

(19) World Intellectual Property
Organization
International Bureau



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

PCT

(10) International Publication Number
WO 2004/100164 A1

(51) International Patent Classification⁷: **G11B 27/028**,
27/10, 27/28

(21) International Application Number:
PCT/IB2004/001336

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

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PI 20031469 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) Agent: **ALBAN TAY MAHTAN & DE SILVA**; 39 Robinson Road, #07-01 Robinson Point, Singapore 068911 (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).

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.

(54) Title: VOICE SCRIPT SYSTEM

(57) Abstract: A voice script system for providing a scrolling text on video content including the steps: using an audio track corresponding to the video content and a script corresponding to the audio track to produce time-stamped text data; aligning the time-stamped text data with the audio track and the video content; and translating the aligned, time-stamped text data to a bitmap graphic on a scrolling track of the video content. The time-stamped text data is time-stamped by phrase or word start. The text may be displayed on the scrolling track in a plurality of lines scrolling from the bottom, or on the scrolling track in a single line scrolling from left to right.



WO 2004/100164 A1

Voice Script System

Field of the Invention

5 The present invention relates to a voice script system and refers particularly, though not exclusively, to such a voice process system to assist with post-production voice processes of an audio/video presentation.

Reference to Related Applications

10

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

15

Background to the Invention

Historically much of the voice track of an audio/video presentation is re-recorded during postproduction. This is because the audio track recorded on location may include much
20 background noise that interferes with the sound quality, may not have been able to be recorded at the correct quality, and so forth. In the re-recording process, the director and actors are in a recording studio, and the video is played on a screen. They listen to the location audio track through headphones, and try to re-record exactly what was recorded on location in synch with the video, but with the correct enunciation and emotion. This may take several attempts for each
25 sentence, phrase, or other language structure.

This is very time consuming and may impact on how fast the audio/video presentation may be released, and the cost of that audio/video presentation. The timing may be important in current affairs broadcasts on television.

30

The normal post-production processing costs about one third of the total budget. Re-recording the audio track is about one third of that cost. To facilitate the post-production audio processing may therefore not only reduce costs, but may also reduce the time taken, and thus allow the audio/video presentation to be on the market earlier.

35

The audio/video presentation may be for a motion picture, a television broadcast, streaming video, semi-professional or home video particularly digital video, or portable media including DVD, VCD, and so forth.

5 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.

10

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

Summary of the Invention

15

According to one aspect of the invention there is provided a voice script system for providing a scrolling text on video content including the steps:

- (a) using an audio track corresponding to the video content and a script corresponding to the audio track to produce time-stamped text data;
- 20 (b) aligning the time-stamped text data with the audio track and the video content; and
- (c) translating the aligned, time-stamped text data to a bitmap graphic on a scrolling track of the video content.

The time-stamped text data may be time-stamped by phrase start. Between steps (a) and (b) there may be included a further process step of obtaining a machine translation of the phonemes of the script. After the machine translation is obtained, manual correction for slang and idiomatic language may be performed. The text may be displayed on the scrolling track in a plurality of lines scrolling from the bottom. The scrolling may be by phrase.

30 Alternatively, the time-stamped text may be time-stamped by word start. The text may be displayed on the scrolling track in a single line scrolling from left to right.

The text alignment may be performed by using a punctuation filter to locate the beginnings and ends of sentences.

35

The first process step of producing time-stamped text data may be conducted using a track reading process. The track reading process may include the steps:

- (a) breaking the script into phonemes;
- (b) using a speech recognition process on the audio track to produce a processed audio track;
- (c) generating a phrase of the script using the phonemes;
- 5 (d) conducting a recognition process to recognise the phrase in the processed audio track; and
- (e) extracting the phonemes from an output of the recognition process.

10 The phonemes may be obtained using a text-to-speech process, and the script may be broken into words before the phonemes are obtained. Any unrecognised phrases in the audio track after the recognition process may be subsequently processed in the same manner with reset threshold and accuracy of the speech recognition process. Prior to the speech recognition process, parameters and next phrase logic of the speech recognition process may be set.

15 The audio track may be analysed to produce a time location for words in the audio track. The speech recognition process on the script may also produce a description of words and phoneme length timings. The description of words, phonemes and phoneme length timings may be further processed with a timer input to produce a timed word/phoneme list. The timer input may be obtained from the audio track.

20 The audio track may be part of an audio/video input and the audio/video input may be first converted to be MPEG1/MPEG2 compliant before separating the audio track from the audio/video input.

25 The timed phonemes may be subjected to further processing to provide a speaker phoneme set. The speech recognition process may also produce marked-up phonemes. The further processing may be a phoneme analysis. The phoneme analysis may be an MPEG 7 word/phoneme analysis using sound spectrum and sound waveforms. The audio track may also be input into an MPEG 7 audio low level descriptors extraction process and an MPEG 7 emotive analysis for extraction of
30 emotion-related descriptors.

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; as well as
35 to a video having a scrolling text prepared by the above process.

Description of the Drawings

In order that the 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 the track reading process of Figure 1;

Figure 3 is a flow chart for a preferred process for phoneme extraction;

10 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; and

Figure 7 is a flow chart of word/phoneme analysis.

15 **Description of the Preferred Embodiment**

To refer to Figure 1, there is illustrated the overall process, and the initial process steps. At 100 a machine of a user logs on to the relevant web page to initiate the process. However, the process may be performed on a stand-alone machine.

20

The feature selection is made at 98. It is assumed for the following description that the process is for a new feature and thus the entire process is required.

25

The audio, video and text files corresponding to the audio files, requiring post-processing are selected and input to the server. If a multi-track audio dubbing is to be performed, the input audio track is supplemented by a multi-track voice input 96. In that case, auxiliary text files 94 for the multi-track audio may be required.

