(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 November 2004 (18.11.2004)

PCT

(10) International Publication Number
WO 2004/100128 A1

(51) International Patent Classification[7]: G10L 21/06, 15/26

(21) International Application Number: PCT/IB2004/001349

(22) International Filing Date: 15 April 2004 (15.04.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PI 20031466 18 April 2003 (18.04.2003) MY

(71) Applicant *(for all designated States except US)*: **UNISAY SDN. BHD.** [MY/MY]; Suite 301, Block A Phileo Damansara 1, 9 Jalan 16/11 Petaling Jaya, Selangor DE, 46350 (MY).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **MUTUAL, William** [CA/CA]; #501, 1375 Nicola Street, Vancouver, British Columbia V6G 2G1 (CA).

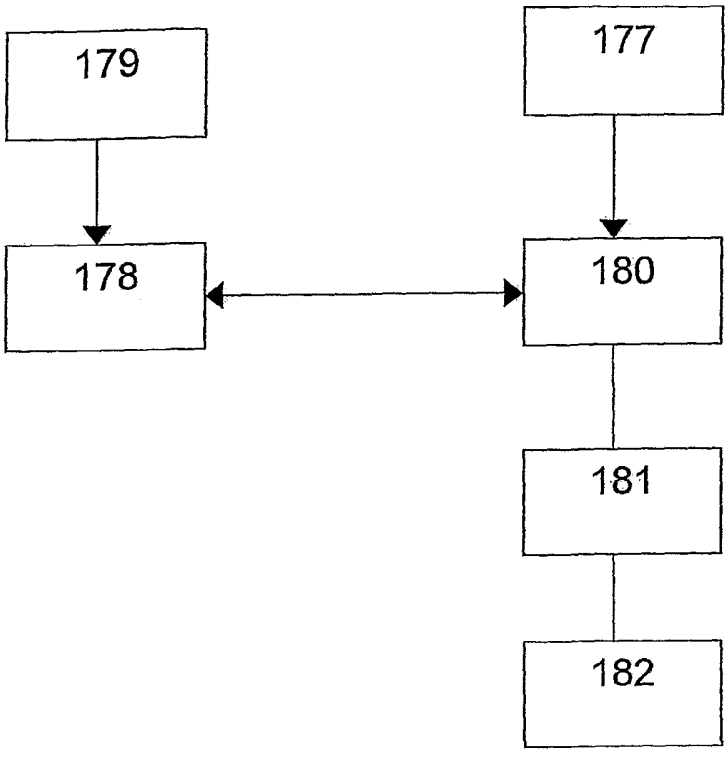
(74) Common Representative: **MUTUAL, William**; c/o ALBAN TAY MAHTANI & DE SILVA, 39 Robinson Road #07-01 Robinson Point, 068911 Singapore (SG).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: SYSTEM FOR GENERATING A TIMED PHOMEME AND VISEM LIST



(57) Abstract: Timed Language Presentation System A system for generating a timed language presentation including receiving an audio/video input and a script input, the script corresponding to an audio track of the audio/video input; separating the audio/video input into the audio track and a separate video stream, only the audio track being used for processing; conducting a track reading process using a text-to-speech process on the script to produce a description of words and phoneme length timings, analysing the audio track to produce a time location for words in the audio track; and processing the description of words, phoneme length timings and time location to produce a timed word/phoneme list.

WO 2004/100128 A1

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM FOR GENERATING A TIMED PHOMEME AND VISEM LIST

Field of the Invention

The present invention relates to a timed phoneme and viseme presentation system and refers particularly, though not exclusively, to such a system for producing timed phonemes and visemes for animation and special effects production.

Reference to related Applications

Reference is made to our co-pending patent applications filed contemporaneously herewith and titled: ""Phoneme Extraction System", "Timed Language Presentation System", "Voice Script System", and "Process for Adding Subtitles to Video Content", the contents of which are hereby incorporated by reference.

Background to the Invention

In producing an animated motion picture, either as a separate production, or as special effects of a motion picture, the audio track is recorded first. The audio track is then manually timed and the script broken down into individual words. A timed script is then produced so the animator knows he has a certain time, and therefore a certain number of frames, for the animated creature to say particular word, make an utterance, make a sound, and so forth. He can then prepare the relevant drawings showing the lip or other movement for the sound, utterance or word, so it will be as close as possible to being in sync with the audio track.

This process is very time consuming, and costly.

Definitions

Throughout this specification a reference to a "machine" (and its grammatical variants) is to be taken as including a reference to one or more of: server, desktop computer, personal computer, laptop computer, notebook computer, tablet computer, and personal digital assistant.

Throughout this specification a reference to a phrase is to be taken as including a reference to a clause.

Summary of the Invention

According to one aspect of the invention there is provided a system for generating a timed phoneme and viseme list comprising the steps:

(a) creating a lookup table of phonemes used and their corresponding visemes;
(b) receiving a timed phoneme list;
(c) analysing the lookup table to extract the visemes corresponding to each phoneme in the timed phoneme list; and
(d) adding the visemes to the timed phoneme list to give a timed phoneme and viseme list.

The phoneme list may include words. Preferably, the timed phoneme list is produced by receiving an audio track input and a script input, the script corresponding to the audio track input; conducting a track reading process using a text-to-speech process on the script to produce a description of words, phonemes and phoneme length timings, analysing the audio track to produce a time location for words in the audio track; and processing the description of words, phonemes, phoneme length timings and time location to produce a timed word/phoneme list.

The track may be part of an audio/video input that is preferably converted to be MPEG1/MPEG2 compliant then the audio track separated.

Before the final process step is performed, an output from the track reading process may be further processed to be MPEG 7 complaint. The timed phonemes may be subjected to further processing to provide a speaker phoneme set prior to input into the final process step.

In the final process step the analysis may be a phoneme analysis. The phoneme analysis may include an MPEG 7 word/phoneme analysis using sound spectrum and sound waveforms.

The audio stream may also be input into an MPEG 7 audio low level descriptors extraction process; as well as into an MPEG 7 emotive analysis for extraction of emotion-related descriptors.

A phrase of the script may be generated using the phonemes, and a recognition process conducted to recognise the phrase in the audio track after processing of the audio track in a speech recognition process. The track reading process may first break the script into words. Any unrecognised phrases in the audio track after the recognition process may subsequently be processed in the same manner with reset threshold and accuracy of the speech process. Prior to the speech recognition process parameters and next phrase logic of the speech recognition process may be set.

In another aspect of the present invention there is provided a computer usable medium comprising a computer program code that is configured to cause one or more processors to execute one or more functions to perform the steps and functions as described above.

