US 20040234250A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2004/0234250 A1**

**Cote et al.** (43) **Pub. Date: Nov. 25, 2004**

(54) **METHOD AND APPARATUS FOR PERFORMING AN AUDIOVISUAL WORK USING SYNCHRONIZED SPEECH RECOGNITION DATA**

(76) Inventors: **Jocelyne Cote**, Longueuil (CA); **Howard Ryshpan**, Longueuil (CA)

Correspondence Address:
OGILVY RENAULT
1981 MCGILL COLLEGE AVENUE
SUITE 1600
MONTREAL, QC H3A2Y3 (CA)

(21) Appl. No.: **10/797,576**

(22) Filed: **Mar. 11, 2004**

**Related U.S. Application Data**

(63) Continuation-in-part of application No. 10/067,131, filed on Sep. 12, 2001. Continuation of application No. PCT/CA02/01386, filed on Sep. 12, 2002.

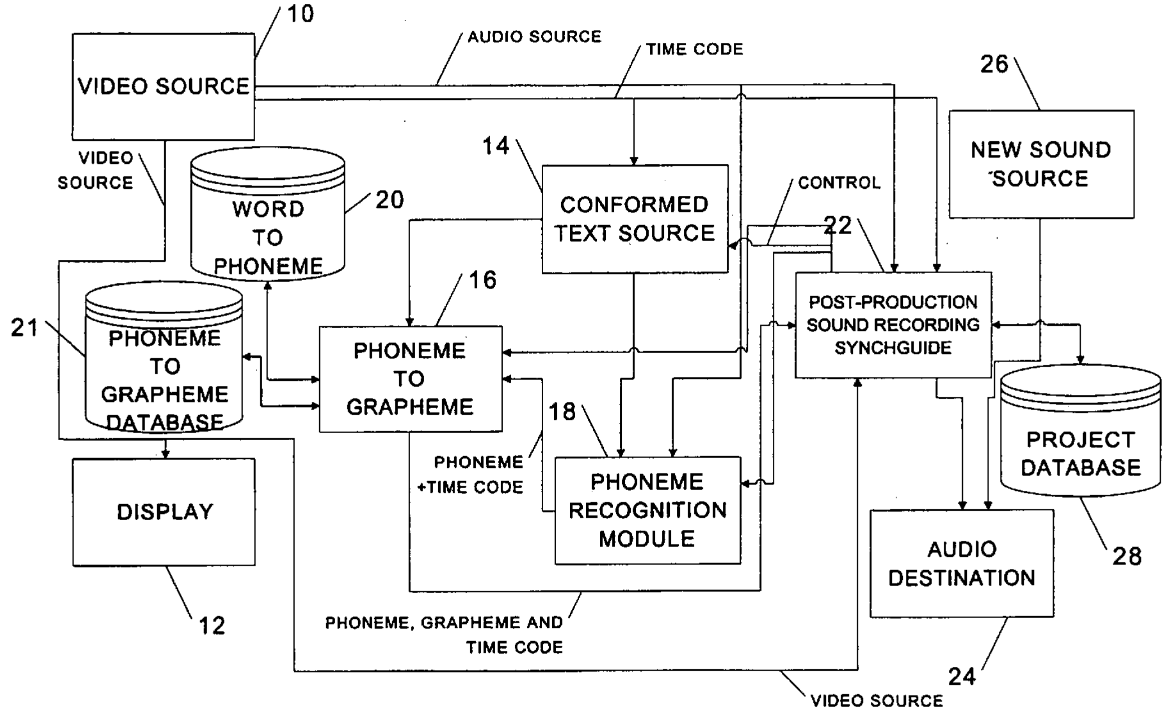
**Publication Classification**

(51) **Int. Cl.**$^7$ .......................... **H04N 5/781**; G10L 11/00; G10L 21/00

(52) **U.S. Cl.** ............................................. **386/96**; 386/125

(57) **ABSTRACT**

A method and apparatus is disclosed for producing an audiovisual work. The method and apparatus is based on speech recognition. Extraction of basic units of speech with related time code is performed. The invention may be advantageously used for performing post-production synchronization of a video source, dubbing assisting, closed-captioning assisting and animation generation assisting.



