortions are difficult to detect and assess by listening to the acoustic output.

The Essex program

We have modelled parts of the SbR system on our lab computer (HP-85). [G] This section describes the start of work on the systematic modification of the lookup tables containing the lower phonetic specification of 69 allophone segments. This file includes information from which boundary markers can be derived and transition shapes constructed. We have also implemented the conjoining rules which calculate transitions between the segments.

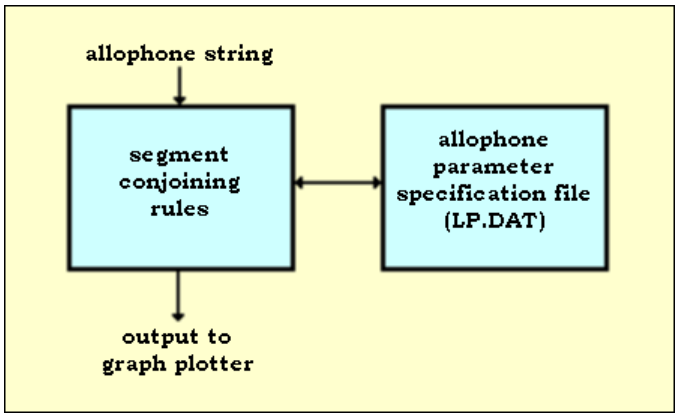

Fig. 6 Program design for producing graphical output of the conjoining rules.

This graphical display has been designed to show what an ideal spectrogram of the output of a synthesizer with a transparent transfer function looks like according to the JSRU algorithm.

This display has the aspect ratio and scaling of a standard 4 kHz spectrogram produced on a Kay Sonagraph. In the following example, the first three formants have been reproduced to show formant transition between segments as they would appear in the SOI file [the .SOI file collects the control parameters for outputting to the hardware synthesizer]. Amplitudes are not interpreted on this display (but see below). A particular convention has been adopted in the examples in this section with respect to excitation source. In the BTRL system, there are only three possibilities: periodic, aperiodic and a fixed ratio mix of the two (see the discussion above). The display convention is as follows:

Graphical conventions

Vowels

Periodic excitation is indicated by vertical hatching of formants. Hatching (and the corresponding dotted hatching used to indicate aperiodic excitation) is used to mark synthesizer control frames. The short vertical Lines are spaced at 10ms intervals, the period of one time frame.

Consonants

Plosives: for [-voice] plosives, the trajectories of the formant parameters are shown during the silent interval (in the stop phase of the plosive).

For [+voice] plosives during the stop phase, Fl is shown hatched to indicate periodic excitation, but with amplitudes only on Fl. F2 and F3 are shown as silent formant frequency trajectories.

The burst and aspiration phases of plosives (xY, xZ - e.g. PY, PZ) are shown with Fl either hatched ([+voice]) or unhatched ([-voice]). The presence of aperiodic excitation is shown by dotted hatching of F2 and F3.

Fricatives: aperiodic excitation applies for all fricatives. This is shown as dotted hatching of F2 and F3. Hatching or no hatching of Fl indicates whether periodic excitation is mixed in with the aperiodic source.

Typical examples:

T - silent F trajectories marked,

TY - burst marked as aperiodic source on F2, F3,

TZ - aspiration marked as aperiodic source on F2, F3,

D, DY - stop phase and burst additional marked on Fl as having aperiodic excitation,

D - aperiodic excitation shown on F2, F3,

Z - mixed excitation shown by marking periodic on Fl aperiodic on FZ, F3,

EE - typical vowel showing periodic excitation on all formants.

Purpose of the Essex transition model

The following display examples have been generated by interaction of the rules with the lookup tables. For the moment, we are not looking at context sensitive durational variations of segments generated by the prosodics component of the SbR program. Segments appear therefore with their boundaries marked (vertical dashed lines) and with untransformed intrinsic durations, not with re-calculated durations as appear in the usual .SOI file.

The object is to have a repeatable display of effects on formant frequencies from modifications made to the parameter file and the rules for joining segments. We have judged that listening to the synthesizer's output is problematical, Our model of this part of the SbR program enables effects of changes to the parametric specification of allophones and changes to conjoining rules to be seen immediately and in a graphical format which is measurable.

Specifying new allophones for the lockup table is also easier. The hit-and-miss 'let's try this and see what it sounds like' approach is therefore avoided.

We are developing and extending this model to include amplitude parameters (see below) and rules for adjusting intrinsic durations. We hope to be able to provide more objectively derived recommendations for changes. The model displays the results of proposed changes and is more efficient than measuring spectrograms of the synthesizer's output. Using the ILS package at BTRL for waveform analysis will eventually provide more objective assessment of the synthesizer's acoustic output enabling comparison with the present displays to evaluate effects of the synthesizer hardware. Our own computerized waveform analysis system should be available by the new year pending availability of ILS on the BTRL computer system.