Description of the Drawings

In order that he invention may be readily understood and put into practical effect, there shall now be described by way of non-limitative example, only a preferred embodiment of the present invention, the description being with reference to the accompanying illustrative drawings in which:

Figure 1 is an overall chart of a preferred embodiment of the present invention;
Figure 2 is a flow chart for the initialising process of Figure 1;
Figure 3 is a flow chart for a preferred process for phoneme extraction;
Figure 4 is a flow chart for a preferred process for timed phoneme extraction;
Figure 5 is a flow chart for audio low level descriptors extraction;
Figure 6 is a flow chart of emotive analysis;
Figure 7 is a flow chart of word/phoneme analysis;
Figure 8 is a flow chart of the recognition process; and
Figure 9 is a flow chart for the viseme extraction process.

Description of the Preferred Embodiment

To refer to Figure 1, there is illustrated the initial part of overall process. At 101 machine of a user connects to the relevant web page to initiate the process. However, the process may be performed on a stand-alone machine.

The audio and, if possible, text files requiring processing for which a timing track is required are selected and at 102 sent to the server operating the web page, for processing. The server performs the necessary processing in 103 and sends the timing track (and other requested outputs) to the machine of the user.

The first stage 101 of the process is illustrated in more detail in Figure 2. As is shown, each of the following is selected:

(a) the script in text format – 104;
(b) the voice track (audio file) – 105;
(c) the number of words to be recognised at the one time – 106;
(d) the output type and the output parameters. This may be either or both words and phonemes – 107, and
(e) output time units as either or both milliseconds and frames 108.

Upon the five selections being made, at 109 the process is initiated. At 110 a preliminary check is made to ensure all parameters are correct. If not, at 111 an incorrect parameter message is generated and sent to the user's machine. The five input parameters (a) to (e) can then be reselected, reset or the error corrected. If all parameters are correct, at 112 all data is sent for processing and the input stage ends.

To now refer to Figure 3, there is illustrated in more detail the process for phoneme extraction 103 of Figure 1.

The process commences with a detailed check at 113 to determine if all parameters are correct. This is a more complete test than that conducted at 110 and is also to ensure the transmission from the sender's machine to the server has been error-free. If there is an error, an error message is generated at 114, sent to the sending machine, and all processing stops.

This script is processed at 115 to break it down into individual words. If there is an error, an error message is generated at 114, sent to the sending machine, and all processing stops.

At 116 the words of the script are broken into phonemes using a text-to-speech engine such as, for example, the "SAPI 5.1" as available from Microsoft Corp, or the "Viavoice" as available from IBM Corp. If there is an error, error message is generated at 114, sent to the sending machine, and all processing stops.

A speech recognition engine is used at 117 to analyse the audio/voice track 105. The parameters for the analysis are first set, and the logic for the next phrase of the audio/voice track 105 is also set. A suitable speech recognition engine would be, for example, "SAPI 5.1" as available from Microsoft Corp, or the "Viavoice" as available from IBM Corp. If at the commencement of the audio track and script, the first phrase of the script is generated from 116. For subsequent processing, the next phrase of the script is generated. At the end of the script, there will be nothing left so the processing is diverted at 119 to the next major processing stage. Otherwise, it proceeds to 120 where the present voice track position is set.

The phonemes used come from the script. The phoneme timing, and time location in the audio track of the phoneme come from the audio track. By analysing on a phrase-by-phrase basis from both the script and the audio track, the phonemes can be timed for duration and location with a relatively high degree of accuracy. The overall length of the audio track is known and this will be the overall duration of the script. Therefore, it is possible to estimate an approximate time in the duration of the script where a particular phrase might appear.

Therefore, in process step 121, the current time location is recognised from the audio track as is the current script phrase as appearing on the audio track using speech recognition. By recognising the script phrase on the audio track a "most likely position" for that script phrase can be determined and thus its likely time location on the audio track. Phonemes missing from the script may be obtained from the audio track using the voice properties of the speaker as obtained from the closely matched complete set, or by

synthesising the missing phonemes that don't have a close match. Manual intervention may be required to correct errors depending on the subject of the text, number of slang expressions used, and the diction on the audio track.

The word times are then updated in 122 using the results of the recognition process, and the next phrase of the script prepared. The process then loops back to 118 and repeats until there is nothing left in the script.

When there is nothing left, the process is diverted at 119 and passes to a second stage of processing. Here, any unrecognised phrases in the script are re-examined. If there are no unrecognised phrases, the output from 122 is directly to 129 where the timed word/phoneme list is generated.

If there are unrecognised phrases in the script (normally due to the speech recognition engine producing on unusable result for phrase), in 123 the speech recognition engine threshold and accuracy are set. The first (or next for subsequent phrases) phrase of the script is generated. If there are no more unrecognised phrases, at 125 the process is diverted to 129 to generate the timed word/phoneme list.

If there is an unrecognised phrase, the voice track position is set at 126, as before. The recognition process of 121 is re-performed at 127 but with the new thresholds and accuracy. Again, at 128 the recogniser results are used to update the word times, and the next unrecognised phrase prepared. The process loops back to 124.

When all processing has completed, the output is generated at 129 and the process ends.

The recognition process at 121 is further illustrated in Figure 8.

At the beginning 159 of the recognition process audio 167 is input at 163 for processing in accordance with the MPEG 7 sound classification model audio classifier. This extracts and outputs at 171 the timed sound classifications. Sound classifications include non-word sounds such as, for example, inhalation, laughter, natural sounds, noises, background sounds and noises, music, and so forth.

Audio 166 is input at 117 as described above for speech recognition analysis preferably using the Hidden Markov Method ("HMM") for raw recognition and timing. The output 172 is an aligned raw phoneme list.

The text 165 is input for processing as described above for text-to-speech (phoneme) and the output 173 is a normally timed phonetic text representation.

The audio and text are input at 164 into process 121 as is described above for phrase-segmented, time-assisted, speech recognition. The output 174 is a verified, aligned word/phoneme list.

Outputs 171, 172 and 173 are input to process 168 for initial text alignment for words and the sound classifications. The output 175 from process 168 is estimated word timings and partial phoneme timings. This is also input into 121 for its output 174. The output 175 is also input to process 169, as is output 174, for final text alignment (including some or all of the sound classifications). The output 176 from 169 is a complete timed/word/phoneme list. The recognition process ends at 170.

To now refer to Figure 4, there is shown the initial processing of an audio/video input 8 as received in the machine operating the system of the present invention. At input 8 there is an audio/video input and a script corresponding to the audio track of the audio/video input as a text input in digital form. The audio/video input is preferably in source language. This is process 101 and 102 of Figure 1; and steps 104 to 112 of Figure 3.

The audio/video input is split into two processing streams 4, 6. In stream 4 the audio/video 10 is first transcoded at 12 to MPEG1/MPEG2, if not already in the appropriate format. If in the appropriate format, transcoding is not necessary. Audio/video separation takes place and the video is output at 14 and by-passes all subsequent processing at this stage. The audio/video separation is a standard technique using standard applications/engines.

In stream 6, the script undergoes a track reading process at 18. These are process steps 113 to 122 of Figure 3. Also input at 18 is a time input to enable each of the phrases, words and phonemes to be timed for likely duration.