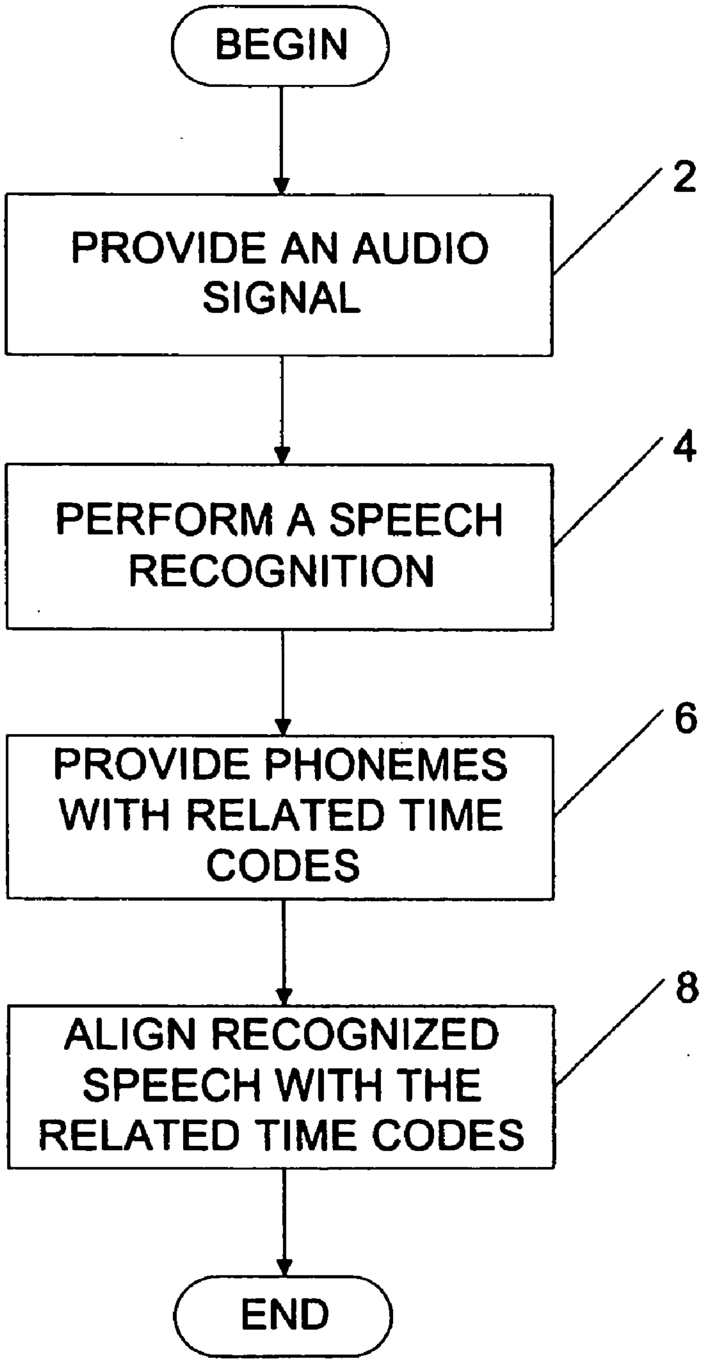

BEGIN
PROVIDE AN AUDIO SIGNAL
2
PERFORM A SPEECH RECOGNITION
4
PROVIDE PHONEMES WITH RELATED TIME CODES
6
ALIGN RECOGNIZED SPEECH WITH THE RELATED TIME CODES
8
END

FIG. 1

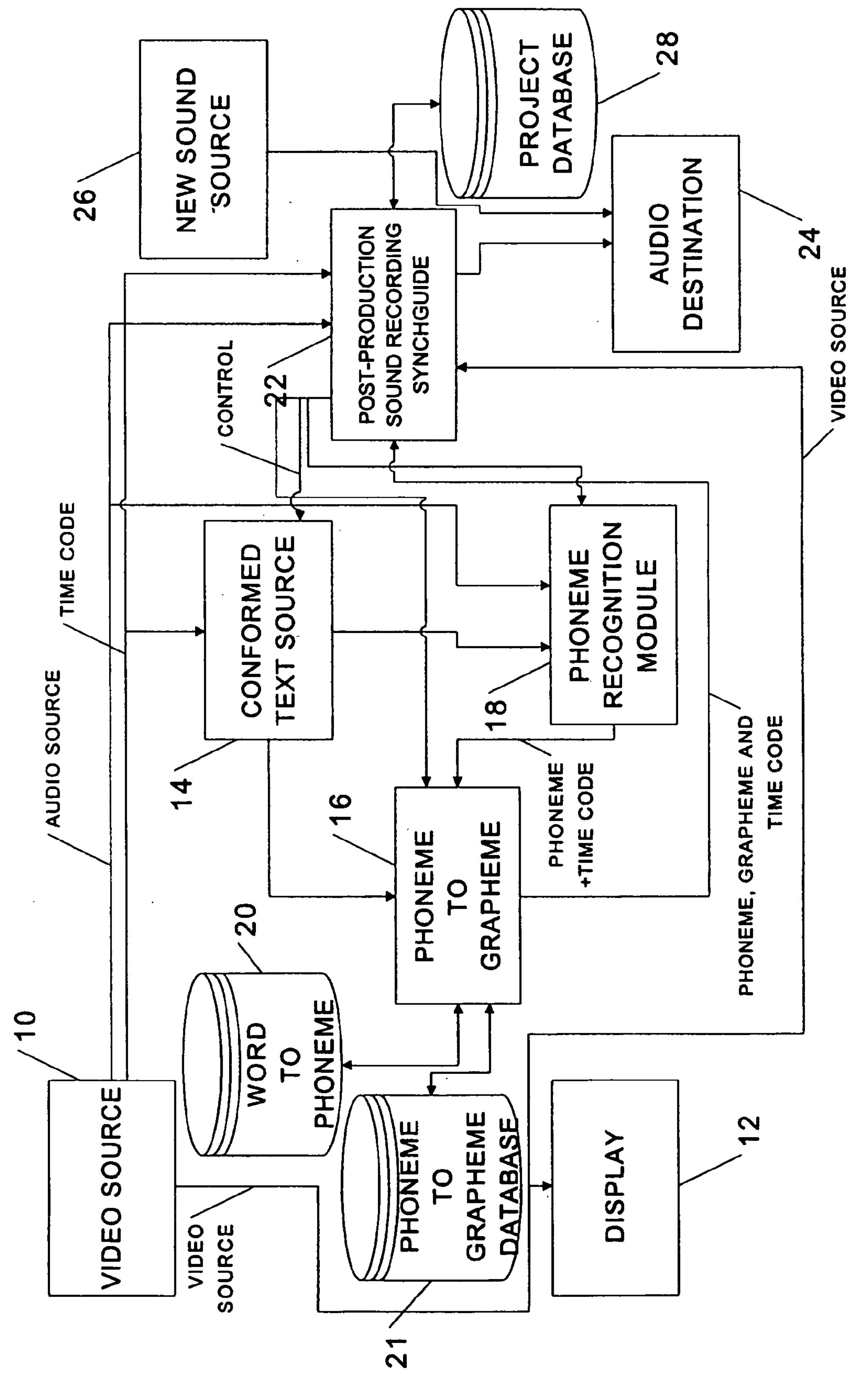

VIDEO SOURCE
10
VIDEO SOURCE
AUDIO SOURCE
TIME CODE
CONFORMED TEXT SOURCE
14
CONTROL
22
POST-PRODUCTION SOUND RECORDING SYNCHGUIDE
NEW SOUND SOURCE
26
PROJECT DATABASE
28
AUDIO DESTINATION
24
VIDEO SOURCE
PHONEME RECOGNITION MODULE
18
PHONEME +TIME CODE
PHONEME, GRAPHEME AND TIME CODE
PHONEME TO GRAPHEME
16
WORD TO PHONEME
20
PHONEME TO GRAPHEME DATABASE
21
DISPLAY
12