> [The programming of the model was done by Marcel Tatham in semi-compiled Hewlett-Packard extended technical BASIC and HPGL. The program runs on the Advanced Speech Technology Laboratory's HP-85A desktop computer, with output to VDU and/or a Hewlett-Packard graph plotter.]

E. OBSERVATIONS ON TRANSITION HANDLING.

The observations in this section have been made by examining graphical output from the Essex model of the SbR transition algorithm. The graphs do not include adjustments made for duration of segments by the SbR prosodics program.

A few examples are given for comparison with digitized spectrograms of real speech. These were digitized by hand from real spectrograms and are an example of a further section of the current programs being developed. aspect ratios, etc., have been adjusted to enable direct comparison with the output of the SbR program and with real spectrograms.

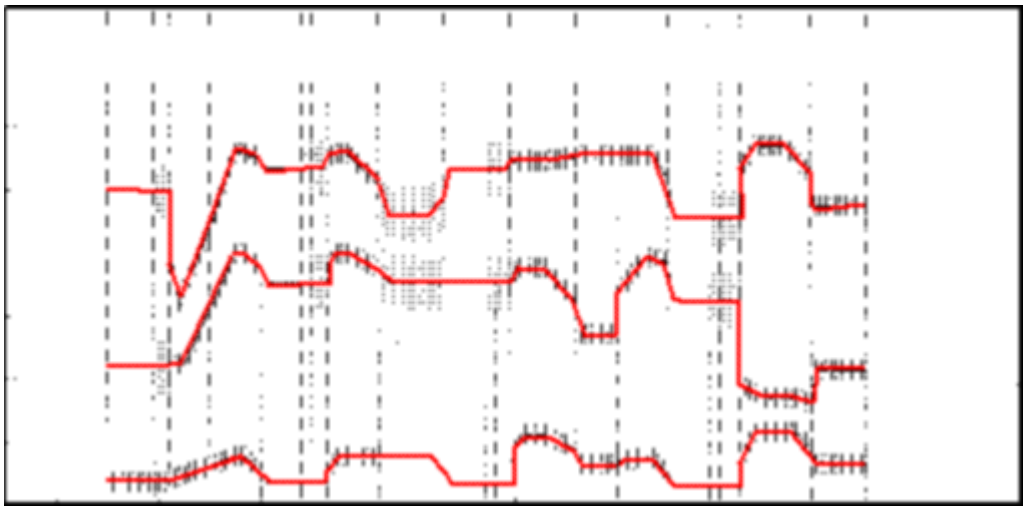

Fig. 7 'British Telecom' - synthetic: y-axis 4 kHz, x-axis 200 frames.

Fig. 7 illustrates the kind of output our program produces. The display has a time axis of up to 2s, marked off in 10 ms frames (the frame duration is the same as that of the JSRU system). The y-axis is frequency in Hz, up to 4 kHz since this is the range of the parameter specification in the synthesizer. Vertical dashed lines on the graph mark segment boundaries, and in all examples in this report the durations of segments are the standard intrinsic durations without prosodic modification. In this display no account has been taken of amplitude other than a binary representation: where there is no amplitude then no hatching appears. Hatching indicates some amplitude, but not how much. There are two types of hatching indicating whether the excitation is periodic (as in vowels) or periodic (as in voiceless fricatives).

Two things in particular should be noted in the illustrations in this section of the report:

- discontinuities of the formant frequency trajectory through the utterance,
- change of 'shape' of the trajectory, which can be abrupt.

Generally speaking, changes of resonance frequency in human speech are due to tongue movement which is adjusting the size and shape of the resonating cavity. Because of mechanical constraints, it is not possible for tongue movement to be discontinuous or abrupt beyond certain limits.

For example, in Fig. 7 we can look at the juncture between segments KZ and O in the word *Telecom*. KZ is the aspiration phase of K, and O is the final vowel in the word. There is a discontinuity on F2 at the boundary between these segments with a drop in frequency from around 2.7 kHz to 1 kHz in apparently zero time - something which is mechanically impossible for tongue movement. Aspiration can be thought of as devoiced vowel, so we would expect the tongue to be approximating the correct vowel position during this phase. In real speech we would expect a relatively smooth trajectory of F2 from the end of the I (the middle vowel in the word) to KZ and from there to O.

General smoothness is therefore important, and spectrograms of natural speech are characterized by this (see digitized spectrogram of real speech). It is very clear that the jagged trajectory of F3 during the R in British is wrong. Although this segment in this context would in natural speech show considerable movement of F3, it would not be abrupt.