The time input may be in seconds down to a preferred level of three decimal places (i.e. milliseconds). The output from process 18 is timed phonemes. It may also have marked-up phonemes and descriptions of the words. Preferably, all output to 20 is in MPEG 7 representation. In the stage of processing at 20 the phoneme timing are completed in accordance with process steps 123 to 128 of Figure 3 to have a speaker phoneme set that is as complete as possible. The output 30 from process 20 is passed to data storage 22 for storage and to be used later in other processes.

The audio output 32 from process 12, and the outputs from process steps 18, 20, are all passed for three parallel enhancement process steps.

In the first enhancement step 24, the audio output 32 from process 12 is passed for MPEG 7 audio low level descriptors ("LLD") extraction. LLDs may include such characteristics as the spectrum of the sound, silence, and so forth. The LLD analysis may proceed in parallel to the track reading process 18, or may be subsequent to that process. If subsequent, the output from process 18 may be also input to process 24.

The audio output 32 from process 12 and the result 34 of the track reading process 18 are both input to the second enhancement step 26 . In this step MPEG 7 emotive analysis is performed by extracting of the audio using emotion-related descriptors. An emotive analysis based on MPEG 7 HLDs (High Level Descriptors) is then performed.

The third enhancement step 28 has as its input the text and phoneme timing output 30 of process 20, the audio output 32 from process 12, and a timer 36 . The timer 36 may be in seconds down to a preferred level of three decimal places (i.e. milliseconds) and may start from the commencement of the audio track, or otherwise as prescribed or required. The audio track 32 is combined with the output 30 and the timer 36 and analysed and described into a timed word/phoneme list. The output 30 gives the phonemes and their likely duration. The audio track 32 and the timer 36 give the time location of the phonemes By matching the timed phonemes and the audio track 32, based on a likely match, a timed word/phoneme list can be prepared. This is output 38 from process 28 and stored in data storage 22 for later use. The outputs 40, 42 from processes 26, 24 respectively are also stored in data storage 22 for subsequent use. The processing ends.

Figure 5 is a flow chart of the audio LLD extraction process 24 of Figure 4. Here, the spectrum, and logarithm of the spectrum, are calculated at 130. Some or all of a large number of low level descriptors are then determined in either parallel (as shown) or sequentially. These include, but are not limited to:

131 audio spectrum envelope descriptor;
132 audio spectrum centroid descriptor;
133 audio spectrum spread descriptor
134 audio spectrum flatness descriptor;
135 audio spectrum basis descriptor;
136 audio waveform descriptor;
137 audio power descriptor;
138 silence descriptor;
139 audio spectrum projection descriptor;
140 audio fundamental frequency descriptor;
141 audio harmonicity descriptor;
142 audio spectrum flatness type descriptor;
143 log attack time descriptor;
144 harmonic spectral centroid descriptor
145 harmonic spectral deviation descriptor;
146 harmonic spectral spread descriptor;
147 harmonic spectral variation descriptor;
148 spectral centroid descriptor; and
149 temporal centroid descriptor.

The result is then output as is described above.

In Figure 6 is illustrated the process for emotive high level descriptors extraction. This is process step 26 in Figure 4. Here, four different processes are conducted in parallel:

150 the fundamental frequency (pitch change) is measured along each phoneme/word as a function of time;
151 audio amplitude changes detection (loudness change) is measured along each phoneme/word as a function of time;

152 rhythm detection (auto-correlation of the audio power) is measured along each phoneme/word as a function of time; and

153 spectral slope along each phoneme/word is measured along each phoneme/word as a function of time.

The outputs of all four are then combined at 154 to create prosodic descriptors.

The final enhancement process 28 of Figure 4 is illustrated in Figure 7 – the phoneme/word analysis. This may be performed using HLDs. The phoneme segmentation of the audio track 155 and the diphone segmentation of the audio track 156 are combined and at 157 the phoneme and diphone data is extracted and subjected to post-processing. Post-processing may include one or more of normalisation of amplitude via peak or RMS methodologies, noise reduction, and frequency filtering (either adaptive or manually set). The phoneme/diphone data descriptors and then encoded and passed for storage.

In Figure 9, there is shown the viseme extraction process. Visemes are the visual representations of the speaker mouth positions corresponding to the phonemes being pronounced. The mouth positions include one or more of tongue, teeth and lips. In the animation industry number of the visemes used varies according to many factors including, but not limited to production cost, feature type, quality, and so forth. Quality cartoons usually have twelve distinct mouth positions (visemes). Others may use no more than five visemes. Given that there is a direct correspondence between a phoneme and a viseme in a particular viseme set, the process of phoneme to viseme conversion may be in a direct lookup table, where the table is selected for a particular viseme set. As such, a set of visemes is created at 179. All phonemes used are then tabulated and the viseme corresponding to each phoneme is also tabulated to create a phoneme-to-viseme lookup table 178.

Upon the phoneme-to-viseme lookup table 178 being created, at 177 the complete timed word/phoneme list is input to the lookup table 178 and the corresponding visemes extracted at 180 and added to the timed word/phoneme list. The complete timed word/phoneme/viseme list is then output at 181 and the process ends at 182.

The present invention also extends to a computer usable medium comprising a computer program code that is configured to cause one or more processors to execute one or more functions to perform the steps and functions described above.

Whilst there has been described in the foregoing description a preferred embodiment of the present invention, it will be understood by those skilled in the technology concerned that many variations or modifications in details of design, construction or operation may be made without departing from the present invention.

The present invention extends to all features disclosed both individually, and in all possible permutations and combinations.

The claims:

1. A system for generating a timed phoneme and viseme list comprising the steps:
   (a) creating a lookup table of phonemes used and their corresponding visemes;
   (b) receiving a timed phoneme list;
   (c) analysing the lookup table to extract the visemes corresponding to each phoneme in the timed phoneme list; and
   (d) adding the visemes to the timed phoneme list to give a timed phoneme and viseme list.

2. A system as claimed in claim 1, wherein the timed phoneme list includes words.

3. A system as claimed in claim1 or claim 2,wherein the timed phoneme list is produced by:
   (a) receiving an audio track input and a script input, the script corresponding to the audio track input;
   (b) analysing the audio track to produce a time location for words in the audio track;
   (c) conducting a track reading process using a text-to-speech process on the script to produce a description of words, phonemes and phoneme length timings; and
   (d) processing the description of words, phonemes and phoneme length timings and time location to produce a timed word/phoneme list.

4. A system as claimed in claim 3, wherein the audio track is part of an audio/video input, subsequent to step (a) and before step (b) there is performed the step of converting the audio/video input to be MPEG1/MPEG2 compliant, and separating the audio track from the audio/video input.

5. A system as claimed in claim 3 or claim 4, wherein process steps (b) is conducted in parallel with process step (c).

6. A system as claimed in any one of claims 3 to 5, wherein before process step (d) is performed, an output from process step (c) is further processed to be MPEG 7 complaint.

