

SPEECH SYNTHESIS

One method of synthesising speech involves the concatenation of small units of speech in the time domain. Thus representations of speech waveform
5 may be stored, and small units such as phonemes, diphones or triphones - i.e. units of less than a word - selected according to the speech that is to be synthesised, and concatenated. Following concatenation, known techniques may be employed to adjust the composite waveform to ensure continuity of pitch and signal phase. However, another factor affecting the perceived quality of the
10 resulting synthesised speech is the amplitude of the units; preprocessing of the waveforms - i.e. adjustment of amplitude prior to storage - is not found to solve this problem, inter alia because the length of the units extracted from the stored data may vary.

According to the present invention there is provided a speech synthesiser
15 comprising

- a store containing representations of speech waveform;
- selection means responsive in operation to phonetic representations input thereto of desired sounds to select from the store units of speech waveform representing portions of words corresponding to the desired sounds;
- 20 - means for concatenating the selected units of speech waveform characterised by means for adjusting the amplitude of at least the voiced portion relative to a predetermined reference level.

One example of the invention will now be described, by way of example, with reference to the accompanying drawings, in which:

25 Figure 1 is a block diagram of one example of speech synthesis according to the invention;

Figure 2 is a flow chart illustrating operation of the synthesis; and

Figure 3 is a timing diagram.

In the speech synthesiser of Figure 1, a store 1 contains speech
30 waveform sections generated from a digitised passage of speech, originally recorded by a human speaker reading a passage (of perhaps 200 sentences) selected to contain all possible (or at least, a wide selection of) different sounds.



Accompanying each section is stored data defining "pitchmarks" indicative of points of glottal closure in the signal, generated in conventional manner during the original recording.

An input signal representing speech to be synthesised, in the form of a
5 phonetic representation is supplied to an input 2. This input may if wished be generated from a text input by conventional means (not shown). This input is processed in known manner by a selection unit 3 which determines, for each unit of the input, the addresses in the store 1 of a stored waveform section corresponding to the sound represented by the unit. The unit may, as mentioned
10 above, be a phoneme, diphone, triphone or other sub-word unit, and in general the length of a unit may vary according to the availability in the waveform store of a corresponding waveform section.

The units, once read out, are concatenated at 4 and the concatenated waveform subjected to any desired pitch adjustments at 5.

15 Prior to this concatenation, each unit is individually subjected to an amplitude normalisation process in an amplitude adjustment unit 6 whose operation will now be described in more detail. The basic objective is to normalise each voiced portion of the unit to a fixed RMS level before any further processing is applied. A label representing the unit selected allows the reference level store 8
20 to determine the appropriate RMS level to be used in the normalisation process. Unvoiced portions are not adjusted, but the transitions between voiced and unvoiced portions may be smoothed to avoid sharp discontinuities. The motivation for this approach lies in the operation of the unit selection and concatenation procedures. The units selected are variable in length, and in the
25 context from which they are taken. This makes preprocessing difficult, as the length, context and voicing characteristics of adjoining units affect the merging algorithm, and hence the variation of amplitude across the join. This information is only known at run-time as each unit is selected. Postprocessing after the merge is equally difficult.

30 The first task of the amplitude adjustment unit is to identify the voiced portions(s) (if any) of the unit. This is done with the aid of a voicing detector 7 which makes use of the pitch timing marks indicative of points of glottal closure in



the signal, the distance between successive marks determining the fundamental frequency of the signal. The data (from the waveform store 1) representing the timing of the pitch marks are received by the voicing detector 7 which, by reference to a maximum separation corresponding to the lowest expected
5 fundamental frequency, identifies voiced portions of the unit by deeming a succession of pitch marks separated by less than this maximum to constitute a voiced portion. A voiced portion whose first (or last) pitchmark is within this maximum of the beginning (or end) of the speech unit is, respectively, considered to begin at the beginning of the unit or end at the end of the unit. This
10 identification step is shown as step 10 in the flowchart shown in Figure 2.

