

WAVEFORM SPEECH SYNTHESIS

The present invention relates to speech synthesis, and is particularly
5 concerned with speech synthesis in which stored segments of digitised waveforms
are retrieved and combined.

According to the present invention there is provided a method of speech
synthesis comprising the steps of:

retrieving a first sequence of digital samples corresponding to a first
10 desired speech waveform and first pitch data defining excitation instants of the
waveform;

retrieving a second sequence of digital samples corresponding to a second
desired speech waveform and second pitch data defining excitation instants of the
second waveform;

15 forming an overlap region by synthesising from at least one sequence an
extension sequence, the extension sequence being pitch adjusted to be
synchronous with the excitation instants of the respective other sequence;

forming for the overlap region weighted sums of samples of the original
sequence(s) and samples of the extension sequence(s).

20 In another aspect of the invention provides an apparatus for speech
synthesis comprising the steps of:

means storing sequences of digital samples corresponding to portions of
speech waveform and pitch data defining excitation instants of those waveforms;

control means controllable to retrieve from the store means 1 sequences
25 of digital samples corresponding to desired portions of speech waveform and the
corresponding pitch data defining excitation instants of the waveform;

means for joining the retrieved sequences, the joining means being
arranged in operation (a) to synthesise from at least the first of a pair of retrieved
sequences an extension sequence to extend that sequence into an overlap region
30 with the other sequence of the pair, the extension sequence being pitch adjusted
to be synchronous with the excitation instants of that other sequence and (b) to
form for the overlap region weighted sum of samples of the original sequence(s)
and samples of the extension sequence(s).



Other aspects of the invention are defined in the sub-claims.

Some embodiments 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 form of speech synthesiser in accordance with the invention;

Figure 2 is a flowchart illustrating the operation of the joining unit 5 of the apparatus of Figure 1; and

Figure 3 to 9 are waveform diagrams illustrating the operation of the joining unit 5.

10 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 all possible (or at least, a wide selection of) different sounds. Thus each entry in the waveform store 1 comprises digital samples of a portion of speech corresponding to one or more phonemes, with marker information indicating the boundaries between the phonemes. 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.

20 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 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 length of a unit may vary according to the availability in the waveform store of a corresponding waveform section. Where possible, it is preferred to select a unit which overlaps a preceding unit by one phoneme. Techniques for achieving this are described in our CO-pending International patent application no. PCT/GB/9401688 and US patent application no. 166,988 of 16 December 1993.

30 The units, once read out, are each individually subjected to an amplitude normalisation process in an amplitude adjustment unit 4 whose operation is described in our co-pending European patent application no. 95301478.4.

The units are then to be joined together, at 5. A flowchart for the operation of this device is shown in Figure 2. In this description a unit and the unit which follows it are referred to as the left unit and right unit respectively. Where the units overlap - i.e. when the last phoneme of the left unit and the first phoneme of the right unit are to represent the same sound and form only a single phoneme in the final output - it is necessary to discard the redundant information, prior to making a "merge" type join; otherwise an "abut" type join is appropriate.

In step 10 of Figure 2, the units are received, and according to the type of merge (step 11) truncation is or is not necessary. In step 12, the corresponding pitch arrays are truncated; in the array corresponding to the left unit, the array is cut after the first pitchmark to the right of the mid-point of the last phoneme so that all but one of the pitchmarks after the mid-point are deleted whilst in the array for the right unit, the array is cut before the last pitchmark to the left of the mid-point of the first phoneme so that all but one of the pitchmarks before the mid-point are deleted. This is illustrated in Figure 2.

Before proceeding further, the phonemes on each side of the join need to be classified as voiced or non-voiced, based on the presence and position of the pitchmarks in each phoneme. Note that this takes place (in step 13) after the "pitch cutting" stage, so the voicing decision reflects the status of each phoneme after the possible removal of some pitchmarks. A phoneme is classified as voiced if:

1. the corresponding part of the pitch array contains two or more pitchmarks; and
2. the time difference between the two pitchmarks nearest the join is less than a threshold value; and
 - 3a. for a merge type join, the time difference between the pitchmark nearest the join and the midpoint of the phoneme is less than a threshold value;
 - 3b. for an abut type join, the time difference between the pitchmark nearest the join and the end of the left unit (or the beginning of the right unit) is less than a threshold value.