7. A system as claimed in claim 6, wherein the output from process step (c) includes marked-up phonemes.

8. A system as claimed in any one of claims 3 to 7, wherein the timed phonemes are subjected to further processing to provide a speaker phoneme set prior to input into process step (d).

9. A system as claimed in any one of claims 3 to 8, wherein in process step (d) the analysis is a phoneme analysis.

10. A system as claimed is claim 9, wherein the phoneme analysis is an MPEG 7 word/phoneme analysis using sound spectrum and sound waveforms.

11. A system as claimed in any one of claims 3 to 10, wherein the audio track is also input into an MPEG 7 audio low level descriptors extraction process.

12. A system as claimed in any one of claims 3 to 11, wherein the audio track is also input into an MPEG 7 emotive analysis for extraction of emotion-related descriptors.

13. A system as claimed in claim 11, wherein the low level descriptors are one or more selected from the group consisting of: audio spectrum envelope descriptor, audio spectrum centroid descriptor, audio spectrum spread descriptor, audio spectrum flatness descriptor, audio spectrum basis descriptor, audio waveform descriptor, audio power descriptor, silence descriptor, audio spectrum projection descriptor, audio fundamental frequency descriptor, audio harmonicity descriptor, audio spectrum flatness type descriptor, log attack time descriptor, harmonic spectral centroid descriptor, harmonic spectral deviation descriptor, harmonic spectral spread descriptor, harmonic spectral variation descriptor, spectral centroid descriptor, and temporal centroid descriptor.

14. A system as claimed in claim 12, wherein the emotive analysis is conducted by measuring:
    (a) fundamental frequency along each phoneme/word as a function of time;
    (b) audio amplitude changes along each phoneme/word as a function of time;
    (c) rhythm detection along each phoneme/word as a function of time; and
    (d) spectral slope along each phoneme/word as a function of time.

15. A system as claimed in claim 10, wherein the word/phoneme analysis is conducted by using high level descriptors to analyse a phoneme segmentation of the audio track and a diphone segmentation of the audio track, combining the results, and extracting the phoneme and diphone data and subjecting the extracted data to post-processing.

16. A system as claimed in any one of claims 3 to 15, wherein a phrase of the script is generated using the phonemes, and a recognition process conducted to recognise the phrase in the audio track after processing of the audio track in a speech recognition process.

17. A system as claimed in claim 16, wherein the track reading process first breaks the script into words.

18. A system as claimed in claim 16 or claim 17, wherein any unrecognised phrases in the audio track after the recognition process are subsequently processed in the same manner with reset threshold and accuracy of the speech process.

19. A system as claimed in any one of claims 16 to 18, wherein prior to the speech recognition process parameters of the speech recognition process are set.

20. A system as claimed in any one of claims 16 to 19, wherein prior to the speech recognition process a next phrase logic of the speech recognition process is set.

21. A computer usable medium comprising a computer program code that is configured to cause one or more processors to execute one or more functions to perform the steps and functions as claimed in any one of claims 1 to 20.