30

At 98 the server operating the web page performs a track reading process. This is more completely described in relation to Figures 2, 3 and 4 below. This produces time-stamped text data. At 90, the output of time-stamped text data 92 may be sent to one or both of two streams 88 and 86. A first stream 88 is used when the server has been advised by the user's machine that the text is to be in a different, nominated language. The time-stamped text data is converted to XML at 84 and time-stamped at the start of each phrase. Any translation required is then performed at 82 using a machine translation. The translated text from 82 may be checked for idiomatic language, slang expressions, and corrections at 80.

35

Scrolling from the bottom to the top, each phrase of the translated text is translated to bitmap graphics on two lines at the bottom of each frame that constitute the subtitle track. As the bitmap graphics of the text scroll from the bottom to the top they can easily be read by the actors. As complete phrases are shown, the actors can read the complete phrase at the time the phrase commences and can read the phrase knowing the video context so they can include all necessary enunciation, emotions and emphasis required to match the video and thus have correct audio/lip sync. As there may be more than one actor present at the one time, multiple tracks may be separately displayed. This may be in a side-by-side relationship or in consecutive lines on the scrolling text in the same manner as they would appear on a printed script. As the actors would be reading the top line of the displayed script, if another actor has the following line, their cue time is as long as it takes the first actor to read their line.

If desired, a background colour and the font, font size and font colour, can be selected at 76. As the text as displayed is time-stamped, it will be in sync with the video and will scroll at an appropriate speed to keep pace with the audio and video. This has the script being displayed in what may be termed teleprompter mode (74).

In this way the actor's position is uniform as they are looking at the screen for both the visual cues, and for the text. This means they are a uniform distance from the microphone, thus aiding high quality recording. Also, because they are reading the text from the screen, they do not need to follow a printed text. This avoids the inherent noise of page turns.

The output may be a standard video file that is able to be imported in to digital audio workstations such as "Pro Tools". It also enables the output to be used as a synchronised voice prompter.

The stream 86 is generally the same except the time stamp is by word rather than phrase (72) – all other aspects are the same as 84). In addition, there is no translation, and the word text scrolls left-to-right (70). As the script text displayed is scrolling, the opacity of the display may also need to be selected in 68 – all other selections are the same as 76. As the video has the text corresponding to the audio (and matching the video) scrolling left-to-right it will be in sync when it crosses an imaginary line in the centre of the display at 66.

Again, the actor's position is uniform as they are looking at the screen for both the visual cues, and for the text. This means they are a uniform distance from the microphone, thus aiding high quality recording. Also, because they are reading the text from the screen, they do not need to follow a printed text. This avoids the inherent noise of page turns.

The first stage of the track reader 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;
- 5 (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.

10 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 of track reading ends.

15

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.

20 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.

25 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, "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
30 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
35 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.

5 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.

10

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.

15

20 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.

25

If there are unrecognised phrases in the script (normally due to the speech recognition engine producing an 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.

30

35 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.

5

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
10 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
15 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.

20

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.

25 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
30 recognition process ends at 170.

To now refer to Figure 4, there is shown the overall track reader 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
35 text input in digital form. The audio/video input is preferably in source language. This is process 96 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.

5

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.

- 10 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
- 15 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.

20

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 [any others?]. The LLD analysis may proceed in parallel to the track reading process 18, or may be subsequent to that process. If subsequent,

25 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
- 30 LLDs 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
- 35 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 spectrums, 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; log attack time descriptor;
- 143 harmonic spectral centroid descriptor
- 144 harmonic spectral deviation descriptor;
- 145 harmonic spectral spread descriptor;
- 146 harmonic spectral variation descriptor;
- 147 spectral centroid descriptor; and
- 148 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:

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

- 151 rhythm detection (auto-correlation of the audio power) is measured along each phoneme/word as a function of time; and
- 152 spectral slope along each phoneme/word is measured along each phoneme/word as a function of time.
- 5 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 normalization of amplitude via peak or RMS methodologies, noise reduction, and frequency filtering (either adaptive or manually set). The phoneme/diphone data descriptors are then encoded and passed for storage as well as for input into the translation process 90.

- 15 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; as well as to a video having subtitles prepared by the above system.
- 20 The translation is preferably conducted using a machine translation. Suitable machine translation engines include Systrans 4.0. The translation may be into any suitable language in any format – using the Roman alphabet, Roman numerals, Arabic numerals, and any regional or national alphabets, scripts and symbols.
- 25 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.
- 30 The present invention extends to all features disclosed both individually, and in all possible permutations and combinations.