FIG. 2

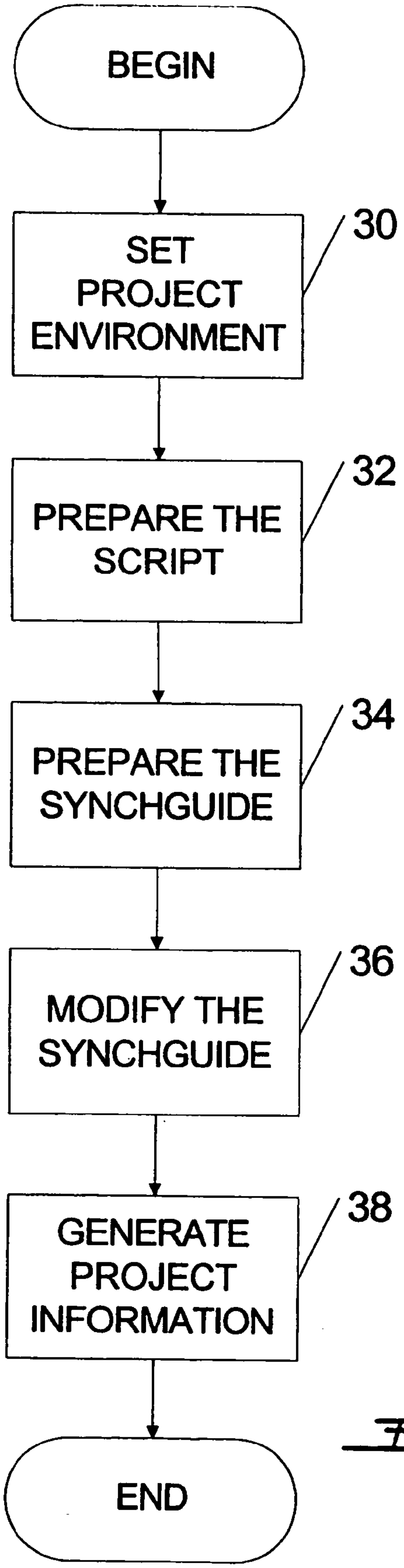

BEGIN
SET PROJECT ENVIRONMENT
30
PREPARE THE SCRIPT
32
PREPARE THE SYNCHGUIDE
34
MODIFY THE SYNCHGUIDE
36
GENERATE PROJECT INFORMATION
38
END