Otherwise it is classified as unvoiced.

Rules 3a and 3b are designed to prevent excessive loss of speech samples in the next stage.

In the case of a merge type join (step 14), speech samples are discarded (step 15) from voiced phonemes as follows:

- 5 Left unit, last phoneme - discard all samples following the last pitchmark ;
 Right unit, first phoneme - discard all samples before the first pitchmark;
and from unvoiced phonemes by discarding all samples to the right or left of the midpoint of the phoneme (for left and right units respectively).

In the case of an abut type join (steps 16, 15), the unvoiced phonemes
10 have no samples removed whilst the voiced phonemes are usually treated in the same way as for the merge case, though fewer samples will be lost as no pitchmarks will have been deleted. In the event that this would cause loss of an excessive number of samples (e.g. more than 20 ms) then no samples are removed and the phoneme is marked to be treated as unvoiced in further processing.

- 15 The removal of samples from voiced phonemes is illustrated in Figure 3. The pitchmark positions are represented by arrows. Note that the waveforms shown are for illustration only and are not typical of real speech waveforms.

The procedure to be used for joining two phonemes is an overlap-add process. However a different procedure is used according to whether (step 17)
20 both phonemes are voiced (a voiced join) or one or both are unvoiced (unvoiced join).

The voiced join (step 18) will be described first. This entails the following basic steps: the synthesis of an extension of the phoneme by copying portions of its existing waveform but with a pitch period corresponding to the other phoneme
25 to which it is to be joined. This creates (or, in the case of a merge type join, recreates) an overlap region with, however, matching pitchmarks. The samples are then subjected to a weighted addition (step 19) to create a smooth transition across the join. The overlap may be created by extension of the left phoneme, or of the right phoneme, but the preferred method is to extend both the left and the
30 right phonemes, as described below. In more detail:

1. a segment of the existing waveform is selected for the synthesis, using a Hanning window. The window length is chosen by looking at the last two pitch periods in the left unit and the first two pitch periods in the

right unit to find the smallest of these four values. The window width - for use on both sides of the join - is set to be twice this.

2. the source samples for the window period, centred on the penultimate pitchmark of the left unit or the second of the right unit, are extracted and multiplied by the Hanning window function, as illustrated in Figure 4. Shifted versions, at positions synchronous with the other phoneme's pitchmarks, are added to produce the synthesised waveform extension. This is illustrated in Figure 5. The last pitch period of the left unit is multiplied by half the window function and then the shifted, windowed segments are overlap added at the last original pitchmark position, and successive pitchmark positions of the right unit. A similar process takes place for the right unit.

3. the resulting overlapping phonemes are then merged; each is multiplied by a half Hanning window of length equal to the total length of the two synthesised sections as depicted in Figure 6, and the two are added together (with the last pitchmark of the left unit aligned with the first pitchmark of the right); the resulting waveform should then show a smooth transition from the left phoneme's waveform to that of the right, as illustrated in Figure 7.

4. the number of pitch periods of overlap for the synthesis and merge process is determined as follows. The overlap extends into the time of the other phoneme until one of the following conditions occurs -

- (a) the phoneme boundary is reached;
- (b) the pitch period exceeds a defined maximum;
- (c) the overlap reaches a defined maximum (e.g. 5 pitch periods).

If however condition (a) would result in the number of pitch periods falling below a defined minimum (e.g. 3) it may be relaxed to allow one extra pitch period.

An unvoiced join is performed, at step 20, simply by shifting the two units temporally to create an overlap, and using a Hanning weighted overlap-add, as shown in step 21 and in Figure 8. The overlap duration chosen is, if one of the phonemes is voiced, the duration of the voiced pitch period at the join, or if they are both unvoiced, a fixed value

[typically 5ms]. The overlap (for abut) should however not exceed half the length of the shorter of the two phonemes. It should not exceed half the remaining length if they have been cut for merging. Pitchmarks in the overlap region are discarded. For an abut type join, the boundary between the two phonemes is considered, for the purposes of later processing, to lie at the mid-point of the overlap region.