The amplitude adjustment unit 6 then computes (step 11) the RMS value of the waveform over the voiced portion, for example the portion B shown in the timing diagram of Figure 3, and a scale factor S equal to a fixed reference value divided by this RMS value. The fixed reference value may be the same for all
15 speech portions, or more than one reference value may be used specific to particular subsets of speech portions. For example, different phonemes may be allocated different reference values. If the voiced portion occurs across the boundary between two different subsets, then the scale factor S can be calculated as a weighted sum of each fixed reference value divided by the RMS value.
20 Appropriate weights are calculated according to the proportion of the voiced portion which falls within each subset. All sample values within the voiced portion are (step 12 of Figure 2) multiplied by the scale factor S . In order to smooth voiced/unvoiced transitions, the last 10ms of unvoiced speech samples prior to the voiced portion are multiplied (step 13) by a factor S_1 which varies linearly from
25 1 to S over this period. Similarly, the first 10ms of unvoiced speech samples following the voiced portion are multiplied (step 14) by a factor S_2 which varies linearly from S to 1. Tests 15, 16 in the flowchart ensure that these steps are not performed when the voiced portion respectively starts or ends at the unit boundary.

30

Figure 3 shows the scaling procedure for a unit with three voiced portions A, B, C, D, separated by unvoiced portions. Portion A is at the start of the unit,

This Sheet Has Been Replaced



THIS SHEET HAS BEEN
REPLACED

PCT/GB96/00529

WO 96/27870

4

so it has no ramp-in segment, but has a ramp-out segment. Portion B begins and ends within the unit, so it has a ramp-in and ramp-out segment. Portion C starts within the unit, but continues to the end of the unit, so it has a ramp-in, but no ramp-out segment.

- 5 This scaling process is understood to be applied to each voiced portion in turn, if more than one is found.

Although the amplitude adjustment unit may be realised in dedicated hardware, preferably it is formed by a stored program controlled processor operating in accordance with the flowchart of Figure 2.

10



CLAIMS



1. A speech synthesiser comprising
 - a store containing representations of speech waveform;
 - 5 - selection means responsive in operation to phonetic representations input thereto of desired sounds to select from the store units of speech waveform representing portions of words corresponding to the desired sounds;
 - means for identifying voiced portions of the selected units
 - means for concatenating the selected units of speech waveform;
- 10 characterised by means arranged to adjust the amplitude of the voiced portions of the units relative to a predetermined reference level and to leave unchanged at least part of any unvoiced portion of the unit.
2. A speech synthesiser according to Claim 1 in which the adjusting means is arranged to scale the or each voiced portion by a respective scaling factor, and
- 15 to scale the adjacent part of any abutting unvoiced portion by a factor which varies monotonically over the duration of that part between the scaling factor and unity.
3. A speech synthesiser according to Claim 1 or 2 in which a plurality of reference levels is used, the adjusting means being arranged for each voiced
- 20 portion, to select a reference level in dependence upon the sound represented by that portion.
4. A speech synthesiser according to Claim 3 in which each phoneme is assigned a reference level and any voiced portion containing waveform segments from more than one phoneme is assigned a reference level which is a weighted
- 25 sum of the levels assigned to the phonemes contained therein, weighted according to the relative durations of the segments.

Fig.1.