FIG. 3

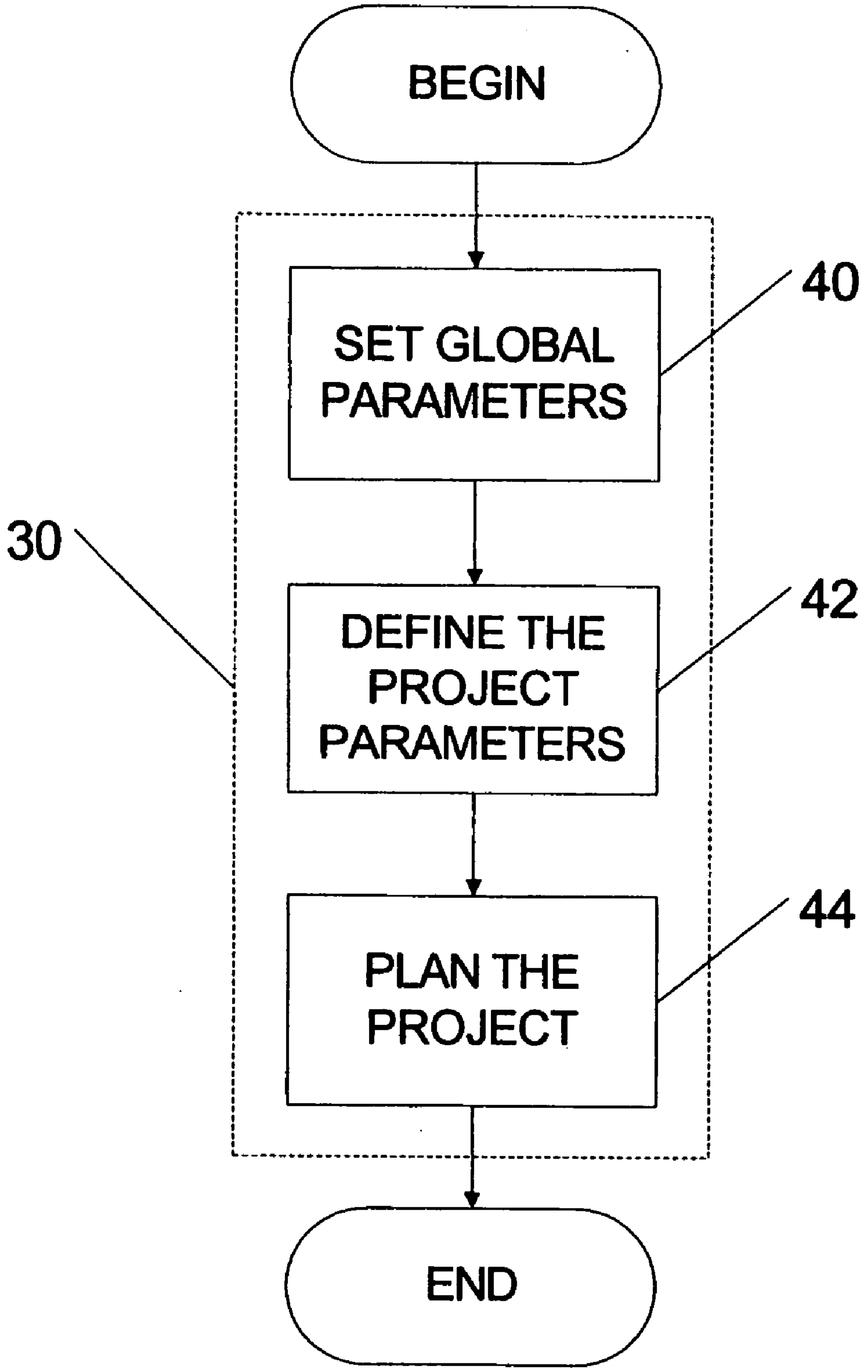

BEGIN
SET GLOBAL PARAMETERS
40
30
DEFINE THE PROJECT PARAMETERS
42
PLAN THE PROJECT
44
END

FIG. 4

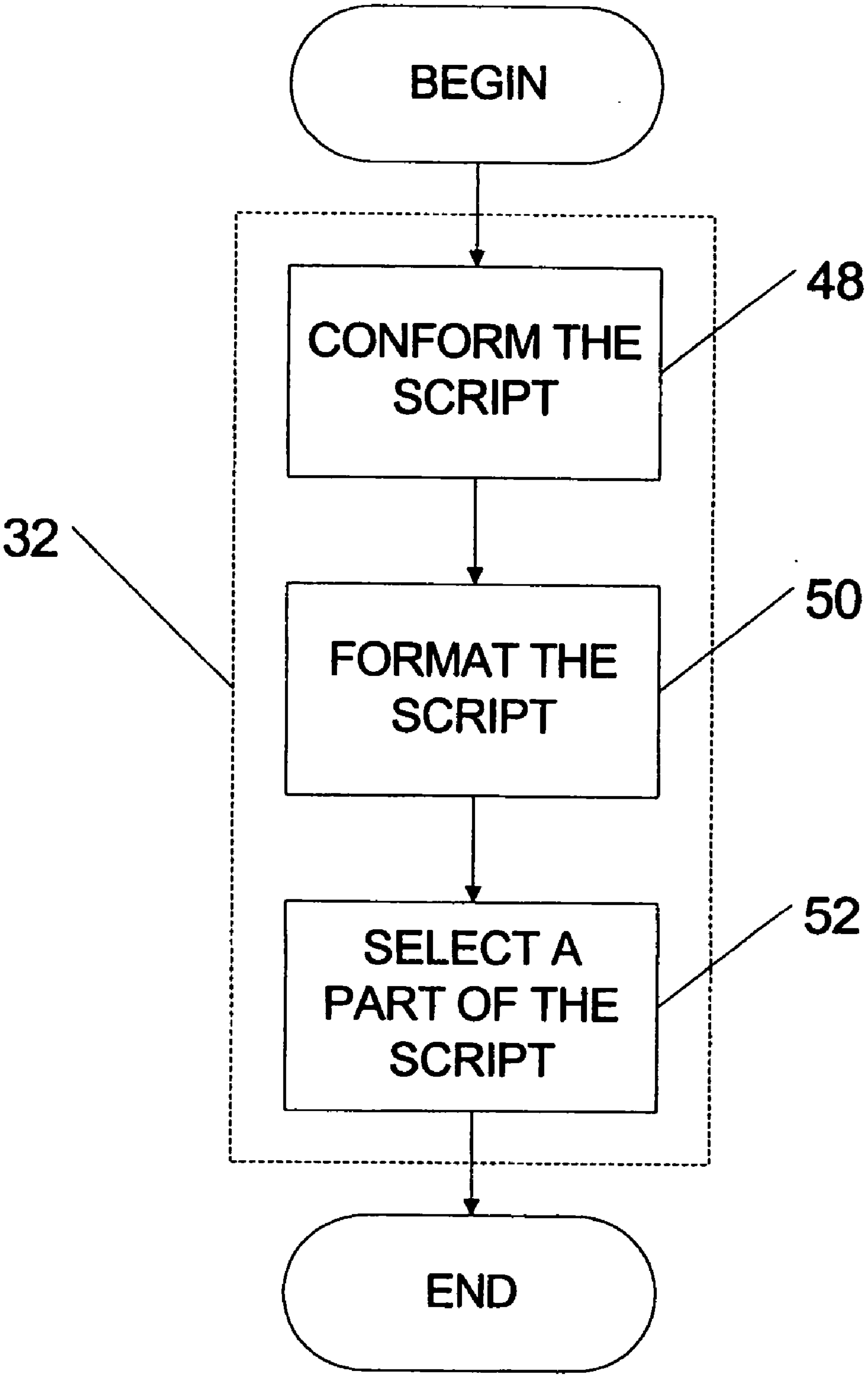

BEGIN
CONFORM THE SCRIPT
48
32
FORMAT THE SCRIPT
50
SELECT A PART OF THE SCRIPT
52
END

FIG. 5

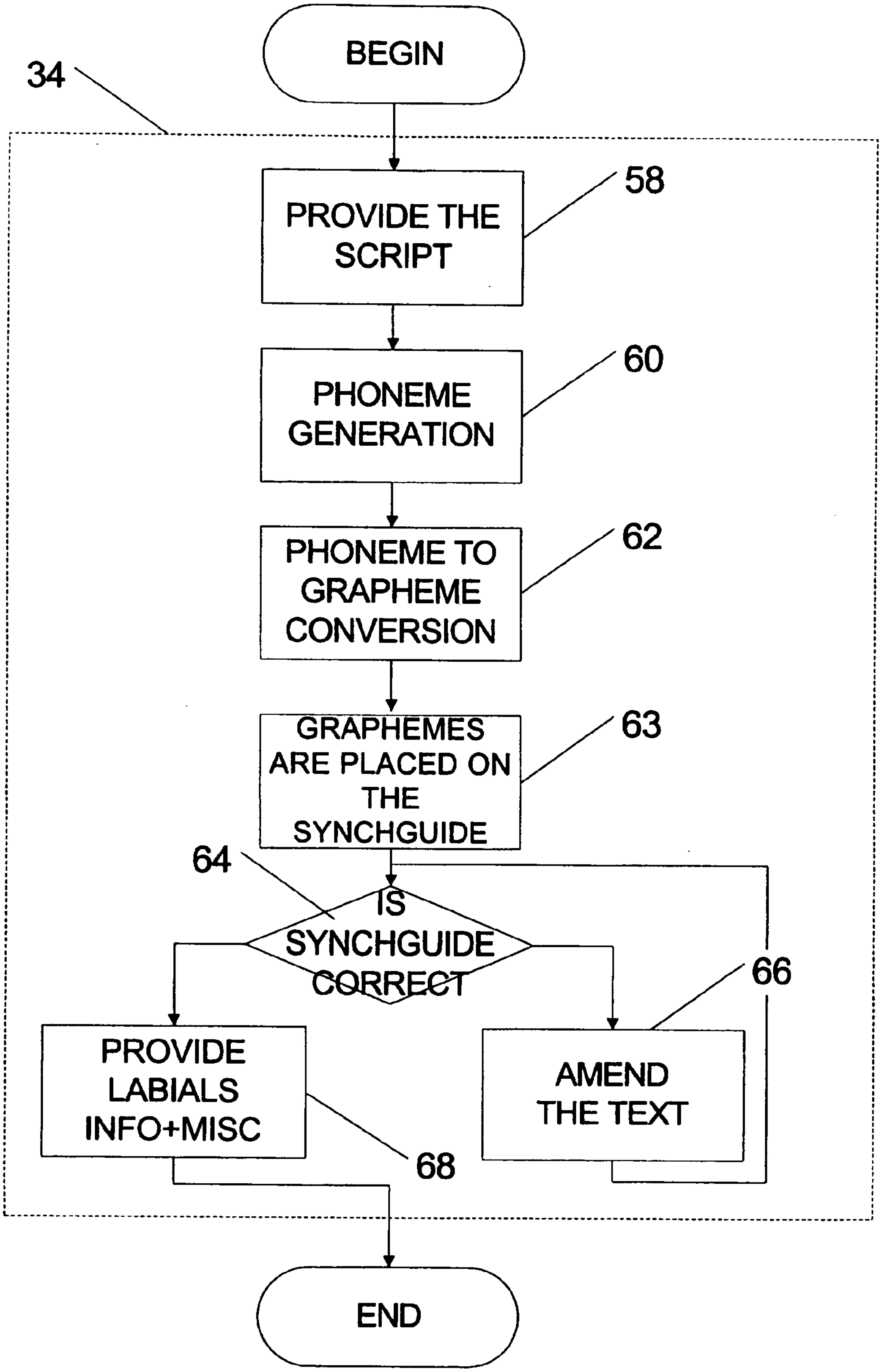