Of course, this method of shifting to create the overlap shortens the duration of the speech. In the case of the merge join, this can be avoided by "cutting" when discarding samples not at the midpoint but slightly to one side so that when the phonemes have their (original) mid-points aligned an overlap results.

The method described produces good results; however the phasing between the pitchmarks and the stored speech waveforms may - depending on how the former were generated - vary. Thus, although pitch marks are synchronised at the join this does not guarantee a continuous waveform across the join. Thus it is preferred that the samples of the right unit are shifted (if necessary) relative to its pitchmarks by an amount chosen so as to maximise the cross-correlation between the two units in the overlap region. This may be performed by computing the cross-correlation between the two waveforms in the overlap region with different trial shifts (e.g. ± 3 ms in steps of 125 μ s). Once this has been done, the synthesis for the extension of the right unit should be repeated.

After joining, an overall pitch adjustment may be made, in conventional manner, as shown at 6 in Figure 1.

The joining unit 5 may be realised in practice by a digital processing unit and a store containing a sequence of program instructions to implement the above-described steps.

CLAIMS

1. A method of speech synthesis comprising the steps of:
- 5 retrieving a first sequence of digital samples corresponding to a first desired speech waveform and first pitch data defining excitation instants of the waveform;
- retrieving a second sequence of digital samples corresponding to a second desired speech waveform and second pitch data defining excitation instants of the second waveform;
- 10 forming an overlap region by synthesising from at least one sequence an extension sequence, the extension sequence being pitch adjusted to be synchronous with the excitation instants of the respective other sequence;
- forming for the overlap region weighted sums of samples of the original sequence(s) and samples of the extension sequence(s).
- 15
2. A method of speech synthesis comprising the steps of:
- retrieving a first sequence of digital samples corresponding to a first desired speech waveform and first pitch data defining excitation instants of the waveform;
- 20 retrieving a second sequence of digital samples corresponding to a second desired speech waveform and second pitch data defining excitation instants of the second waveform;
- synthesising from the first sequence an extension sequence at the end of the first sequence, the extension sequence being pitch adjusted to be synchronous with the excitation instants of the second sequence;
- 25 synthesising from the second sequence an extension sequence at the beginning of the second sequence, the extension sequence being pitch adjusted to be synchronous with the excitation instants of the first sequence;
- 30 whereby the first and second extension sequences define an overlap region;



forming for the overlap region weighted sums of samples of the first sequence and samples of the second extension sequence and weighted sums of samples of the second sequence and samples of the first extension sequence.

5

3. A method according to claim 2 in which the first sequence has a portion at the end thereof corresponding to a particular sound and the second sequence has a portion at the beginning thereof corresponding to the same sound, and including the step of, prior to the synthesis, removing samples from the end of the said portion of the first waveform and from the beginning of the said portion of the second waveform.

10

4. A method according to claim 1, 2 or 3 in which each synthesis step comprises extracting from the relevant sequence a subsequence of samples, multiplying the subsequence by a window function and repeatedly adding the subsequences with shifts corresponding to the excitation instants of the other one of the first and second sequences.

15

5. A method according to claim 4 in which the window function is centred on the penultimate excitation instant of the first sequence and on the second excitation instant of the second sequence and has a width equal to twice the minimum of selected pitch periods of the first and second sequences, where a pitch period is defined as the interval between excitation instants.

20

25

6. A method according to any one of the preceding claims including the steps of, prior to forming the weighted sums, comparing, over the overlap region, the first sequence and its extension with the second sequence and its extension to derive a shift value which maximises the correlation therebetween, adjusting the second pitch data by the determined shift amount and repeating the synthesis of the second extension sequence.

30



7. An apparatus for speech synthesis comprising the steps of:
means (1) storing sequences of digital samples corresponding to
portions of speech waveform and pitch data defining excitation instants of
those waveforms;

5 control means (2) controllable to retrieve from the store means 1
sequences of digital samples corresponding to desired portions of speech
waveform and the corresponding pitch data defining excitation instants of
the waveform;

10 means (5) for joining the retrieved sequences, the joining means
being arranged in operation (a) to synthesise from at least the first of a
pair of retrieved sequences an extension sequence to extend that
sequence into an overlap region with the other sequence of the pair, the
extension sequence being pitch adjusted to be synchronous with the
excitation instants of that other sequence and (b) to form for the overlap
15 region weighted sum of samples of the original sequence(s) and samples
of the extension sequence(s).