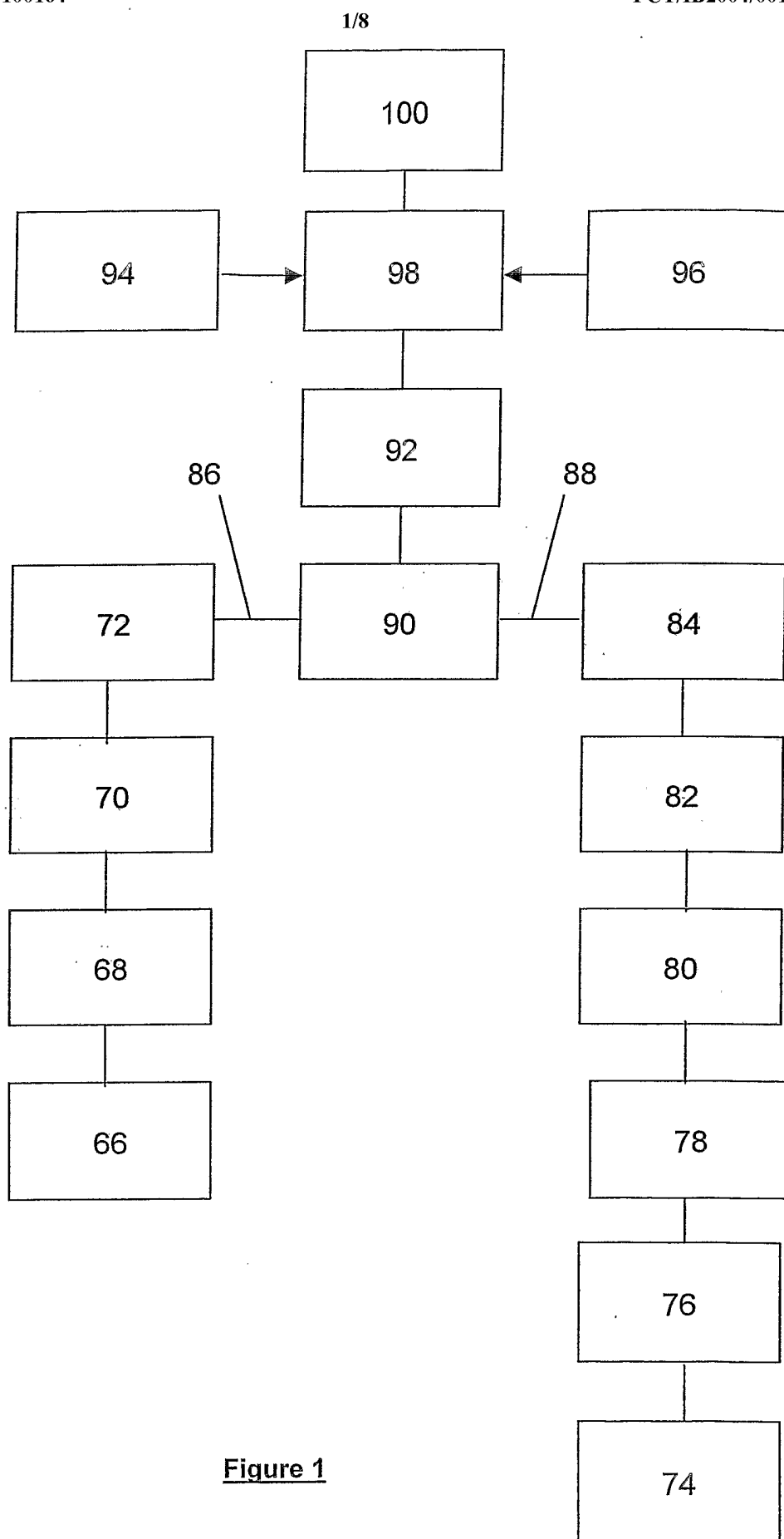
The claims:

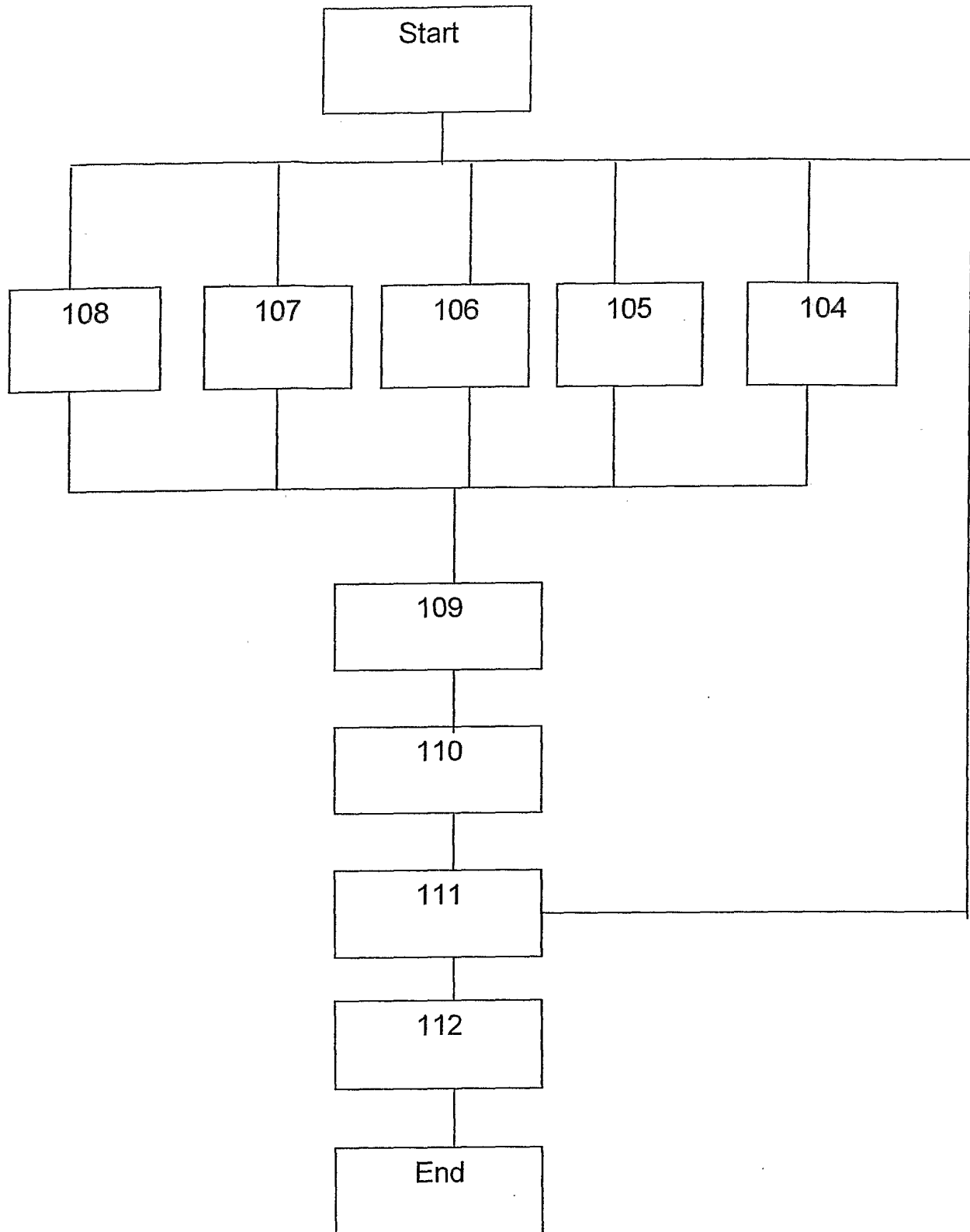
1. A voice script system for providing a scrolling text on video content including the steps:
(d) using an audio track corresponding to the video content and a script corresponding to the audio track to produce time-stamped text data;
5 (e) aligning the time-stamped text data with the audio track and the video content; and
(f) translating the aligned, time-stamped text data to a bitmap graphic on a scrolling track of the video content.
2. A process as claimed in claim 1, wherein the time-stamped text data is time-stamped by
10 phrase start.
3. A process as claimed in claim 2, wherein between steps (a) and (b) there is included a further process step of obtaining a machine translation of the phonemes of the script.
- 15 4. A process as claimed in claim 3, wherein after the machine translation is obtained, manual correction for slang and idiomatic language is performed.
5. A process as claimed in any one of claims 1 to 4, wherein the text is displayed on the scrolling track in a plurality of lines scrolling from the bottom.
20
6. A process as claimed in claim 5, wherein the scrolling is by phrase.
7. A process as claimed in claim 1, wherein the time-stamped text is time-stamped by word start.
25
8. A process as claimed in claim 7, wherein the text is displayed on the scrolling track in a single line scrolling from left to right.
9. A process as claimed in any one of claims 1 to 8, wherein the text alignment is performed
30 by using a punctuation filter to locate the beginnings and ends of sentences.
10. A process as claimed in any one of claims 3 to 6, wherein the translation is into a language as input from a user's machine.
- 35 11. A process as claimed in any one of claims 1 to 10, wherein process step (a) is conducted using a track reading process.