BEGIN
34
PROVIDE THE SCRIPT
58
PHONEME GENERATION
60
PHONEME TO GRAPHEME CONVERSION
62
GRAPHEMES ARE PLACED ON THE SYNCHGUIDE
63
64
IS SYNCHGUIDE CORRECT
66
PROVIDE LABIALS INFO+MISC
AMEND THE TEXT
68
END

FIG. 6

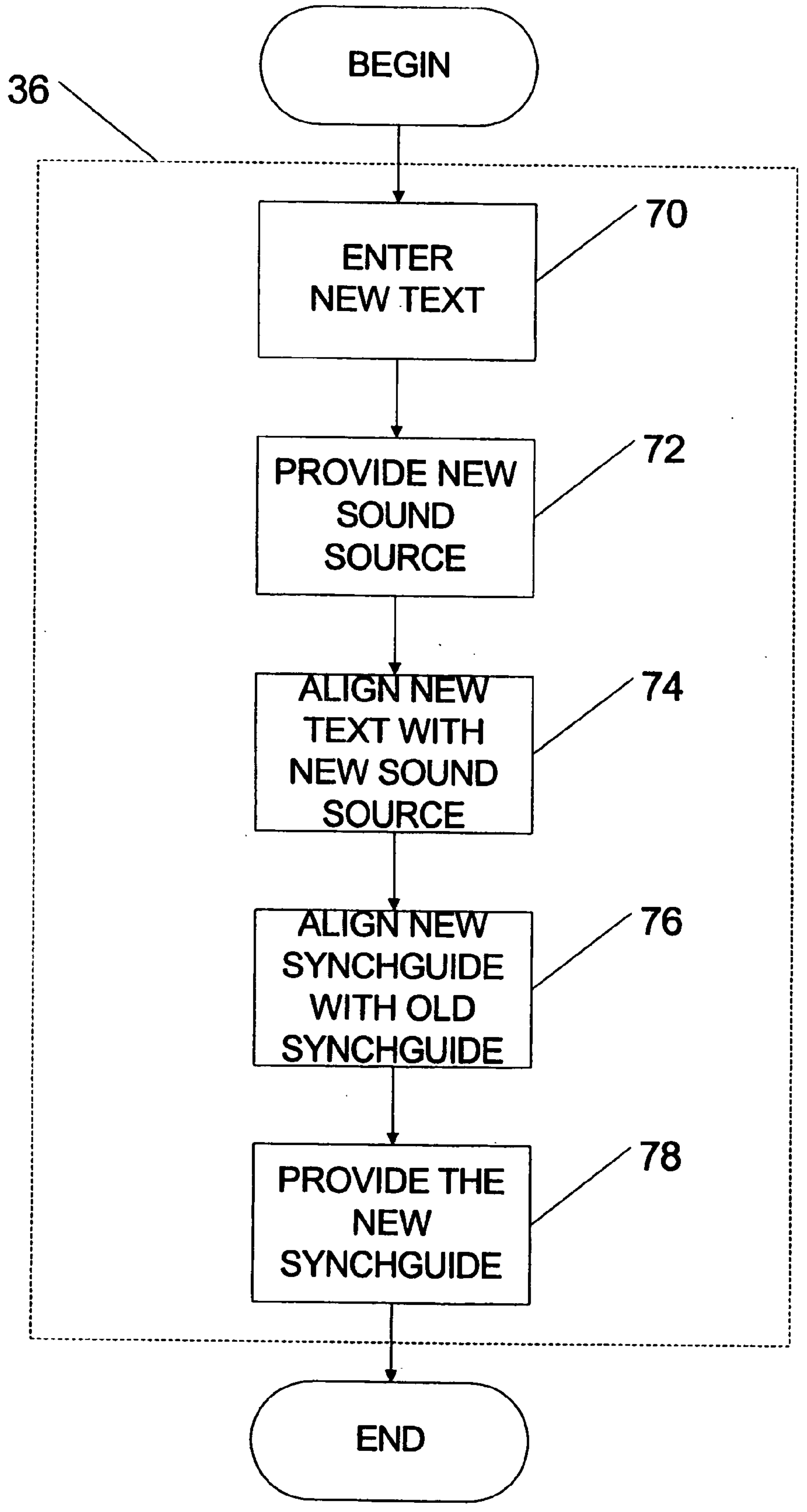
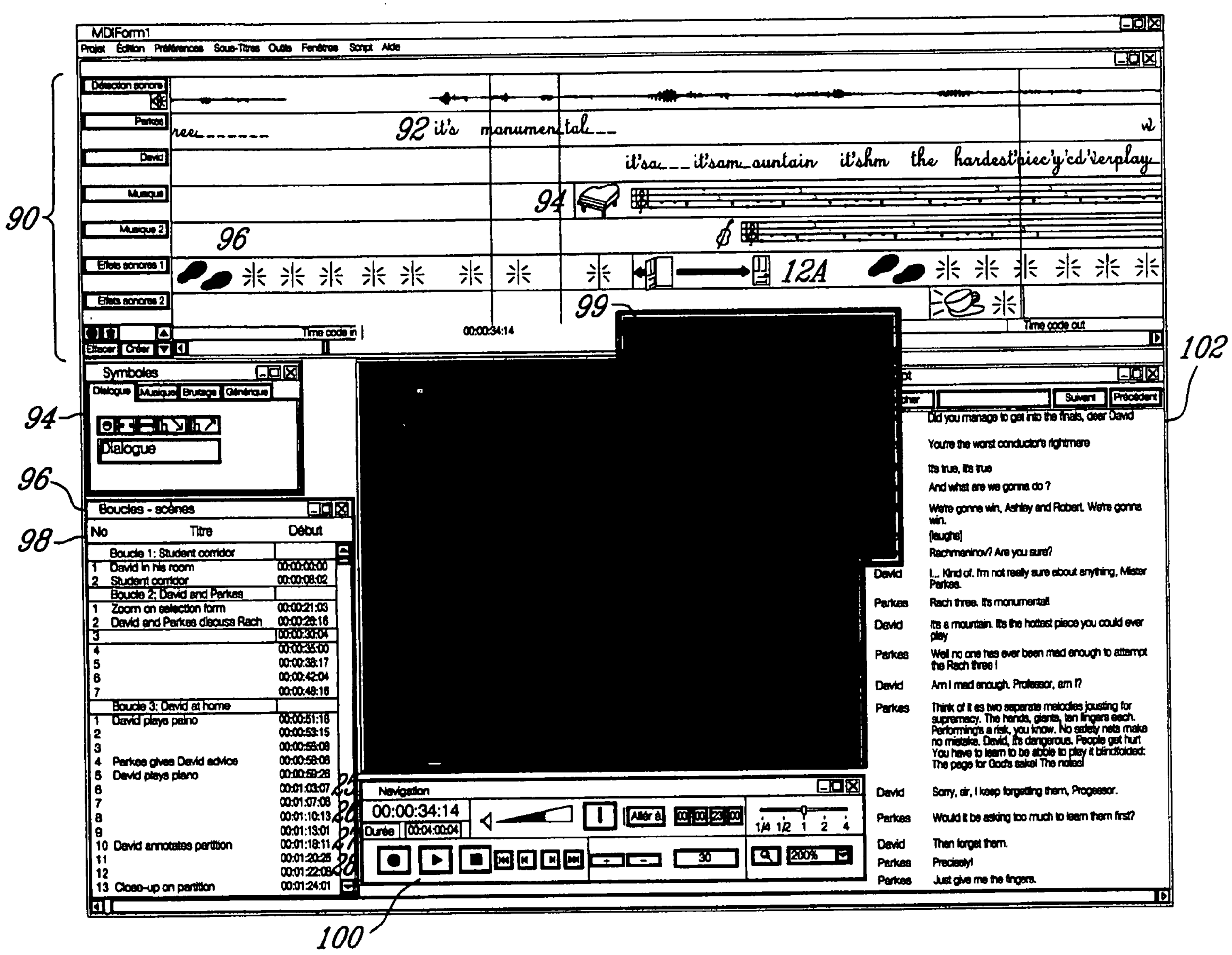