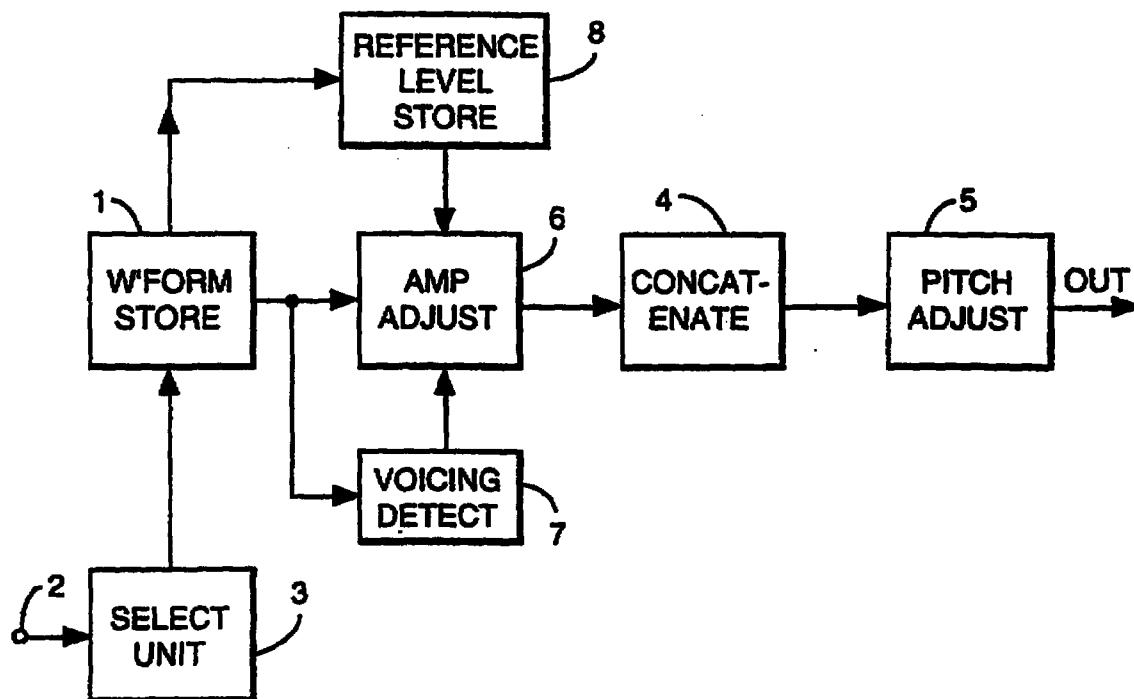


Fig.2.