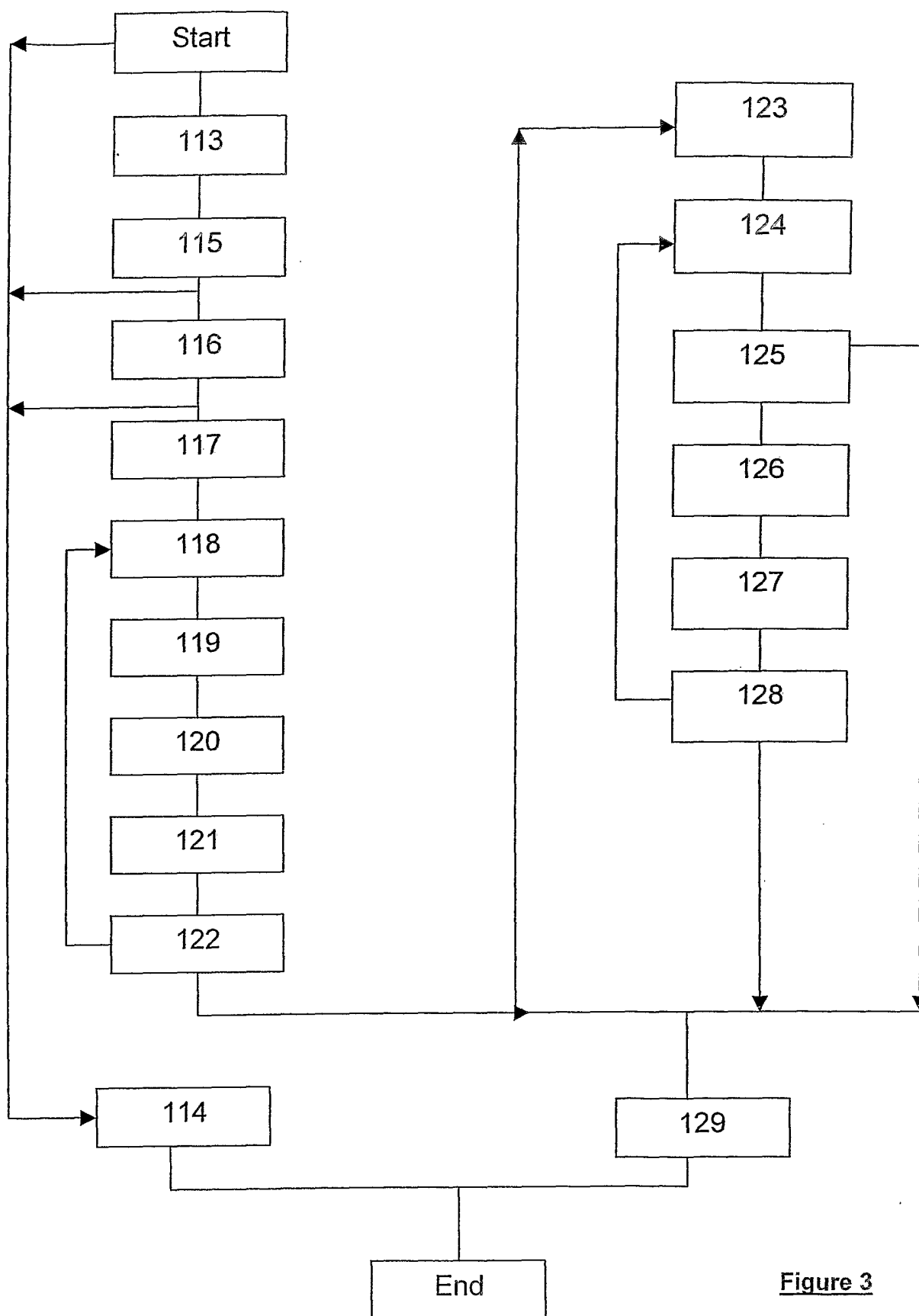
12. A process as claimed in claim 11, wherein the track reading process includes the steps:
- (a) breaking the script into phonemes; using a speech recognition process on the audio track to produce a processed audio track;
 - (b) analysing the audio track to produce a time location for words in the audio track;
 - (c) generating a phrase of the script using the phonemes;
 - (d) conducting a recognition process to recognise the phrase in the processed audio track; and
 - (e) extracting the phonemes from an output of the recognition process.
13. A process as claimed in claim 12, wherein the phonemes are obtained using a text-to-speech process.
14. A process as claimed in claim 12 or claim 13, wherein the script is broken into words before the phonemes are obtained.
15. A process as claimed in any one of claims 12 to 14, 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 recognition process.
16. A process as claimed in any one of claims 12 to 15, wherein prior to the recognition process, parameters of the speech recognition process and a next phrase logic of the speech recognition process, are set.
17. A process as claimed in any one of claims 12 to 16, wherein the audio track is analysed to produce a time location for words in the audio track;
18. A process as claimed in claim 17, wherein the speech recognition process on the script also produces a description of words and phoneme length timings.
19. A process as claimed in claim 18, wherein the description of words, phonemes and phoneme length timings are further processed with a timer input to produce a timed word/phoneme list.
20. A process as claimed in claim 19, wherein the timer input is obtained from the audio track.