Fig. 1.

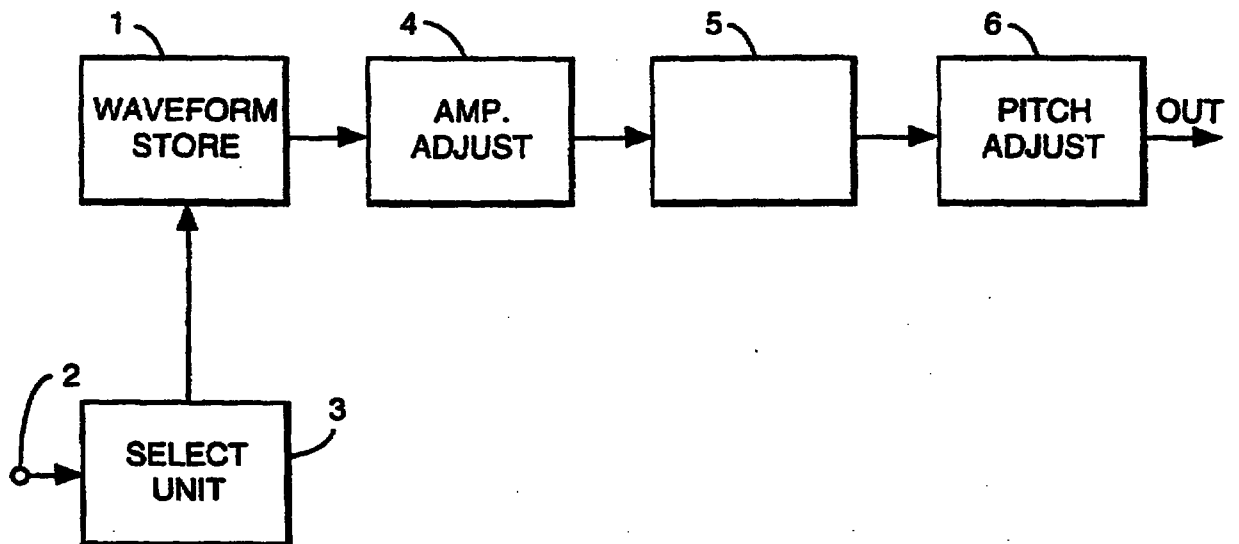
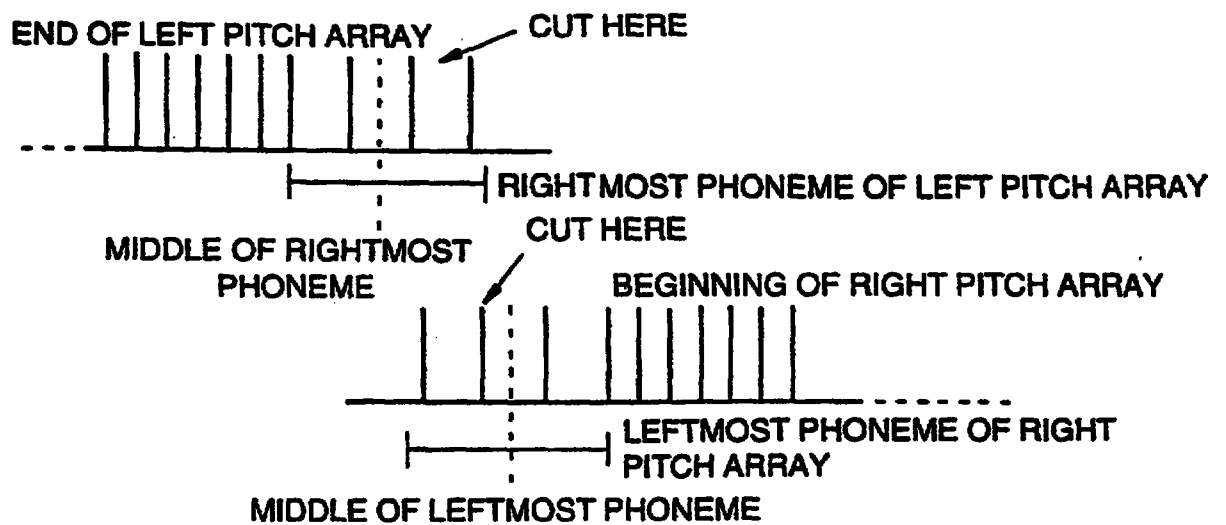