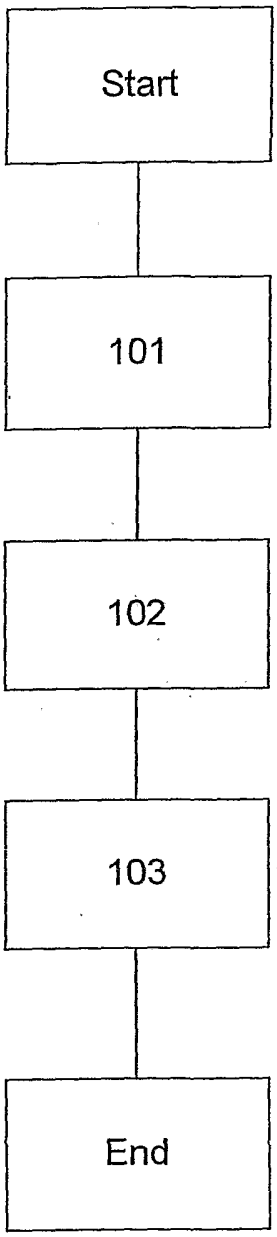


Figure 1

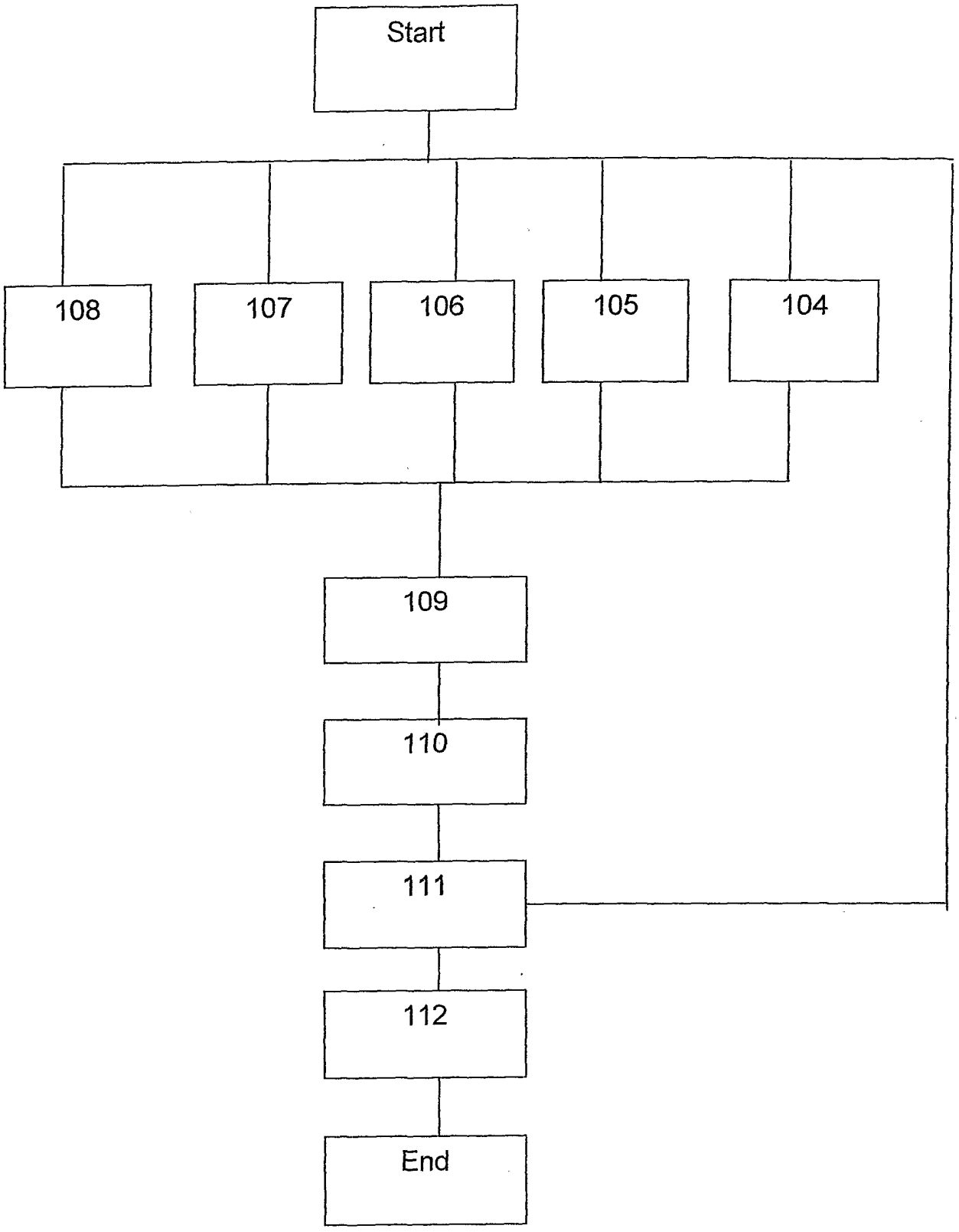


Figure 2

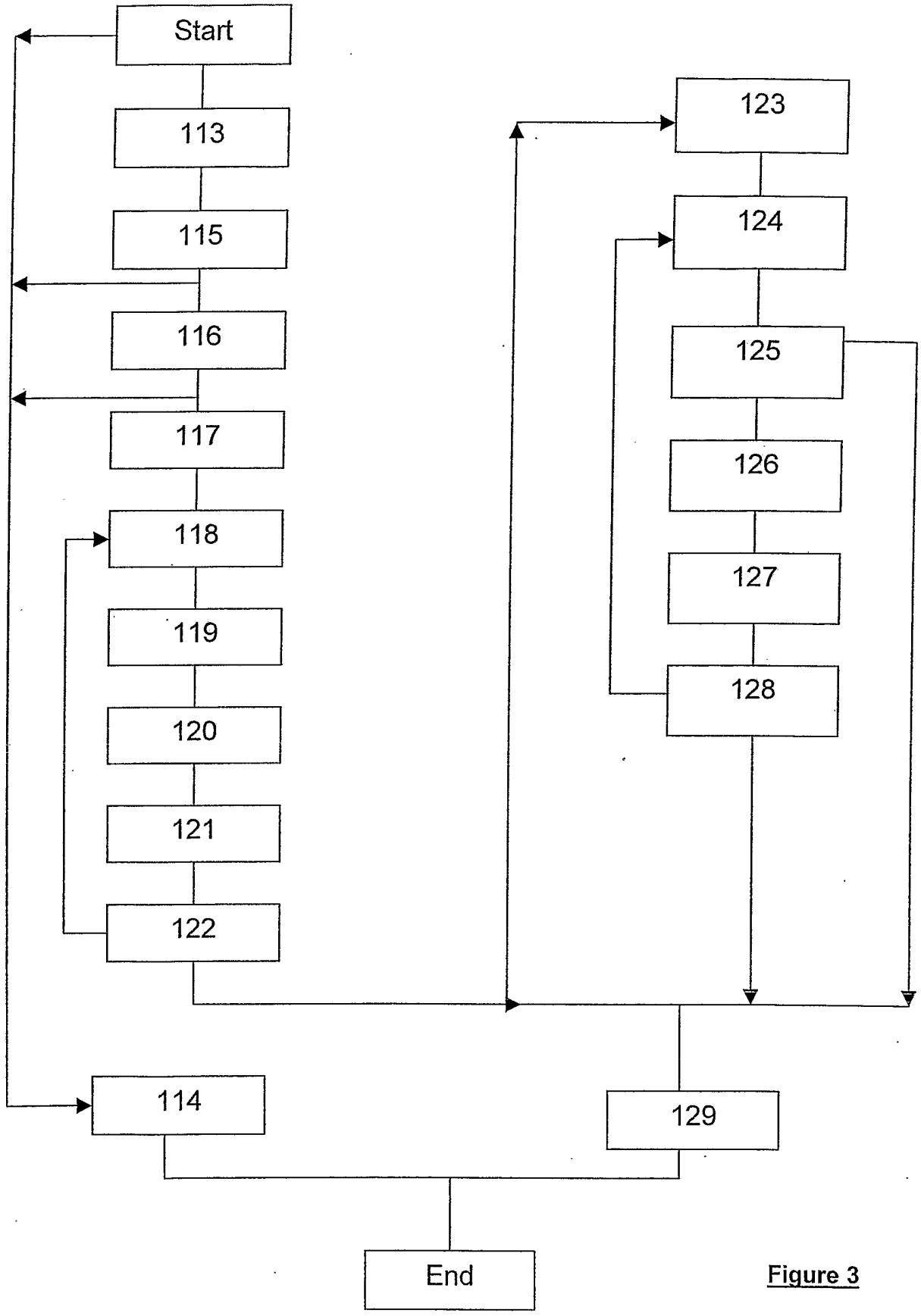

Start
113
115
116
117
118
119
120
121
122
123
124
125
126
127
128
114
129
End

Figure 3

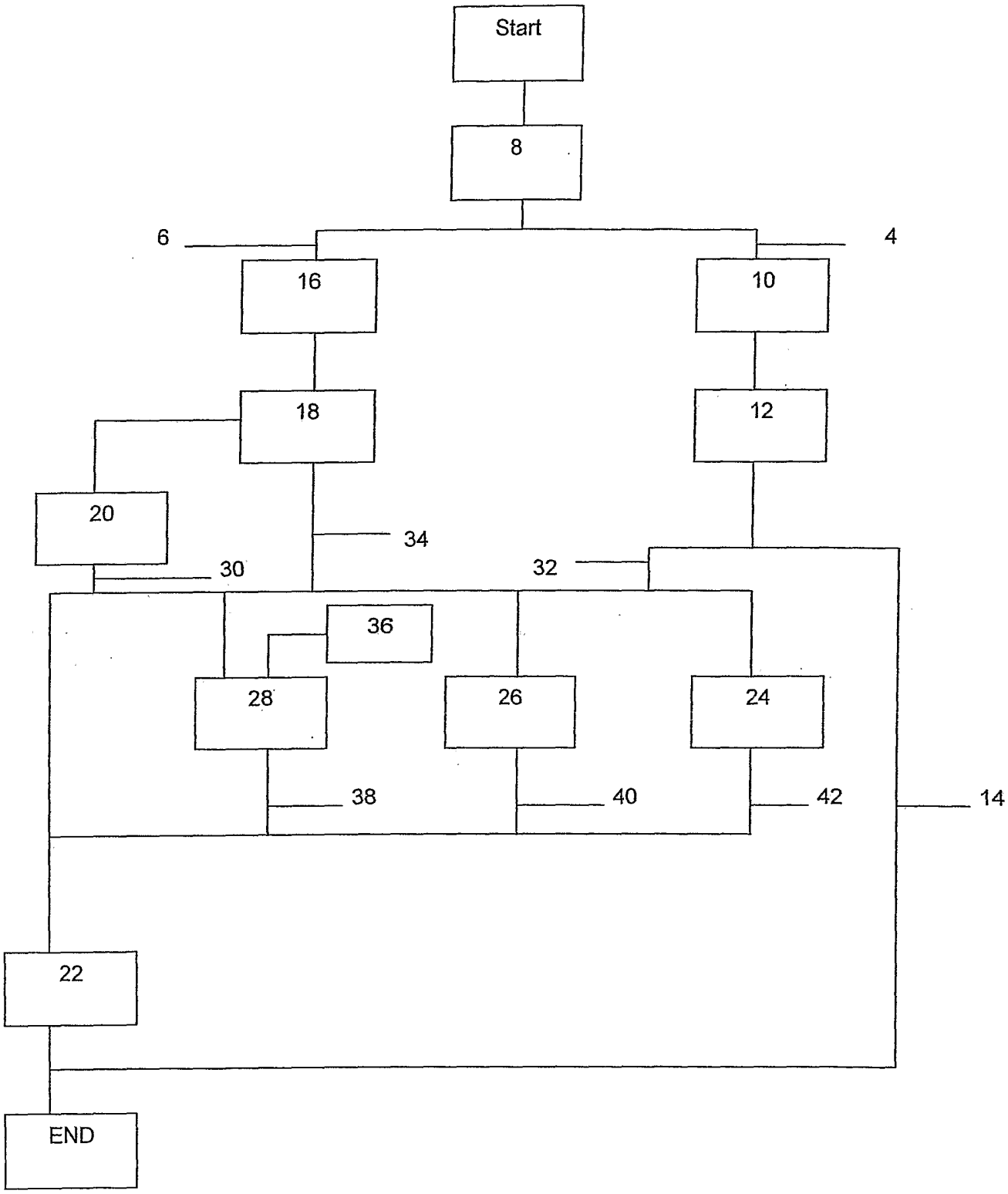


Figure 4

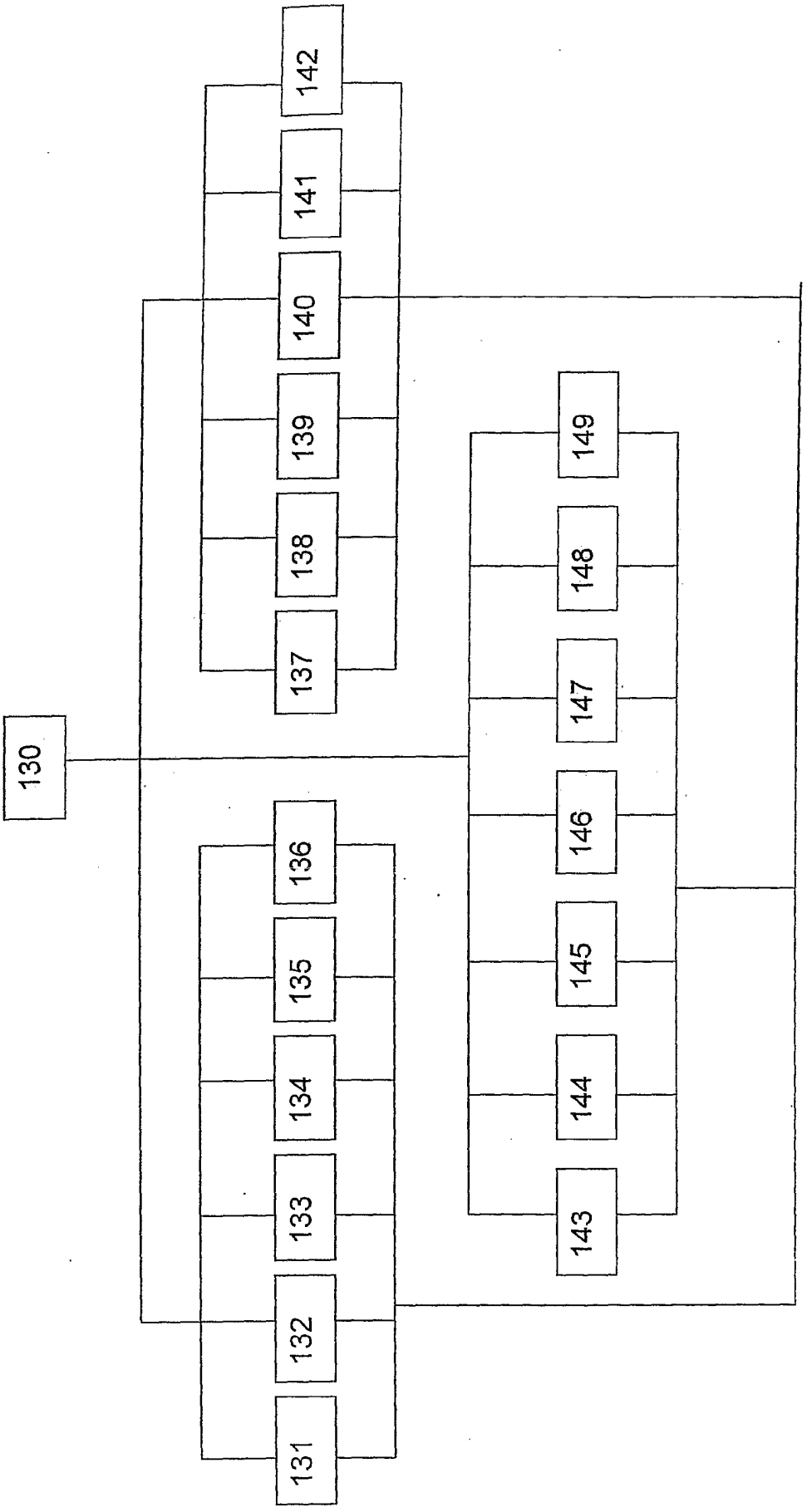

130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149

Figure 5

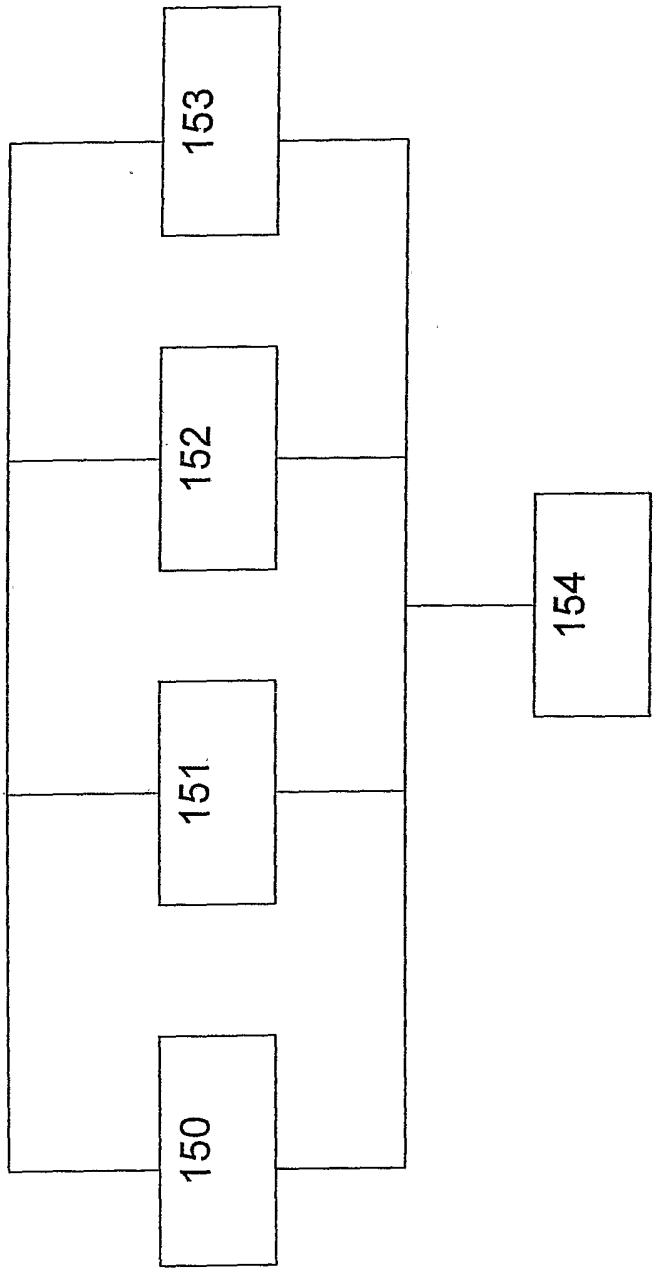


Figure 6

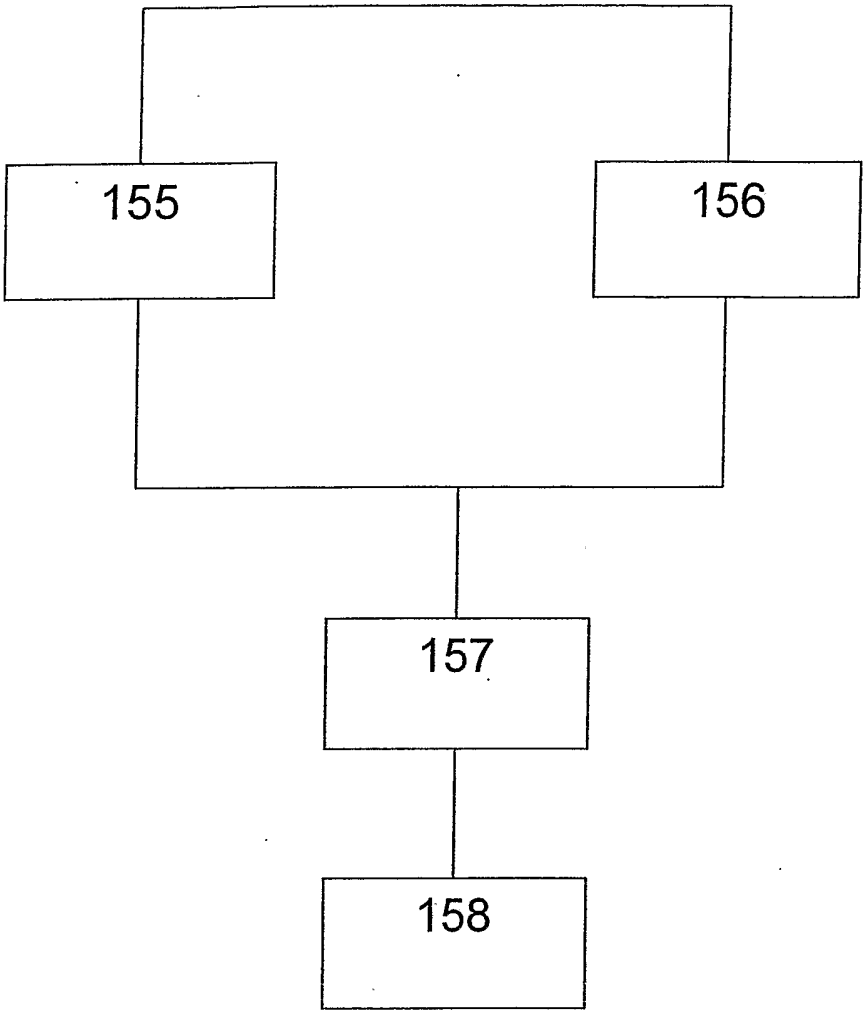


Figure 7

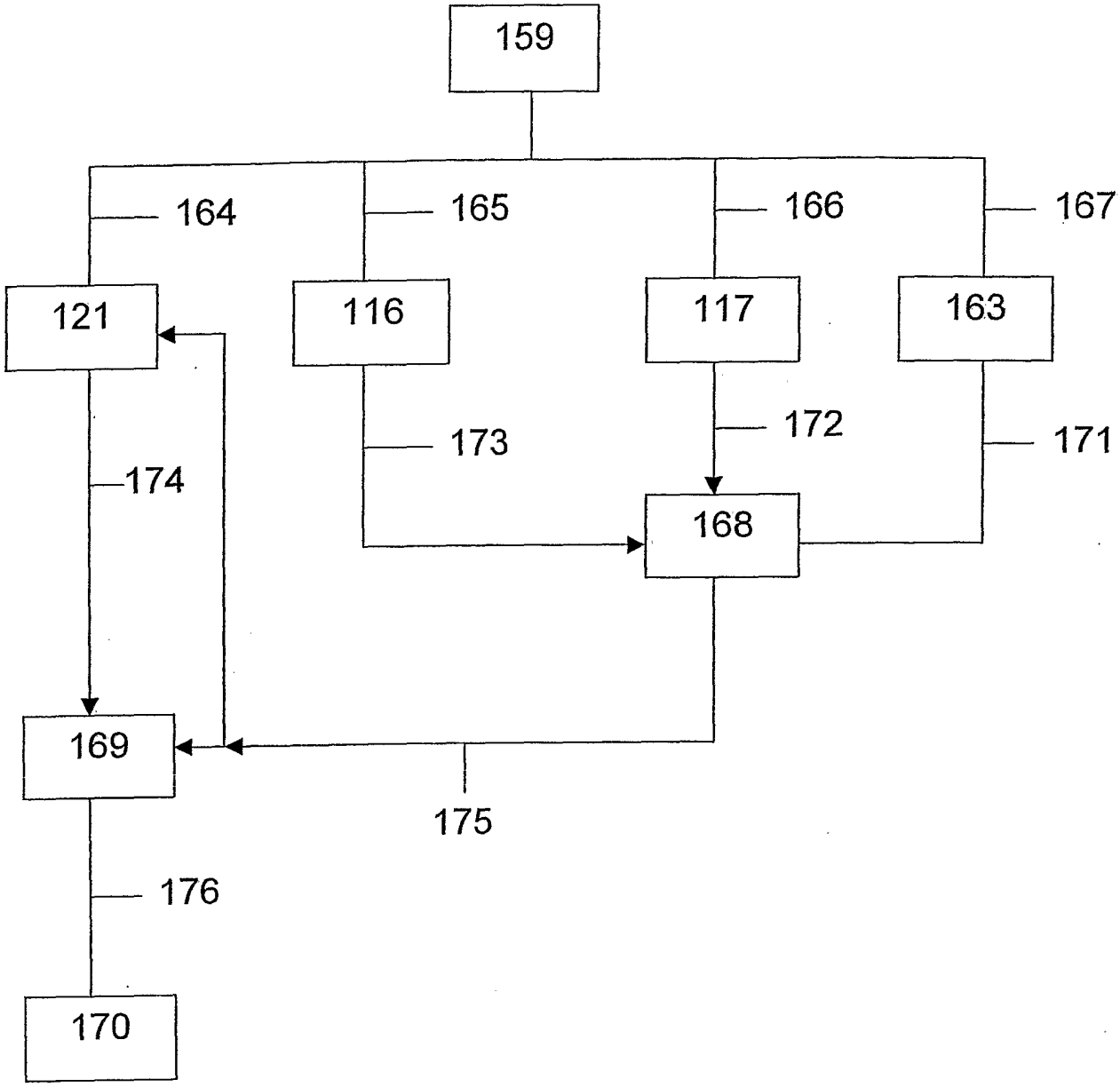

159
164
165
166
167
121
116
117
163
173
172
171
174
168
169
175
176
170

Figure 8

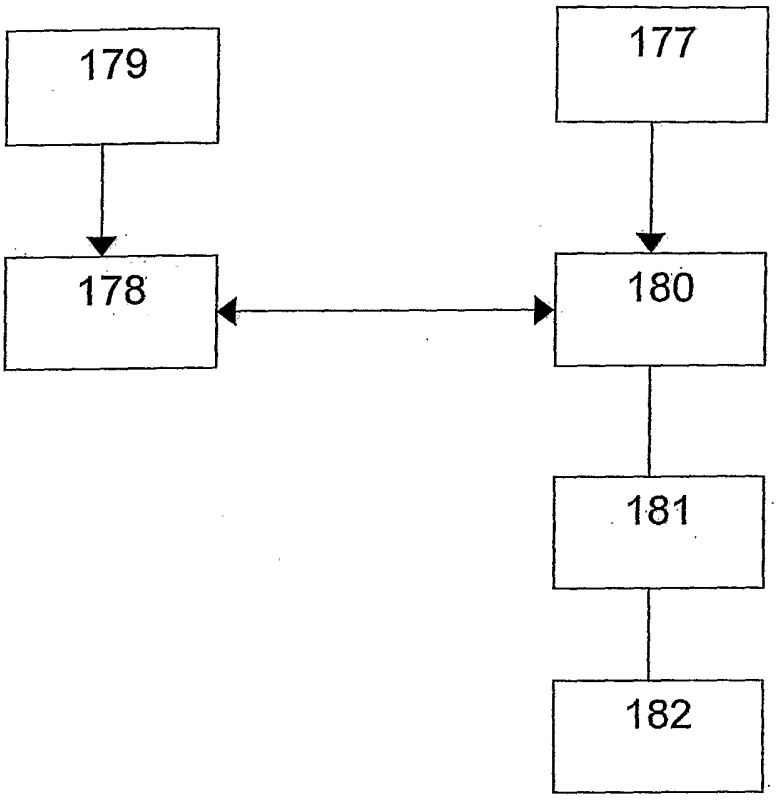


Figure 9

INTERNATIONAL SEARCH REPORT

International Application No
PCT/IB2004/001349