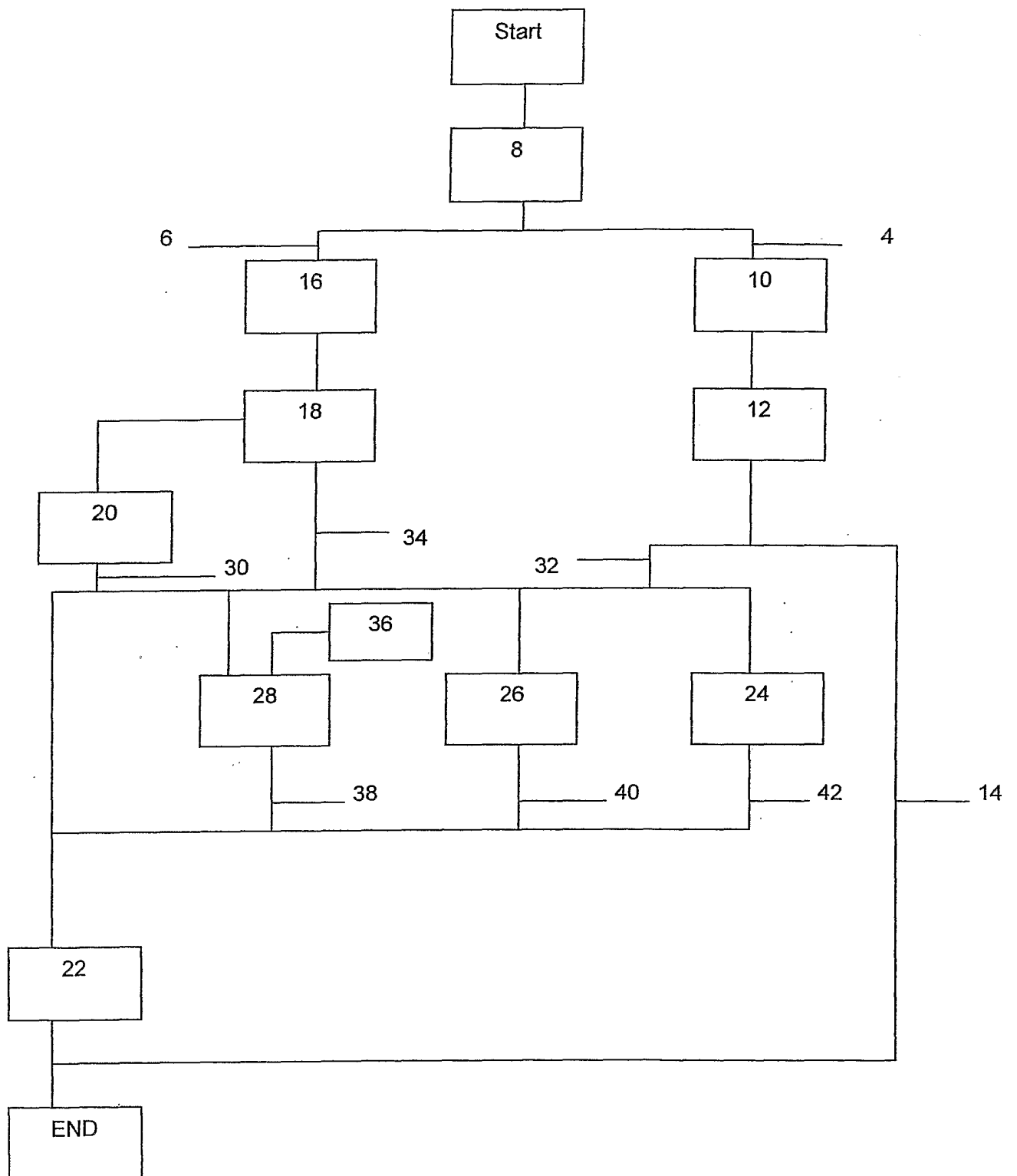
21. A process as claimed in claim 19 or claim 20, wherein the audio track is part of an audio/video input and the audio/video input is first converted to be MPEG1/MPEG2 compliant before separating the audio track from the audio/video input.
- 5 22. A process as claimed in any one of claims 18 to 21, wherein the speech recognition process also produces marked-up phonemes.
23. A process as claimed in any one of claims 19 to 22, wherein the further processing is a phoneme analysis.
- 10 24. A process as claimed in claim 23, wherein the phoneme analysis is an MPEG 7 word/phoneme analysis using sound spectrum and sound waveforms.
25. A process as claimed in any one of claims 12 to 24, wherein the audio track is also input into an MPEG 7 audio low level descriptors extraction process.
- 15 26. A process as claimed in any one of claims 12 to 25, wherein the audio track is also input into an MPEG 7 emotive analysis for extraction of emotion-related descriptors.
- 20 27. A process as claimed in claim 25, 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.
- 25 28. A process as claimed in claim 26, wherein the emotive analysis is conducted by measuring;
- 30 (a) 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
- 35 (d) spectral slope along each phoneme/word as a function of time.