Occasionally a segment target is missed completely due to the effect of adjacent segments. In this phase there are two examples of the phenomenon: in F3 of R and in F2 of I in Telecom. When a target is missed the program marks the value of the missed target with a short horizontal line at the appropriate value and at the time it was expected that the target would be hit. Thus there is a mark just below the jagged part of F3 in R, and just above the central part of F2 in I. Missing targets is extremely common in natural speech, and we were surprised by how rarely the SbR program generates missed targets. This may be a function of the fact that durational adjustments have not been made within this program, and it remains to be seen just how often segments get shortened by the durational rules, thus increasing the probability of missed targets.

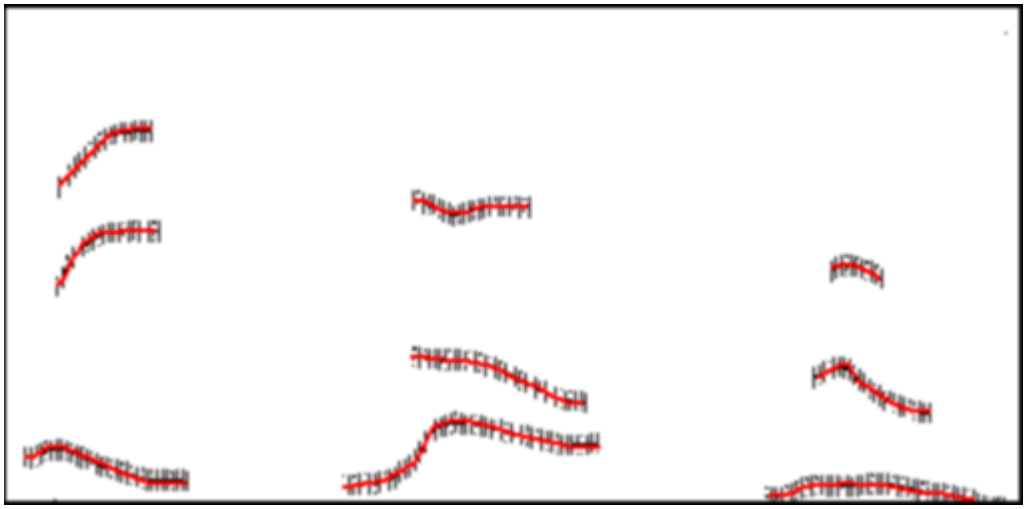

Fig. 8a 'li, la, lu' -digitized spectrogram of natural speech: y-axis 4 kHz, x-axis 200 frames.

Fig. 8a is a digitized spectrogram of [li], [la], [lu] in natural speech. These spectrographic representations show a formant only when there was sufficient amplitude for marking to occur. Therefore, one of the characteristics of [1] is that F2 and F3 have little or no amplitude. Notice the smooth transitions of F2 and F3 in [li]. This is where we would expect the most discontinuity, since the intrinsic values of F2 and F3 are more different between [l] and [i] than they are between this consonant and other vowels.

Notice that at the end of the vowel in each case, Fl persists longer than the other formants. This is characteristic of voiced segments in final position: the amplitude of Fl persists longer than the amplitudes of the other ferments.

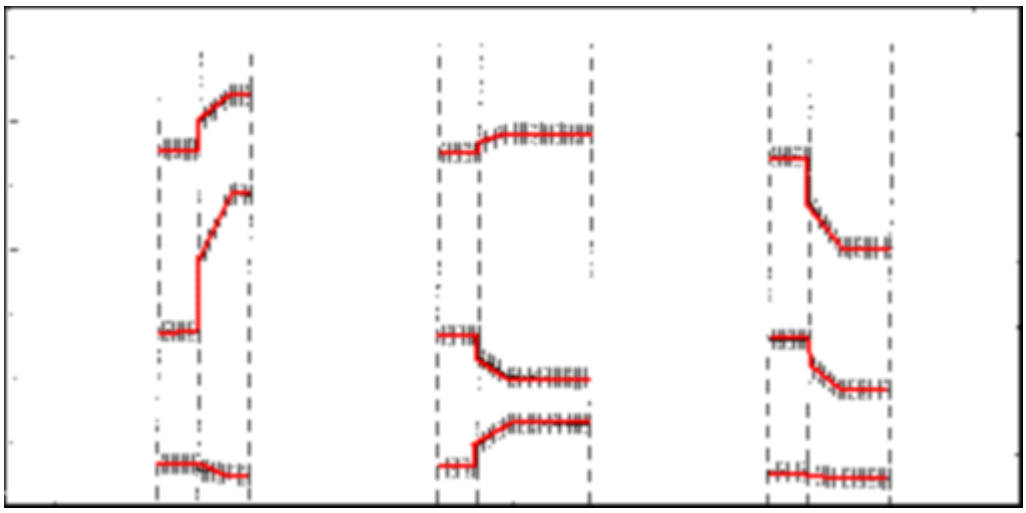

Fig. 8b 'lee, lar, luu' Synthesis by rule output: y-axis 4 kHz, x-axis 200 frames.