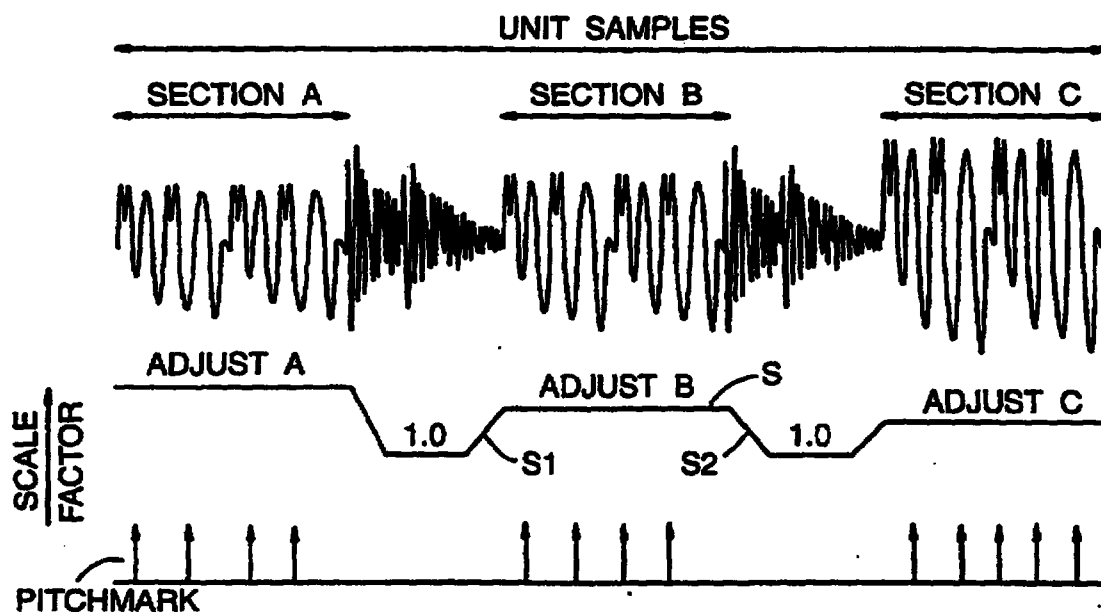
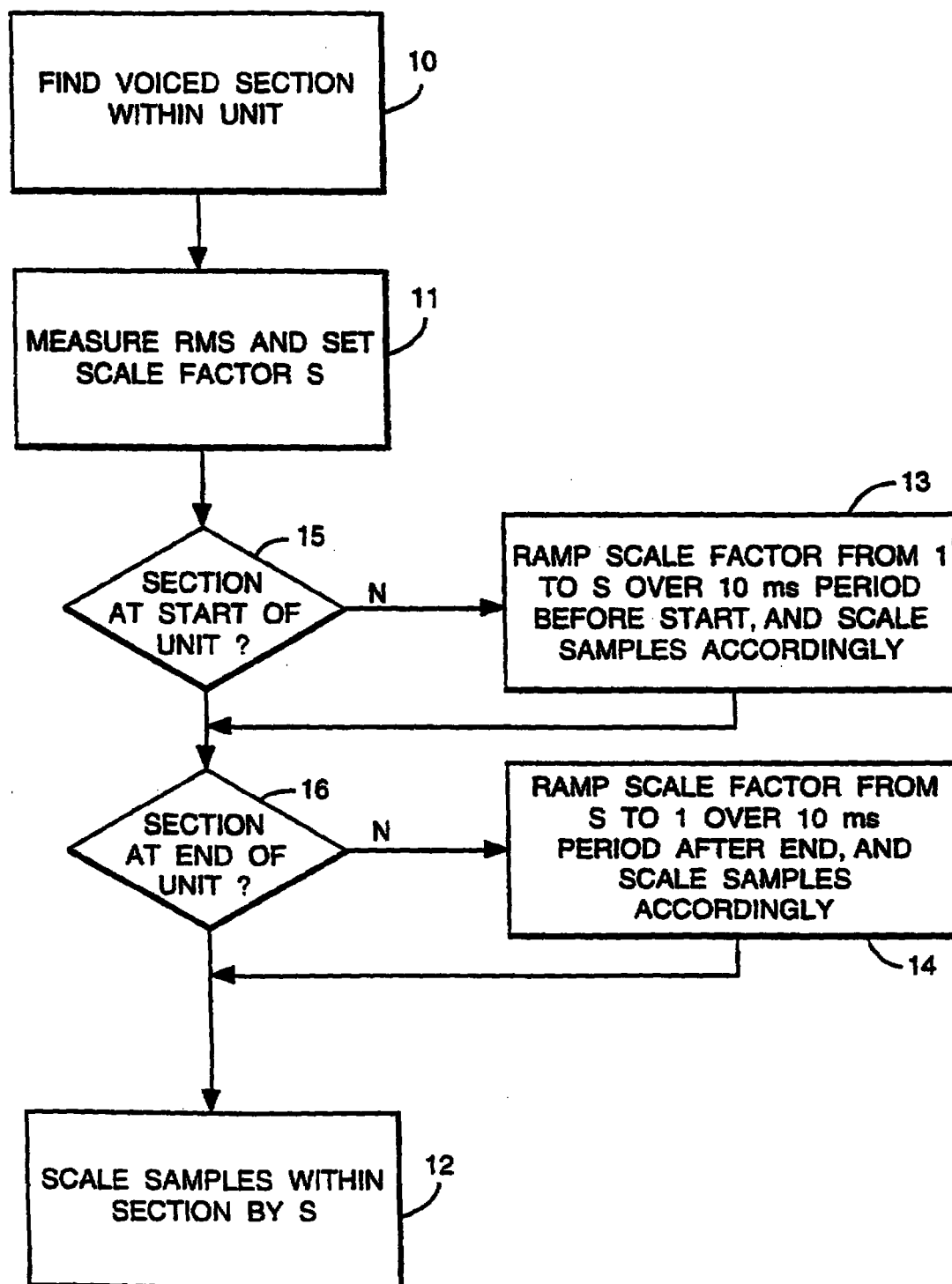


Fig.2.