Fig. 3



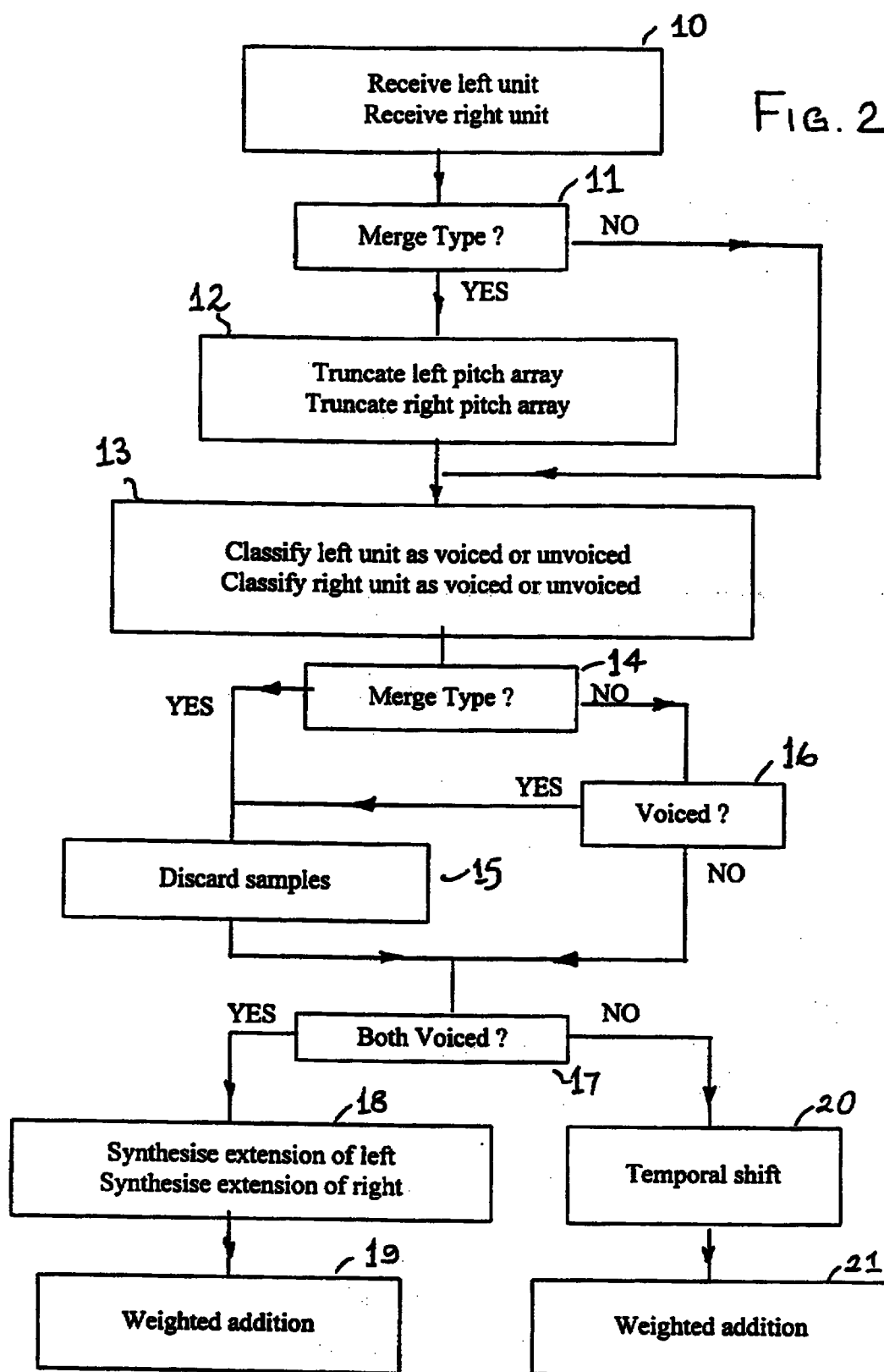


Fig. 7

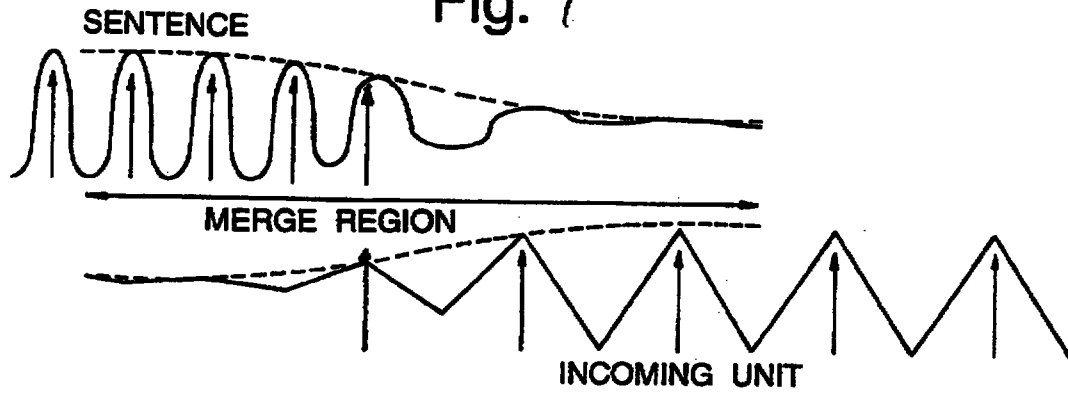