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/06 G10L15/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, IBM-TDB, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	EP 0 689 362 A (AT & T CORP) 27 December 1995 (1995-12-27) column 1, line 14 - line 17 column 2, line 26 - line 37 column 3, line 20 - line 28 column 4, line 34 - line 51 column 5, line 19 - line 24	1,4,8,21 3,9-20
X	WO 01/78067 A (ANANOVA LTD (GB)) 18 October 2001 (2001-10-18) abstract page 4, line 14 - line 24 page 5, line 24 - page 6, line 2 page 10, line 1 - line 2 page 11, line 27 - line 30 page 13, line 16 - line 22	1,2,7,21

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 July 2004	20/08/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Santos Luque, R

INTERNATIONAL SEARCH REPORT

International Application No
PCT/IB2004/001349

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2003/049015 A1 (COTE ET AL) 13 March 2003 (2003-03-13) abstract page 4, paragraph 65 page 3, paragraphs 32,45,49 figure 1	3,9-20
Y	US 5 734 794 A (WHITE) 31 March 1998 (1998-03-31) abstract column 2, line 37 - line 48 column 4, line 46 - line 63	10-14
A	BAGLEY AND GRACER: "Method for Computer Animation of Lip Movements. March 1972." IBM TECHNICAL DISCLOSURE BULLETIN, vol. 14, no. 10, 1 March 1972 (1972-03-01), pages 3039-3040, XP002289401 New York, US the whole document	3,9
A	QUACKENBUSH S ET AL: "OVERVIEW OF MPEG-7 AUDIO" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE INC. NEW YORK, US, vol. 11, no. 6, June 2001 (2001-06), pages 725-729, XP001059867 ISSN: 1051-8215 the whole document	6,7, 10-15
A	COWIE R ET AL: "Emotion recognition in human-computer interaction" IEEE SIGNAL PROCESSING MAGAZINE, IEEE INC. NEW YORK, US, vol. 18, no. 1, January 2001 (2001-01), pages 32-80, XP002191161 ISSN: 1053-5888 tables 7,8	14

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/IB2004/001349

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 0689362	A	27-12-1995	US	5608839 A	04-03-1997
			CA	2149068 A1	22-12-1995
			CN	1132472 A ,B	02-10-1996
			EP	0689362 A2	27-12-1995
			JP	8023530 A	23-01-1996
WO 0178067	A	18-10-2001	AU	4669201 A	23-10-2001
			CN	1426577 T	25-06-2003
			EP	1269465 A1	02-01-2003
			WO	0178067 A1	18-10-2001
			JP	2003530654 T	14-10-2003
			US	2003149569 A1	07-08-2003
US 2003049015	A1	13-03-2003	WO	03023765 A1	20-03-2003
			EP	1425736 A1	09-06-2004
US 5734794	A	31-03-1998	NONE		