REPLACEMENT SHEET

28 FEB 1997

1

SPEECH SYNTHESIS

One method of synthesising speech involves the concatenation of small units of speech in the time domain. Thus representations of speech waveform
5 may be stored, and small units such as phonemes, diphones or triphones - i.e. units of less than a word - selected according to the speech that is to be synthesised, and concatenated. Following concatenation, known techniques may be employed to adjust the composite waveform to ensure continuity of pitch and signal phase. However, another factor affecting the perceived quality of the
10 resulting synthesised speech is the amplitude of the units; preprocessing of the waveforms - i.e. adjustment of amplitude prior to storage - is not found to solve this problem, inter alia because the length of the units extracted from the stored data may vary.

European patent application no. 0 427 485 discloses a speech synthesis
15 apparatus and method in which speech segments are concatenated to provide synthesised speech corresponding to input text. The segments used are so-called VCV (vowel-consonant-vowel) segments and the power of the vowels brought adjacent to one another in the concatenation is normalised to a stored reference power for that vowel.

20 An article entitled 'Speech synthesis by linear interpolation of spectral parameters between dyad boundaries' by Shadle et. al. and published in the Journal of the Acoustics Society of America, vol. 66, no. 5, November 1979, New York, US, describes the degradation caused by interpolating spectral parameters over dyad boundaries in synthesising speech.

25 According to the present invention there is provided a speech synthesiser comprising

- a store containing representations of speech waveform;
- selection means responsive in operation to phonetic representations input thereto of desired sounds to select from the store units of speech waveform
30 representing portions of words corresponding to the desired sounds;
- means for concatenating the selected units of speech waveform characterised by means for adjusting the amplitude of at least the voiced portion relative to a predetermined reference level.

AMENDED SHEET

28 FEB 1997

One example of the invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 1 is a block diagram of one example of speech synthesis according to the invention;

- 5 Figure 2 is a flow chart illustrating operation of the synthesis; and
 Figure 3 is a timing diagram.

In the speech synthesiser of Figure 1, a store 1 contains speech waveform sections generated from a digitised passage of speech, originally recorded by a human speaker reading a passage (of perhaps 200 sentences) selected to contain
10 all possible (or at least, a wide selection of) different sounds. Accompanying each section is stored data defining "pitchmarks" indicative of points of glottal closure in the signal, generated in conventional manner during the original recording.

An input signal representing speech to be synthesised, in the form of a phonetic representation is supplied to an input 2. This input may if wished be
15 generated from a text input by conventional means (not shown). This input is processed in known manner by a selection unit 3 which determines, for each unit of the input, the addresses in the store 1 of a stored waveform section corresponding to the sound represented by the unit. The unit may, as mentioned above, be a phoneme, diphone, triphone or other sub-word unit, and in general the
20 length of a unit may vary according to the availability in the waveform store of a corresponding waveform section.

The units, once read out, are concatenated at 4 and the concatenated waveform subjected to any desired pitch adjustments at 5.

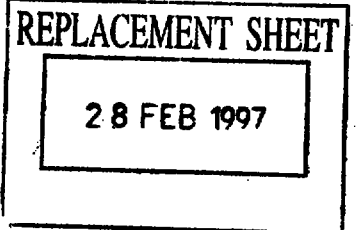
Prior to this concatenation, each unit is individually subjected to an
25 amplitude normalisation process in an amplitude adjustment unit 6 whose operation will now be described in more detail. The basic objective is to normalise each voiced portion of the unit to a fixed RMS level before any further processing is applied. A label representing the unit selected allows the reference level store 8 to determine the appropriate RMS level to be used in the normalisation process.
30 Unvoiced portions are not adjusted, but the transitions between voiced and unvoiced portions may be smoothed to avoid sharp discontinuities. The motivation for this approach lies in the operation of the unit selection and concatenation procedures. The units selected are variable in length, and in the context from

28 FEB 1997

which they are taken. This makes preprocessing difficult, as the length, context and voicing characteristics of adjoining units affect the merging algorithm, and hence the variation of amplitude across the join. This information is only known at run-time as each unit is selected. Postprocessing after the merge is equally
5 difficult.