Fig. 8b shows the SbR output of the same three nonsense words - L EE, L AR, L UU, using JSRU transcription. In the case of L the formant frequency values are held throughout the segments which is not the case in human speech. As a result , the transition to the second segment is abrupt on all three formants. Although the second segment (the vowel) in each case has transitions calculated from the boundary, between the segments they are discontinuous with the end of L in each case. This shows an error in the specification of L, since, from about the halfway point in its duration the formants should start bending toward the boundary values given at the start of the vowel.

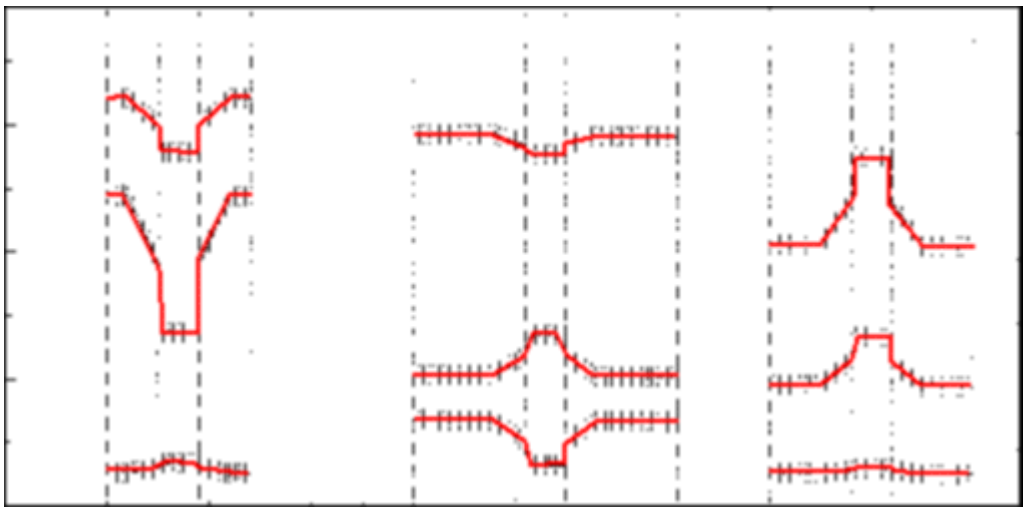

Fig. 8c 'eelee, arlar, uuluu' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 8c shows intervocalic L: in this case between the same vowels shown in Fig. 8b: EE L EE, AR L AR, UU L UU. Once again, as in the other examples, L is characterized by having no transitions within its boundaries. A further point to note is the symmetry of transition shape: this does not always occur in human speech (Ohman 1966a).

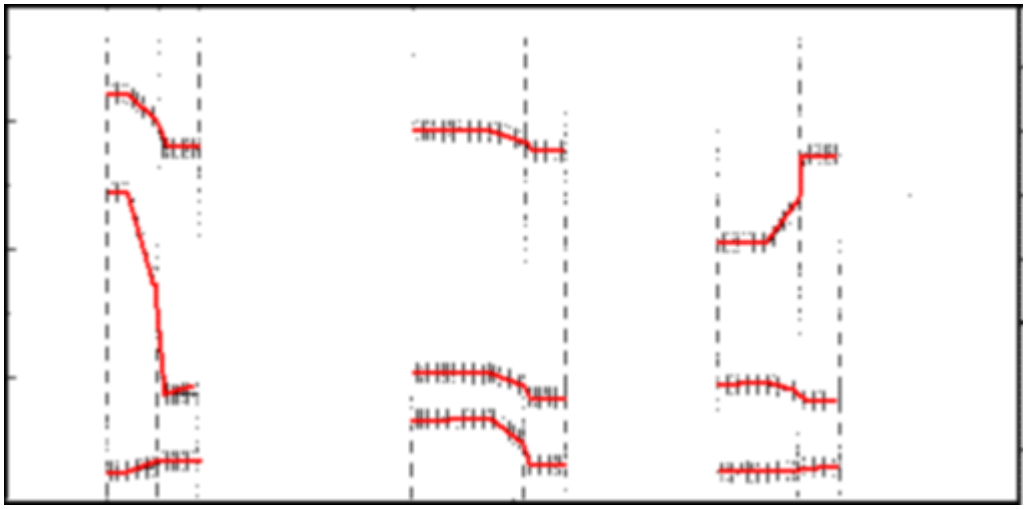

Fig. 8d 'eelp, arlp, uulp' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 8d shows L in final position - occurring again with the same vowels, although this time L is the dark L, LP. There is no difference between these illustrations and the first two segments of each of the graph in Fig. 8c. This kind of symmetry does not occur regularly in human speech.