BEGIN
36
ENTER NEW TEXT
70
PROVIDE NEW SOUND SOURCE
72
ALIGN NEW TEXT WITH NEW SOUND SOURCE
74
ALIGN NEW SYNCHGUIDE WITH OLD SYNCHGUIDE
76
PROVIDE THE NEW SYNCHGUIDE
78
END

FIG. 7



| No | Titre | Début |
|---|---|---|
| | Boucle 1: Student corridor | |
| 1 | David in his room | 00:00:00:00 |
| 2 | Student corridor | 00:00:08:02 |
| | Boucle 2: David and Parkes | |
| 1 | Zoom on selection form | 00:00:21:03 |
| 2 | David and Parkes discuss Rach | 00:00:28:16 |
| 3 | | 00:00:30:04 |
| 4 | | 00:00:35:00 |
| 5 | | 00:00:38:17 |
| 6 | | 00:00:42:04 |
| 7 | | 00:00:48:16 |
| | Boucle 3: David at home | |
| 1 | David plays paino | 00:00:51:18 |
| 2 | | 00:00:53:15 |
| 3 | | 00:00:55:08 |
| 4 | Parkes gives David advice | 00:00:58:08 |
| 5 | David plays piano | 00:00:59:28 |
| 6 | | 00:01:03:07 |
| 7 | | 00:01:07:08 |
| 8 | | 00:01:10:13 |
| 9 | | 00:01:13:01 |
| 10 | David annotates partition | 00:01:18:11 |
| 11 | | 00:01:20:25 |
| 12 | | 00:01:22:08 |
| 13 | Close-up on partition | 00:01:24:01 |

Did you manage to get into the finals, dear David

You're the worst conductor's rightmare

It's true, it's true

And what are we gonna do ?

We're gonna win, Ashley and Robert. We're gonna win.

[laughs]

Rachmaninov? Are you sure?

David I... Kind of. I'm not really sure about anything, Mister Parkes.

Parkes Rach three. It's monumental!

David It's a mountain. It's the hottest piece you could ever play

Parkes Well no one has ever been mad enough to attempt the Rach three !

David Am I mad enough. Professor, am I?

Parkes Think of it as two separate melodies jousting for supremacy. The hands, giants, ten fingers each. Performing's a risk, you know. No safety nets make no mistake. David, it's dangerous. People get hurt You have to learn to be able to play it blindfolded: The page for God's sake! The notes!

David Sorry, sir, I keep forgetting them, Progessor.

Parkes Would it be asking too much to learn them first?

David Then forget them.

Parkes Precisely!

Parkes Just give me the fingers.