The first task of the amplitude adjustment unit is to identify the voiced portions(s) (if any) of the unit. This is done with the aid of a voicing detector 7 which makes use of the pitch timing marks indicative of points of glottal closure in the signal, the distance between successive marks determining the fundamental
10 frequency of the signal. The data (from the waveform store 1) representing the timing of the pitch marks are received by the voicing detector 7 which, by reference to a maximum separation corresponding to the lowest expected fundamental frequency, identifies voiced portions of the unit by deeming a succession of pitch marks separated by less than this maximum to constitute a
15 voiced portion. A voiced portion whose first (or last) pitchmark is within this maximum of the beginning (or end) of the speech unit is, respectively, considered to begin at the beginning of the unit or end at the end of the unit. This identification step is shown as step 10 in the flowchart shown in Figure 2.

The amplitude adjustment unit 6 then computes (step 11) the RMS value
20 of the waveform over the voiced portion, for example the portion B shown in the timing diagram of Figure 3, and a scale factor S equal to a fixed reference value divided by this RMS value. The fixed reference value may be the same for all speech portions, or more than one reference value may be used specific to particular subsets of speech portions. For example, different phonemes may be
25 allocated different reference values. If the voiced portion occurs across the boundary between two different subsets, then the scale factor S can be calculated as a weighted sum of each fixed reference value divided by the RMS value. Appropriate weights are calculated according to the proportion of the voiced portion which falls within each subset. All sample values within the voiced portion
30 are (step 12 of Figure 2) multiplied by the scale factor S . In order to smooth voiced/unvoiced transitions, the last 10ms of unvoiced speech samples prior to the voiced portion are multiplied (step 13) by a factor S_1 which varies linearly from 1 to S over this period. Similarly, the first 10ms of unvoiced speech samples



4

following the voiced portion are multiplied (step 14) by a factor S_2 which varies linearly from S to 1. Tests 15, 16 in the flowchart ensure that these steps are not performed when the voiced portion respectively starts or ends at the unit boundary.

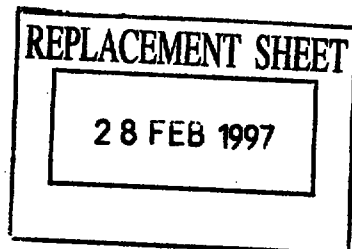
5

Figure 3 shows the scaling procedure for a unit with three voiced portions A, B, C, D, separated by unvoiced portions. Portion A is at the start of the unit, so it has no ramp-in segment, but has a ramp-out segment. Portion B begins and ends within the unit, so it has a ramp-in and ramp-out segment. Portion C starts
10 within the unit, but continues to the end of the unit, so it has a ramp-in, but no ramp-out segment.

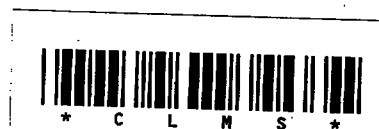
This scaling process is understood to be applied to each voiced portion in turn, if more than one is found.

Although the amplitude adjustment unit may be realised in dedicated
15 hardware, preferably it is formed by a stored program controlled processor operating in accordance with the flowchart of Figure 2.

AMENDED SHEET



5



CLAIMS

1. A speech synthesiser comprising
 - a store containing representations of speech waveform;
 - 5 - selection means responsive in operation to phonetic representations input thereto of desired sounds to select from the store units of speech waveform representing portions of words corresponding to the desired sounds;
 - means for identifying voiced portions of the selected units
 - means for concatenating the selected units of speech waveform;
- 10 characterised by means arranged to adjust the amplitude of the voiced portions of the units relative to a predetermined reference level and to leave unchanged at least part of any unvoiced portion of the unit.
2. A speech synthesiser according to claim 1 wherein said units of the speech waveform vary between phonemes, diphones, triphones and other sub-
15 word units.
3. A speech synthesiser according to Claim 1 in which the adjusting means is arranged to scale the or each voiced portion by a respective scaling factor, and to scale the adjacent part of any abutting unvoiced portion by a factor which varies monotonically over the duration of that part between the scaling factor and unity.
- 20 4. A speech synthesiser according to Claim 1 or 3 in which a plurality of reference levels is used, the adjusting means being arranged for each voiced portion, to select a reference level in dependence upon the sound represented by that portion.
5. A speech synthesiser according to Claim 4 in which each phoneme is
25 assigned a reference level and any voiced portion containing waveform segments from more than one phoneme is assigned a reference level which is a weighted sum of the levels assigned to the phonemes contained therein, weighted according to the relative durations of the segments.