29. A process as claimed in claim 24, 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.
- 5
30. An audio/video presentation having an audio track prepared by the process of any one of claims 1 to 29.
31. A video presentation having a scrolling text display prepared by the process of any one of
- 10
- claims 1 to 29.
32. 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 29.
- 15

**Figure 1**

**Figure 2**

**Figure 3**

**Figure 4**

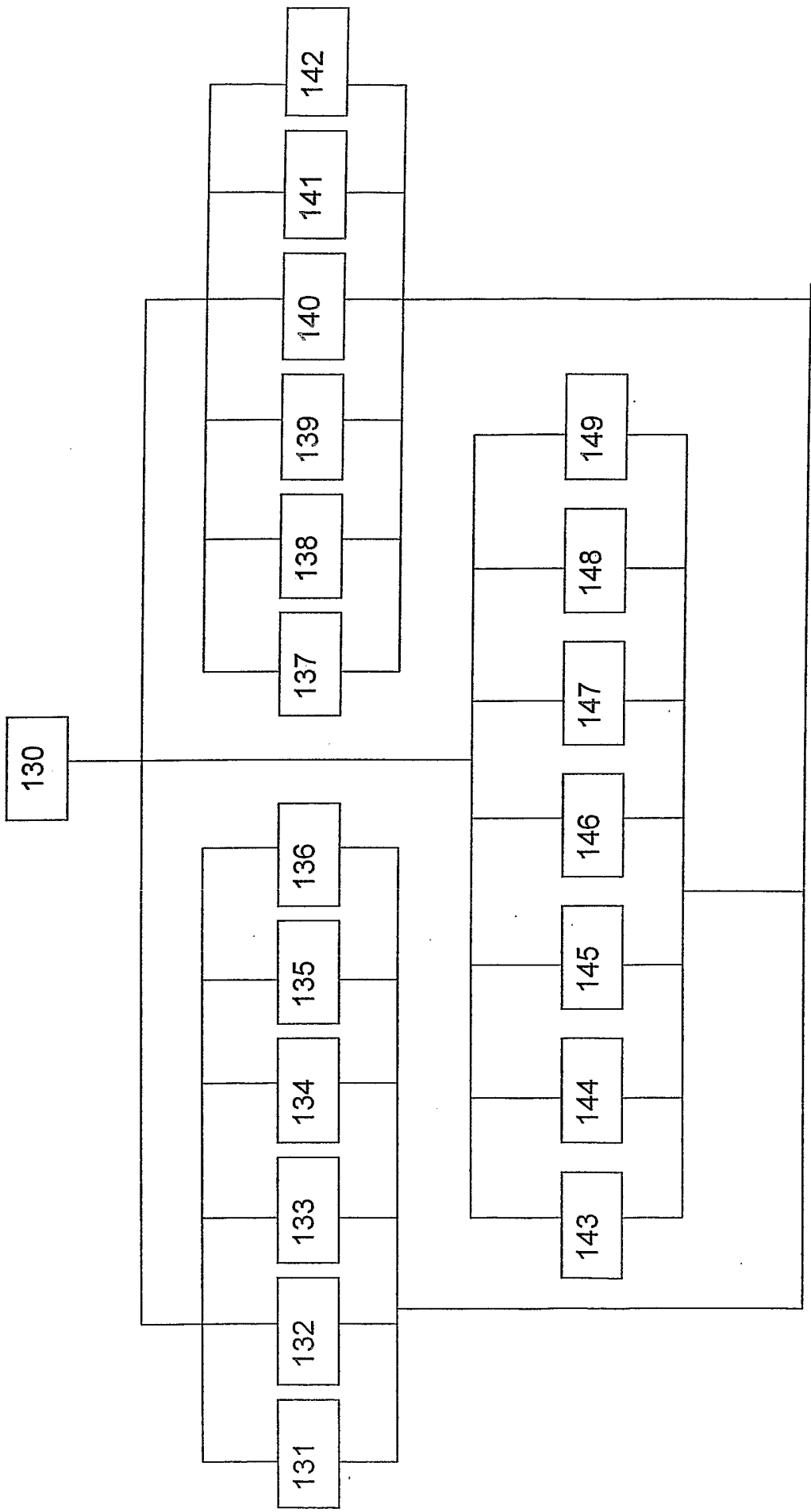


Figure 5

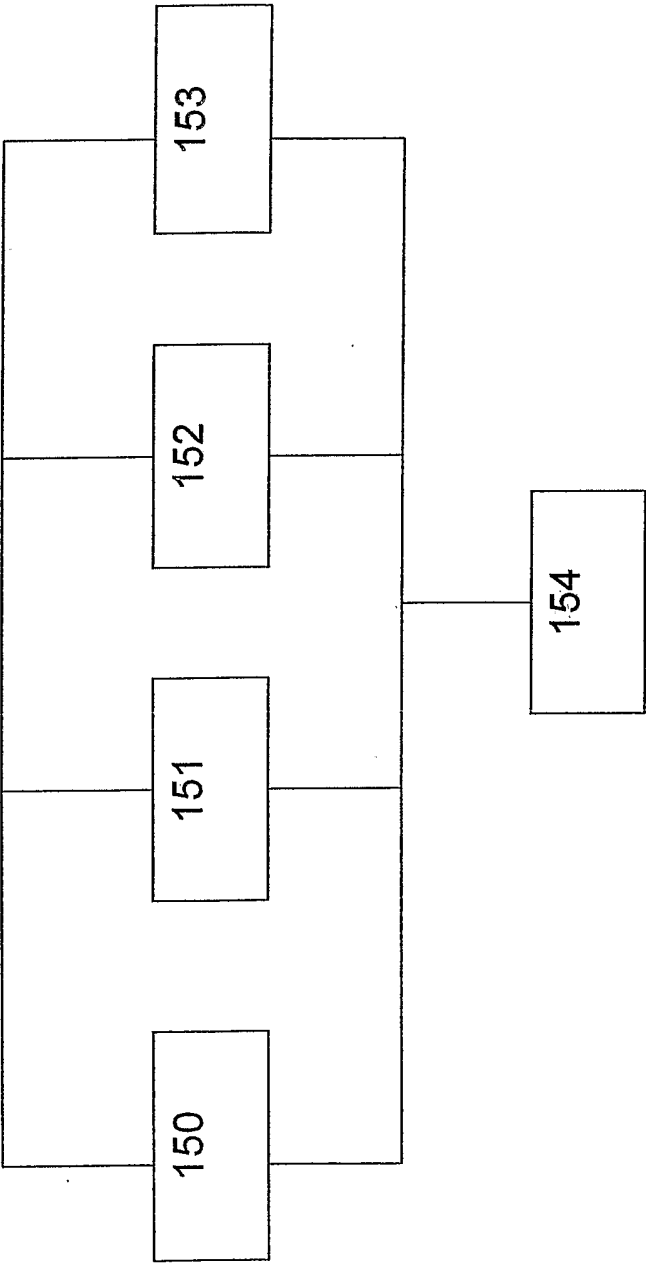


Figure 6

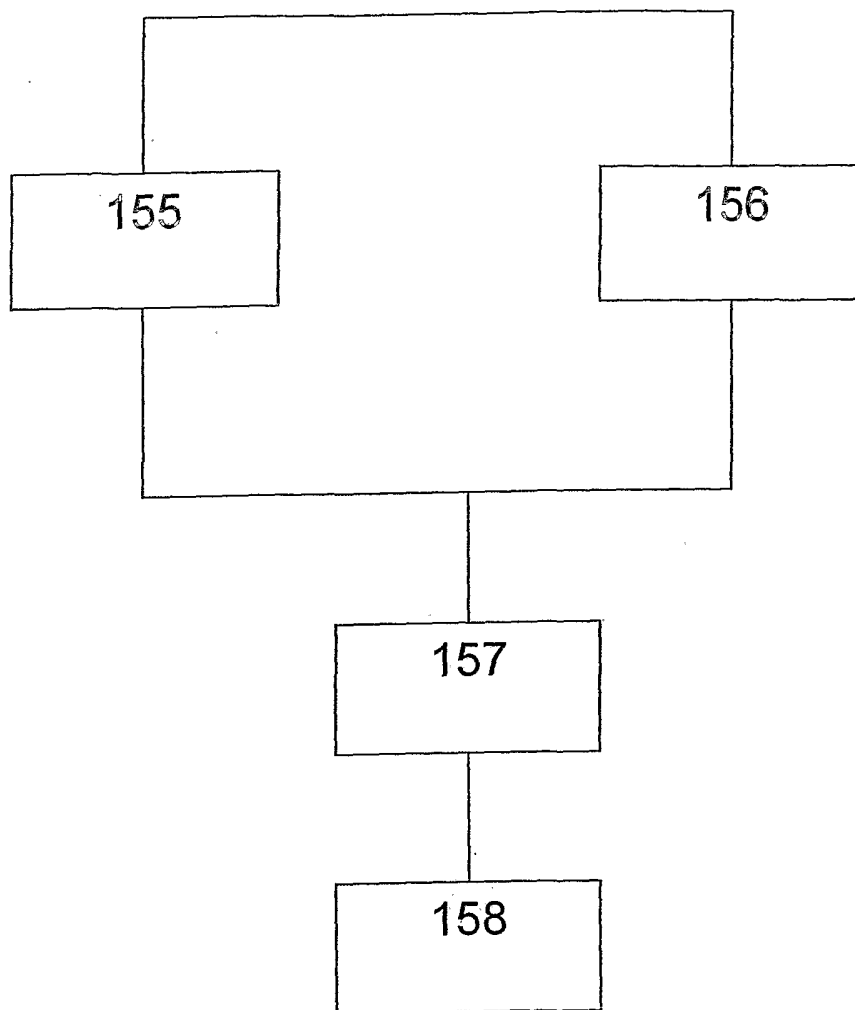
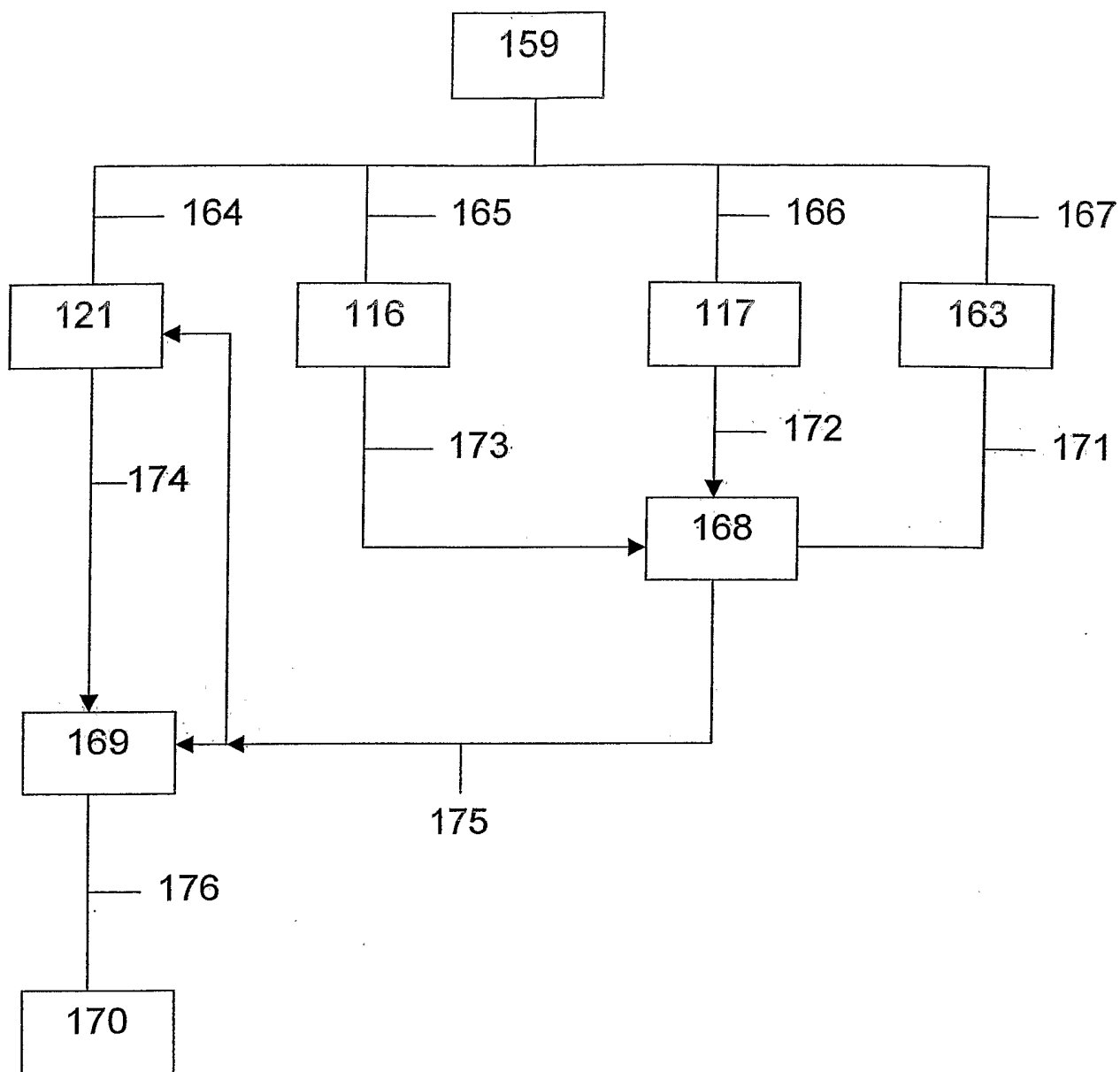