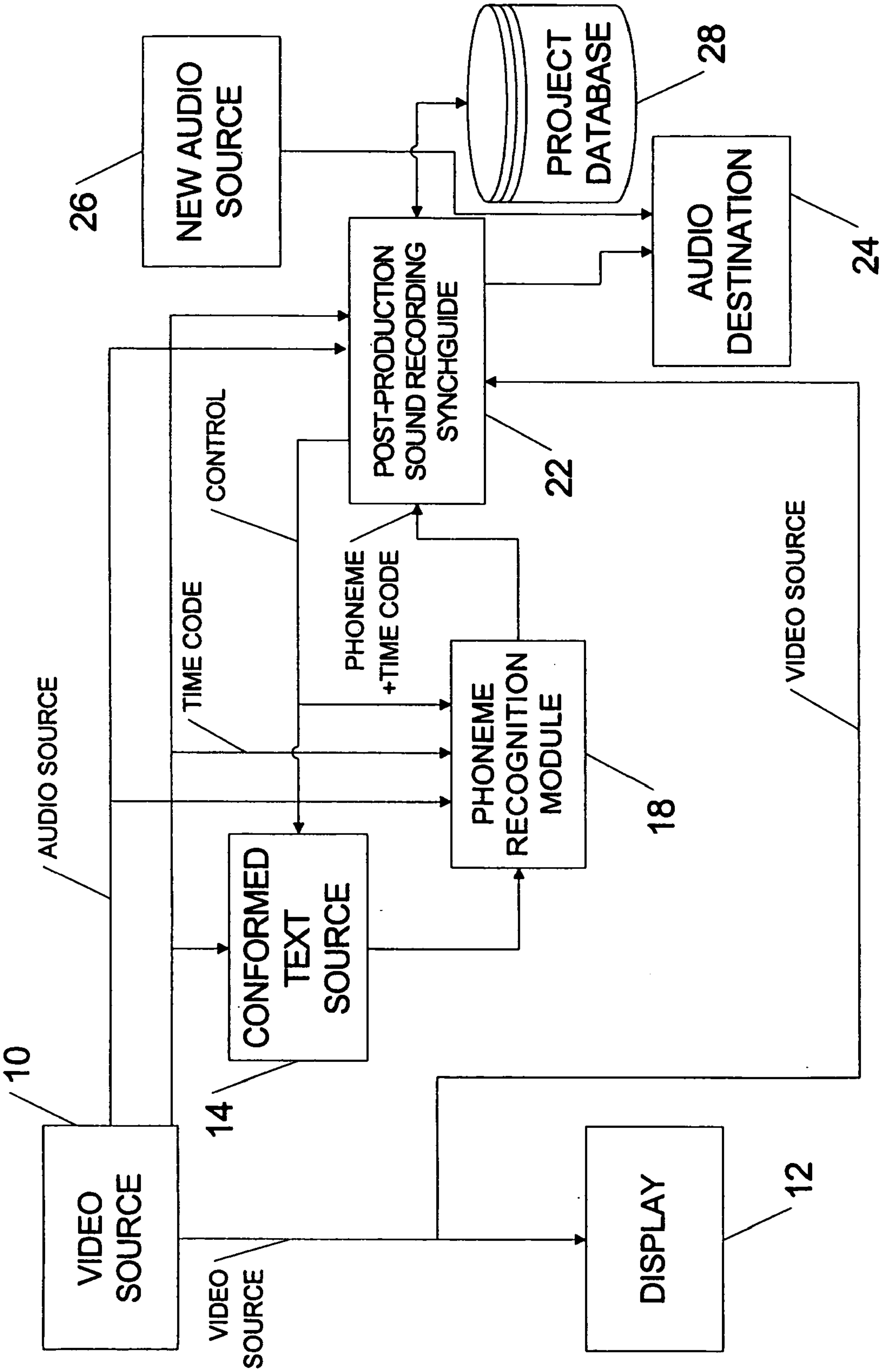

VIDEO SOURCE
10
VIDEO SOURCE
DISPLAY
12
CONFORMED TEXT SOURCE
14
AUDIO SOURCE
TIME CODE
CONTROL
PHONEME RECOGNITION MODULE
18
PHONEME +TIME CODE
POST-PRODUCTION SOUND RECORDING SYNCHGUIDE
22
VIDEO SOURCE
NEW AUDIO SOURCE
26
PROJECT DATABASE
28
AUDIO DESTINATION
24