Fig. 8

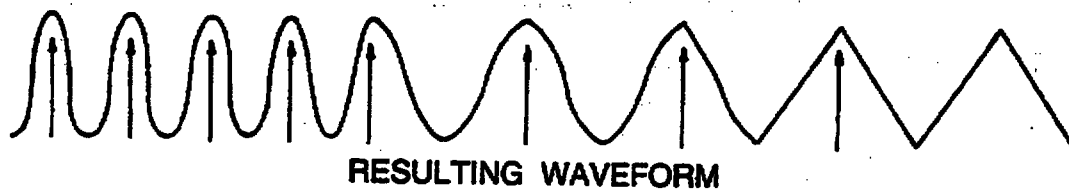
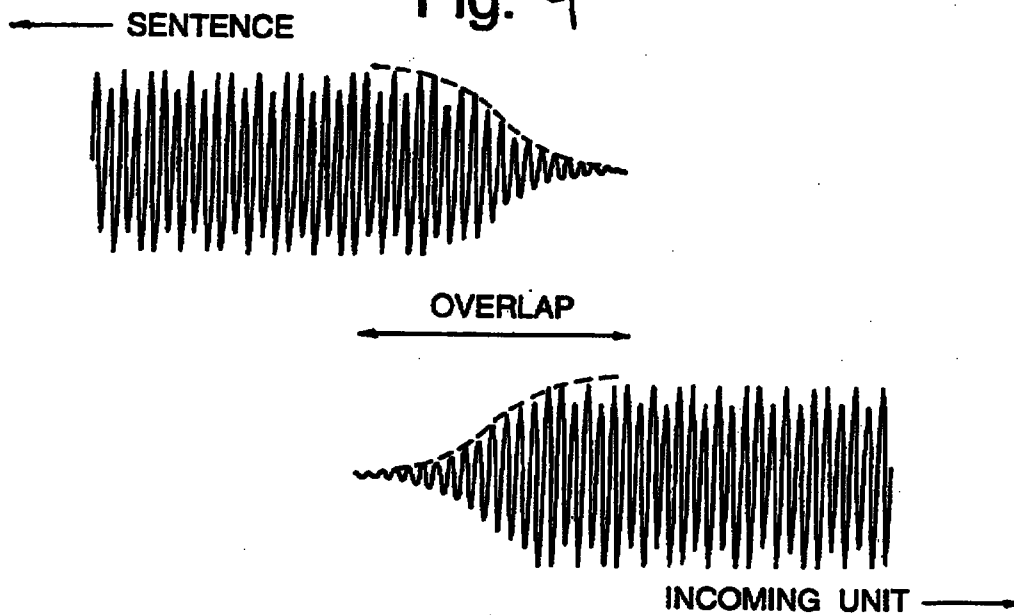