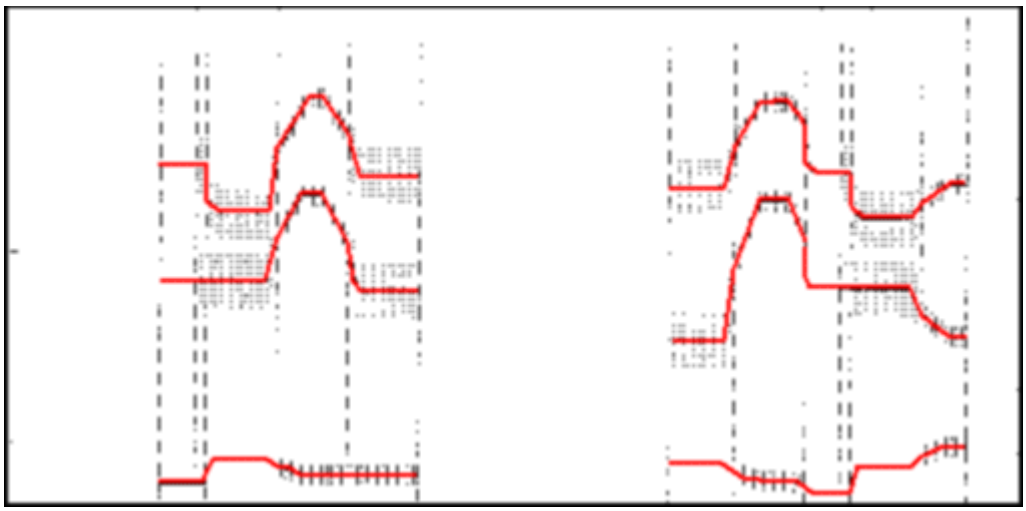

Fig. 9a 'cheese, feature' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 9a - *cheese* and *feature*: voiceless affricates. These two words have been fairly well rendered, although there are some points to note:

- the transitions into and out of the SH segment could have taken up a greater proportion of the segment duration;
- SH (like the other fricatives) does not often show a formant structure in human speech;
- resonance effects can occupy a single area of wider bandwidth than the 100 Hz or so characteristic of a formant. As mentioned elsewhere in this report the JSRU FORTRAN synthesizer interprets the parameters as specified in this illustration to something approximating the natural resonance characteristics of a fricative by combining three formants into a single band. It is necessary to check that this facility is fully implemented in the BTRL synthesizer, since to produce an acoustic signal having the structure shown here would sound wrong.

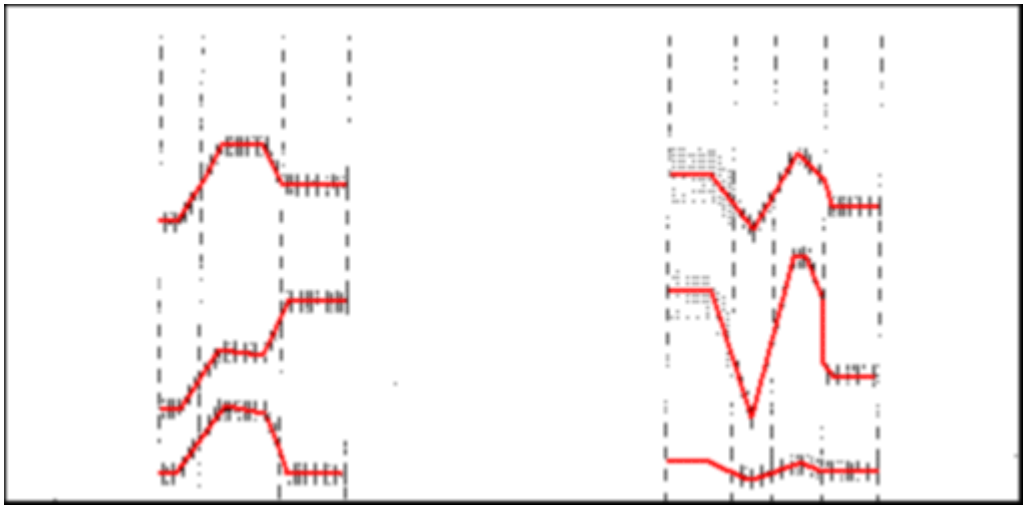

Fig. 9b 'one, swim' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 9b *one* and *swim*: [w] and nasals. In these displays the nasal formant FN has not been displayed. However the other three formants have been relatively accurately handled, though the transitions from the preceding vowel segments should spill slightly into the nasal. This would avoid the discontinuity shown at the boundary between I and M in F2 of *swim*. Initial W in one has been fairly well calculated, but the W in *swim*, particularly in F2 and F3, is abrupt and not characteristic of human speech.