Figure 7

**Figure 8**

INTERNATIONAL SEARCH REPORT

International Application No

PCT/IB2004/001336

A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 G11B27/028 G11B27/10 G11B27/28

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 G11B

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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 649 060 A (ZADROZNY WLODEK WLODZIMIERZ ET AL) 15 July 1997 (1997-07-15) column 1, line 43 - column 3, line 8 column 3, line 47 - column 10, line 44; figures	1-32
X	WO 01/95631 A (BRITISH BROADCASTING CORP ;WIEWIORKA ADAM (GB); KIRBY DAVID GRAHAM) 13 December 2001 (2001-12-13) abstract page 5, line 13 - page 6, line 5 page 7, line 10 - page 8, line 7 page 9, line 23 - page 10, line 6 page 14, line 18 - page 15, line 5 page 17, line 28 - page 20, line 23 page 38, line 5 - page 42, line 21 ----- -/--	1-32



Further documents are listed in the continuation of box C.



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

21 July 2004

Date of mailing of the international search report

27/07/2004

Name and mailing address of the ISA

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

Authorized officer

D'Attilia, M

INTERNATIONAL SEARCH REPORT

International Application No

PCT/IB2004/001336

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WATANABE K ET AL: "Automatic caption generation for video data. Time alignment between caption and acoustic signal" 13 September 1999 (1999-09-13), MULTIMEDIA SIGNAL PROCESSING, 1999 IEEE 3RD WORKSHOP ON COPENHAGEN, DENMARK 13-15 SEPT. 1999, PISCATAWAY, NJ, USA, IEEE, US, PAGE(S) 65-70 , XP010351716 ISBN: 0-7803-5610-1	1,11-14, 17-20, 22,23, 30-32
A	the whole document	2-10,15, 16,21, 25-29
A	EP 0 896 467 A (BRITISH BROADCASTING CORP) 10 February 1999 (1999-02-10) paragraph '0007! - paragraph '0008! paragraph '0014! - paragraph '0017!; figures 1,3	1-32
A	US 6 490 550 B1 (HIRI FARZAD) 3 December 2002 (2002-12-03) column 4, line 27 - column 5, line 26; figure 2	1-32

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/IB2004/001336

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 5649060	A	15-07-1997	DE 69422466 D1	10-02-2000
			DE 69422466 T2	06-07-2000
			EP 0649144 A1	19-04-1995
			JP 2986345 B2	06-12-1999
			JP 7199379 A	04-08-1995
WO 0195631	A	13-12-2001	CA 2411038 A1	13-12-2001
			EP 1295482 A2	26-03-2003
			WO 0195631 A2	13-12-2001
			GB 2379312 A	05-03-2003
			US 2004093220 A1	13-05-2004
EP 0896467	A	10-02-1999	EP 0896467 A1	10-02-1999
			GB 2328069 A ,B	10-02-1999
			US 6226615 B1	01-05-2001
US 6490550	B1	03-12-2002	AU 1747200 A	19-06-2000
			EP 1135921 A1	26-09-2001
			WO 0033552 A1	08-06-2000