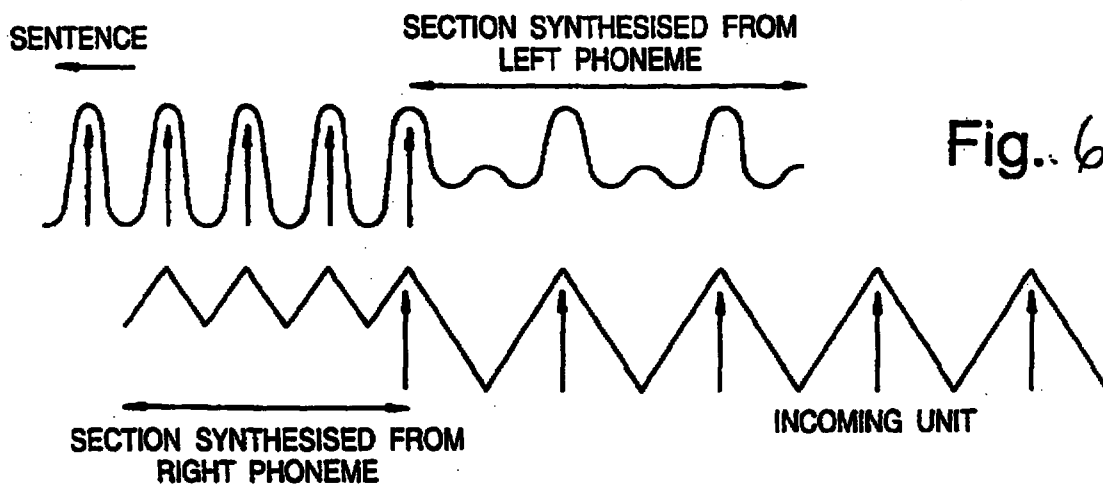
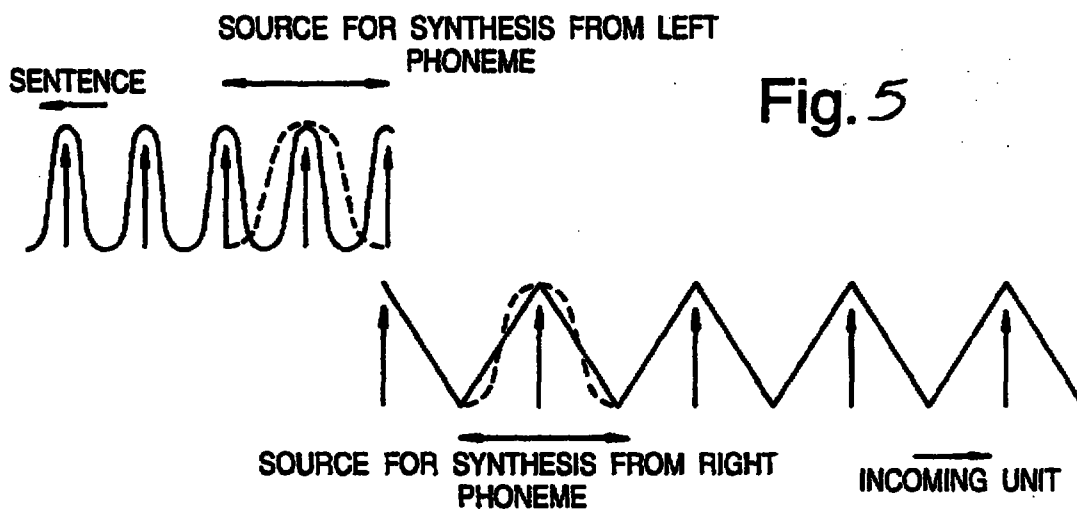
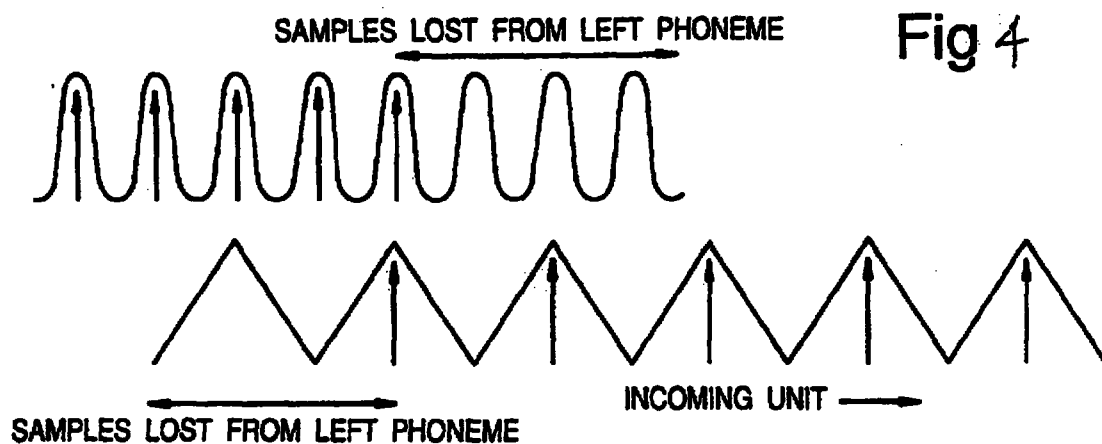