In the JSRU system abrupt changes of direction by control signals to the synthesizer are low-pass filtered to produce a smoother and less discontinuous effect. This may work in practice, though we have yet to examine it systematically. However it could have serious knock-on effects where discontinuity is in fact required. An example here would be the transients associated with plosive release.

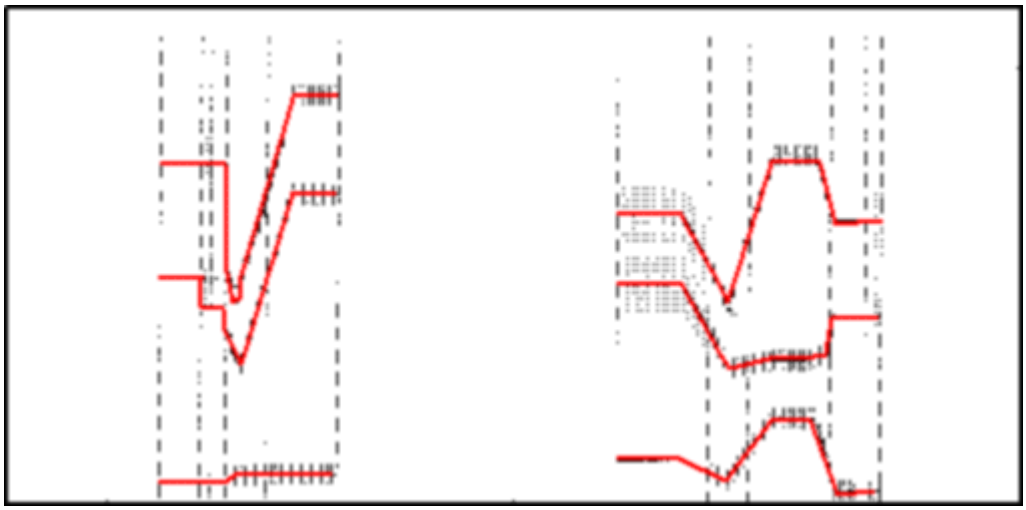

Fig. 9c 'tree, shrug' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 9c - *tree* and *shrug*: R. R is a particularly difficult segment to synthesize. The problem is that the transitions generated span a very wide frequency range (as with L, W, J) and are difficult to calculate without producing the jaggedness shown in these diagrams. Note that for R in both words the target (marked with a short horizontal lime) was missed in F3, as it should have been. The transitions between R and the following vowel are relatively well done.

Liquids and semivowels (R, L, W, J), as mentioned, are particularly difficult to render in synthetic speech. One problem has been the transition shape between segments. We intend to adjust the rules to generate smoother transitions in these cases and try to avoid relying on subsequent smoothing of the control signals within the hardware.

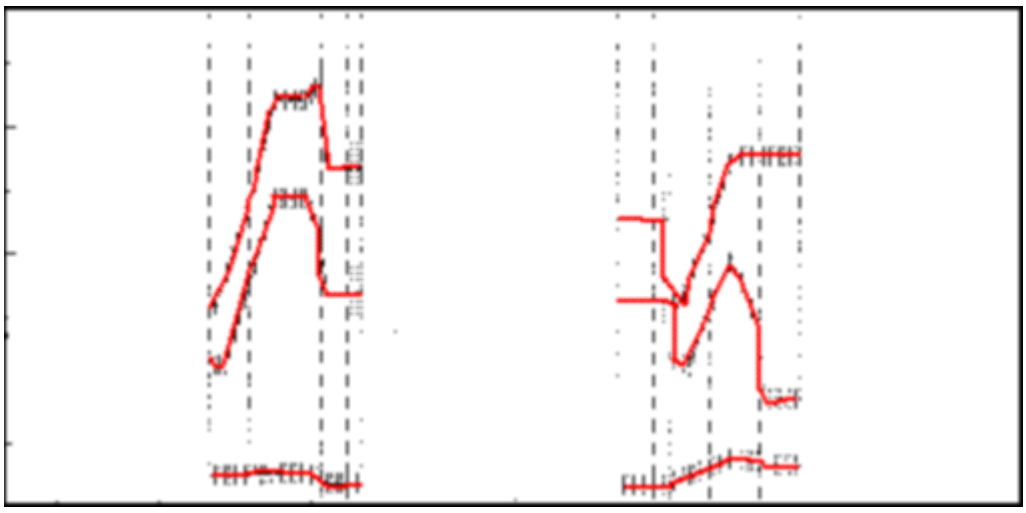

Fig. 9d 'read, grill' - SbR output: y-axis 4 kHz, x-axis 200 frames.