8-035-107

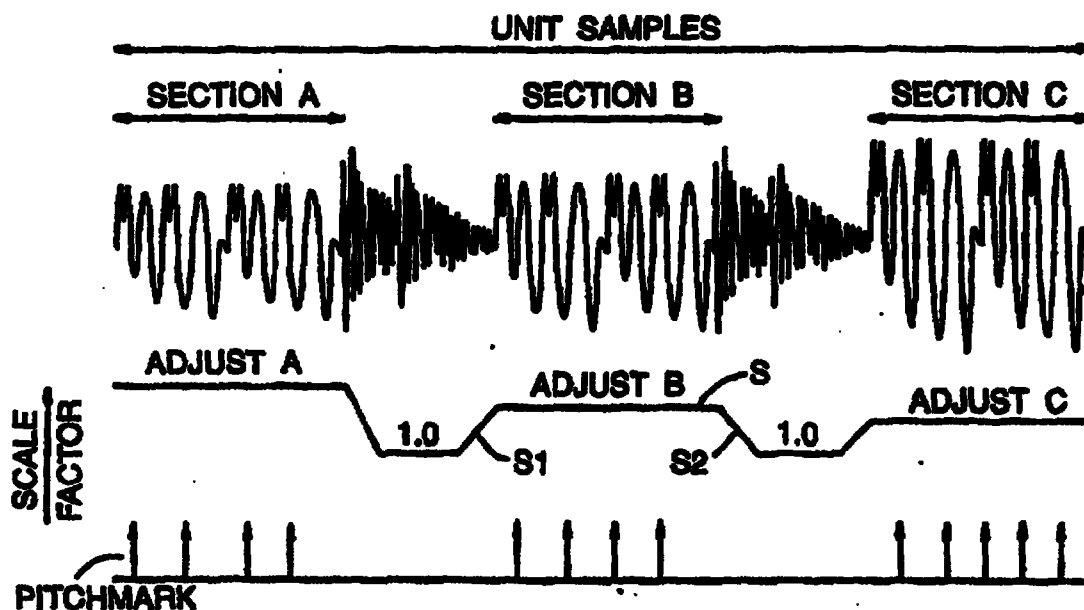
AMENDED SHEET

9703396-3

**PCT**WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G10L 5/04	A1	(11) International Publication Number: WO 96/27870 (43) International Publication Date: 12 September 1996 (12.09.96)
(21) International Application Number: PCT/GB96/00529 (22) International Filing Date: 7 March 1996 (07.03.96) (30) Priority Data: 95301478.4 7 March 1995 (07.03.95) EP (34) Countries for which the regional or international application was filed: GB et al. (71) Applicant (for all designated States except US): BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB). (72) Inventors; and (73) Inventors/Applicants (for US only): LOWRY, Andrew [GB/GB]; 27 Ranelagh Road, Ipswich, Suffolk IP2 0AD (GB). BREEN, Andrew [GB/GB]; 50 Westerfield Road, Ipswich, Suffolk IP4 2UT (GB). JACKSON, Peter [GB/GB]; 36 Manor Road, Martlesham Heath, Ipswich, Suffolk IP5 7SY (GB). (74) Agent: LLOYD, Barry, George, William; BT Group Legal Services, Intellectual Property Dept., 8th floor, 120 Holborn, London EC1N 2TE (GB).	(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG). Published With international search report.	
(54) Title: SPEECH SYNTHESIS		



(57) Abstract

Portions of recorded speech waveform (e.g. corresponding to phonemes) are combined to synthesise words. In order to provide a smoother delivery, each voiced portion of a waveform portion has its amplitude adjusted to a predetermined reference level. The scaling factor used is varied gradually over a transition region between such portions and between voiced and unvoiced portions.