Fig. 9





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 96/32711 (43) International Publication Date: 17 October 1996 (17.10.96)
(21) International Application Number: PCT/GB96/00817 (22) International Filing Date: 3 April 1996 (05.04.96) (30) Priority Data: 95302474.2 12 April 1995 (12.04.95) EP (34) Countries for which the regional or international application was filed: AT 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) Inventor; and (75) Inventor/Applicant (for US only): LOWRY, Andrew [GB/GB]; 27 Ranelagh Road, Ipswich, Suffolk IP2 0AD (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: WAVEFORM SPEECH SYNTHESIS 		
(57) Abstract Portions of speech waveform are joined by forming extrapolations at the end of one and the beginning of the next portion to create an overlap region with synchronous pitchmarks, and then forming a weighted sum across the overlap to provide a smooth transition.		

WAVEFORM SPEECH SYNTHESIS

The present invention relates to speech synthesis, and is particularly
5 concerned with speech synthesis in which stored segments of digitised waveforms
are retrieved and combined.

An example of a speech synthesiser in which stored segments of digitised
waveforms are retrieved and combined is described in a paper by Tomohisa
Hirokawa et al entitled "High Quality Speech Synthesis System Based on
10 Waveform Concatenation of Phoneme Segment" in the IEICE Transactions on
Fundamentals of Electronics, Communications and Computer Sciences 76a(1993)
November, No.11, Tokyo, Japan.

According to the present invention there is provided a method of speech
synthesis comprising the steps of:

15 retrieving a first sequence of digital samples corresponding to a first
desired speech waveform and first pitch data defining excitation instants of the
waveform;

retrieving a second sequence of digital samples corresponding to a second
desired speech waveform and second pitch data defining excitation instants of the
20 second waveform;

forming an overlap region by synthesising from at least one sequence an
extension sequence, the extension sequence being pitch adjusted to be
synchronous with the excitation instants of the respective other sequence;

25 forming for the overlap region weighted sums of samples of the original
sequence(s) and samples of the extension sequence(s).

In another aspect of the invention provides an apparatus for speech
synthesis comprising the steps of:

means storing sequences of digital samples corresponding to portions of
speech waveform and pitch data defining excitation instants of those waveforms;

30 control means controllable to retrieve from the store means 1 sequences
of digital samples corresponding to desired portions of speech waveform and the
corresponding pitch data defining excitation instants of the waveform;

AMENDED SHEET

means for joining the retrieved sequences, the joining means being arranged in operation (a) to synthesise from at least the first of a pair of retrieved sequences an extension sequence to extend that sequence into an overlap region with the other sequence of the pair, the extension sequence being pitch adjusted
5 to be synchronous with the excitation instants of that other sequence and (b) to form for the overlap region weighted sum of samples of the original sequence(s) and samples of the extension sequence(s).