Fig. 9d - *read* and *grill*: R. There are clearly some difficulties next to these high front vowels and their extreme F2 and F3 values. For example, the influence of the EE 3rd formant has caused a missed target the R at the start of the utterance read. There is another missed target for I in grill.

Although we are drawing attention to discontinuities in formant values at segment boundaries (e.g. between I and LP in grill) these discontinuities are not perceptually important if one of the segments has no amplitude. Examples of such an occurrence appear in each of these two words: the boundaries between EE and D in F2 and F3 of read and between GY and R in grill. Although, of course, their effects cannot be heard directly, these discontinuities will affect the transition calculations.

Fig. 9e, f, g. Similar effects to those mentioned above are apparent in these figures.

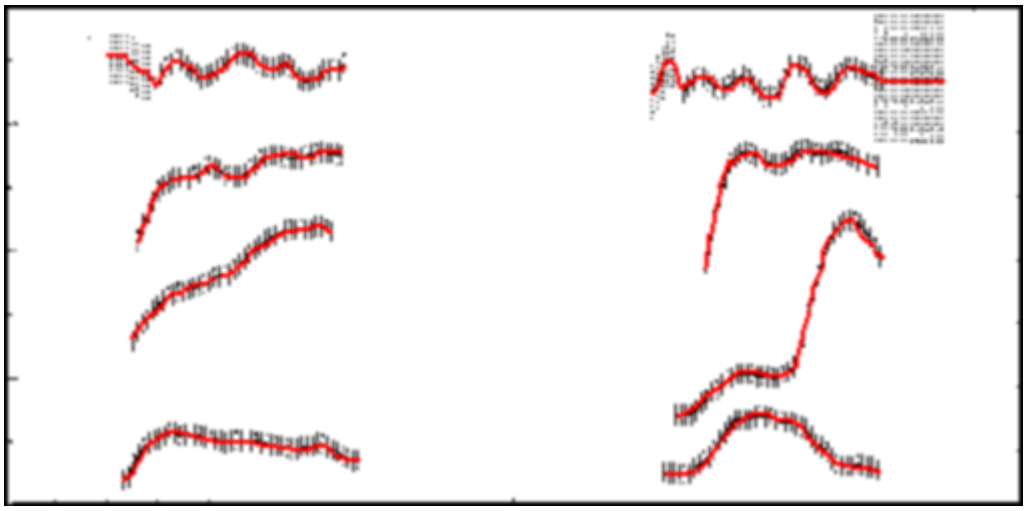

Fig. 10a 'pray, twice' - digitized natural speech: y-axis 4 kHz, x-axis 200 frames.

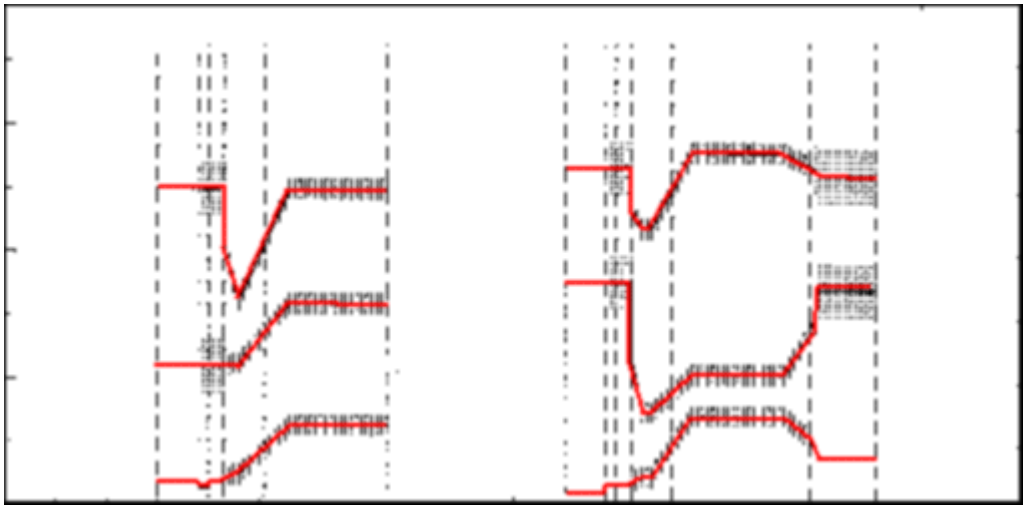

Fig. 10b 'pray, twice' - SbR output: y-axis 4 kHz, x-axis 200 frames.

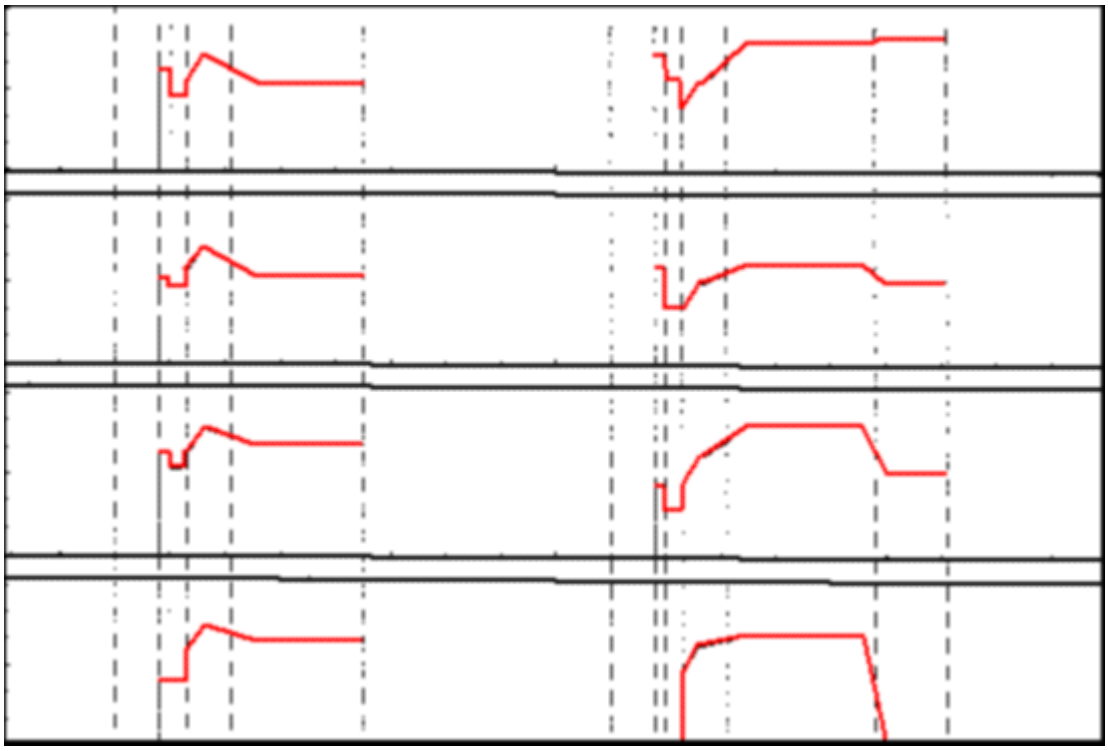

Fig. 10c 'pray, twice' - amplitude plot: y-axis 60 dB, x-axis 200 frames.

Fig. 10a, b, c - *pray* and *twice*. Fig. 10a is a digitized spectrogram of natural speech for these two words. The fourth formant is shown here for completeness. In human speech the fourth formant averages around 3.4 kHz, but nevertheless moves in the frequency domain, whereas in the synthesizer F4 frequency is fixed. Spectrograms of natural speech compared with the SbR output in Fig. 10b (again without durational adjustments) show

- the level frequency sections of the diphthongs in the synthetic version compared with the continuous movement of formants in natural speech;
- durations of transition segments of vowels in synthetic speech are expressed in absolute rather than relative terms in the lookup table. This Leads to very long steady state values centrally in vowels. Some adjustment will have to be made which spreads the transition sections towards the centre of these segments;

- frication (e.g. the fricative in *twice*). Even if 'blending' of F2 and F3 occurs later it is nevertheless the case that a broad band of aperiodic centered around 2.2 kHz is not the same as the band centered around 3.4 kHz as shown in the spectrogram of human speech. There is a high amplitude aperiodic excitation applied in the synthetic speech to F2. This does not appear in natural speech at these frequencies.

Fig. 10c is one example of an amplitude plot of the lookup table entries for segments for these words, processed by the transition algorithm. Amplitude curves are shown separately for Al , A2, A3 and AHF.

F. CONCLUSION

We believe we now have a useful tool to assist the Project. Any combinations of segments can be conjoined to produce a visual output in spectrogram format for direct comparison with spectrograms of natural speech. The technique does not rely on subjective assessment of the waveform produced by the synthesizer, which may have added to it anomalies of the hardware itself. In addition, the results of changes made to the lookup tables can be edited quickly. Experiments on the relative importance of the various parameters can easily be conducted, and it is relatively easy to adjust the rules. We expect to make considerable use of it during the coming year.