FIG - 9

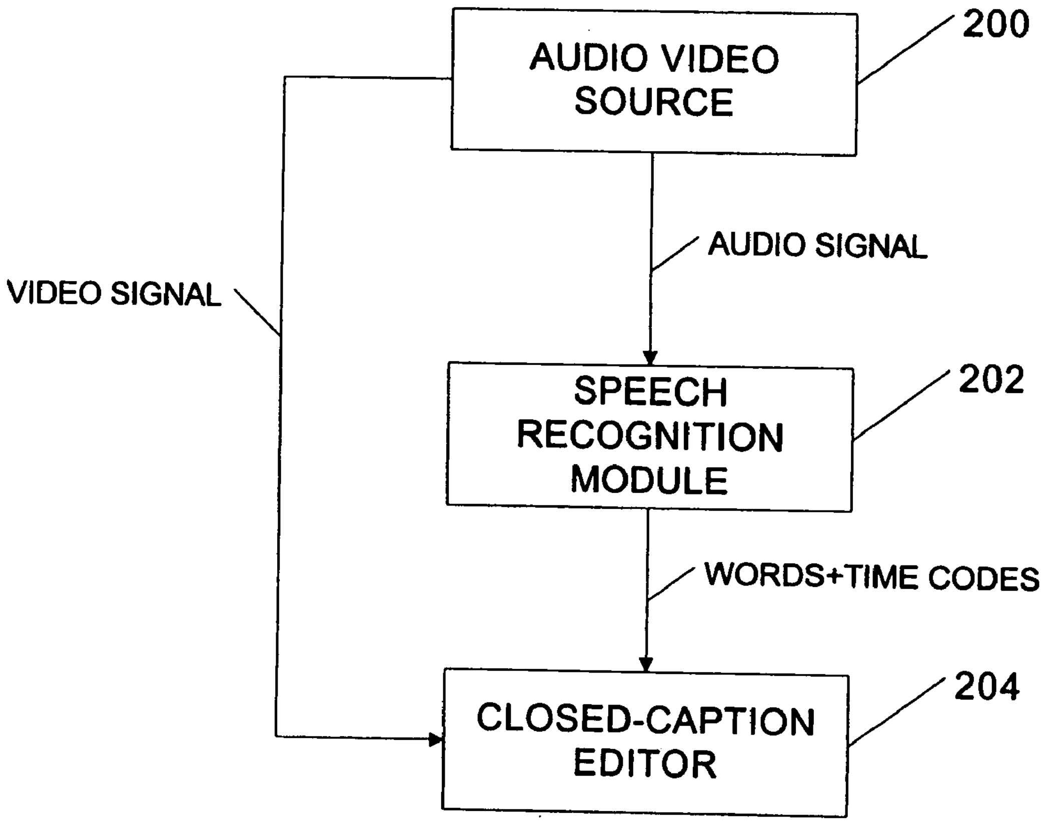

AUDIO VIDEO SOURCE
200
AUDIO SIGNAL
VIDEO SIGNAL
SPEECH RECOGNITION MODULE
202
WORDS+TIME CODES
CLOSED-CAPTION EDITOR
204

FIG. 10

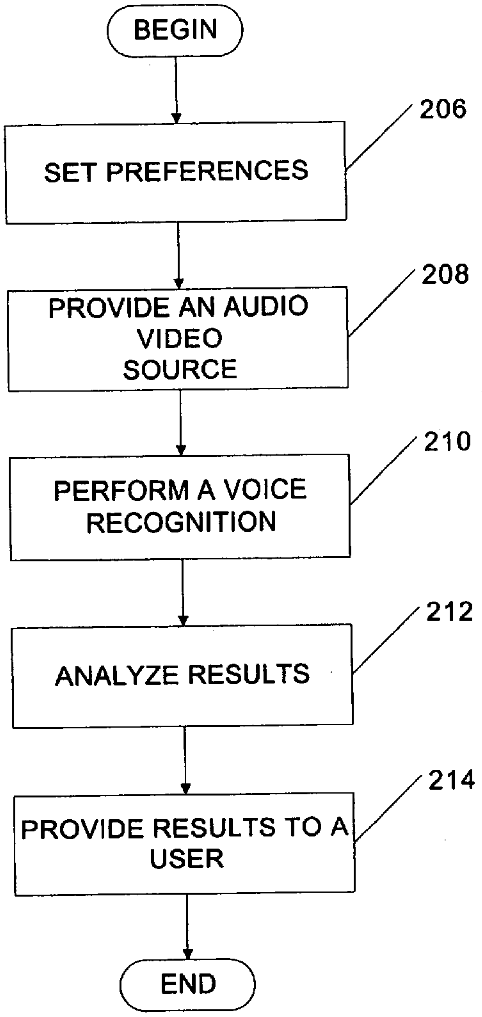

BEGIN
206
SET PREFERENCES
208
PROVIDE AN AUDIO VIDEO SOURCE
210
PERFORM A VOICE RECOGNITION
212
ANALYZE RESULTS
214
PROVIDE RESULTS TO A USER
END

FIG-11

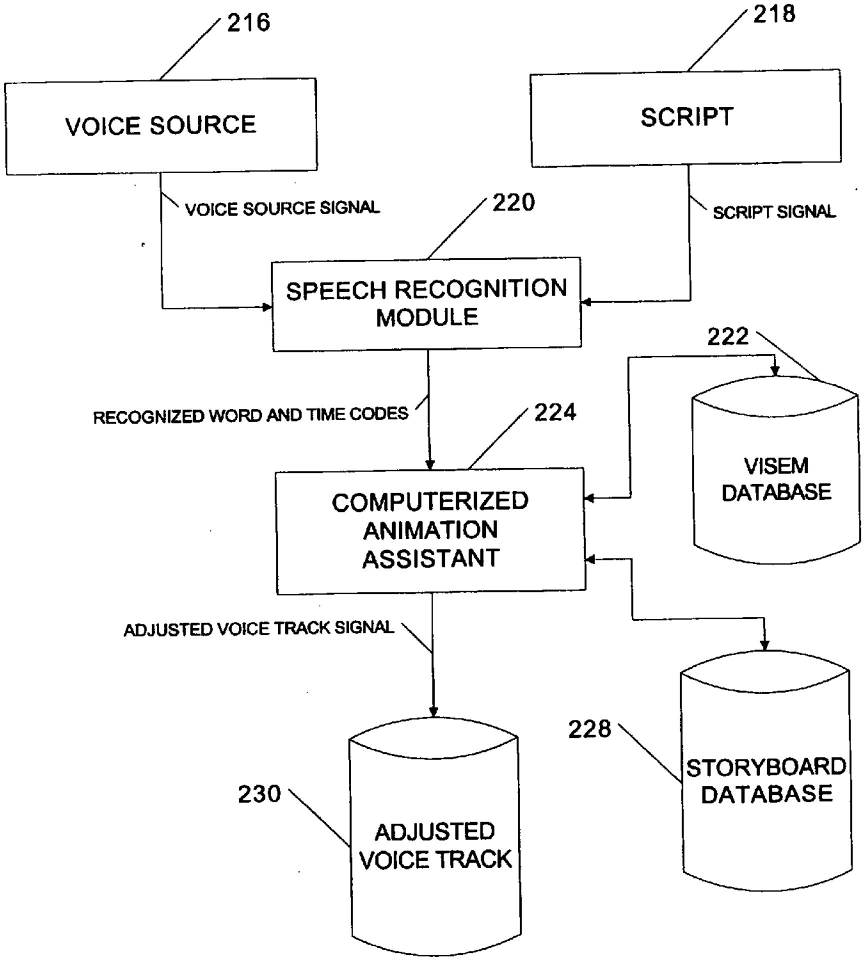

216
VOICE SOURCE
218
SCRIPT
VOICE SOURCE SIGNAL
220
SCRIPT SIGNAL
SPEECH RECOGNITION MODULE
222
RECOGNIZED WORD AND TIME CODES
224
VISEM DATABASE
COMPUTERIZED ANIMATION ASSISTANT
ADJUSTED VOICE TRACK SIGNAL
228
STORYBOARD DATABASE
230
ADJUSTED VOICE TRACK

FIG. 12

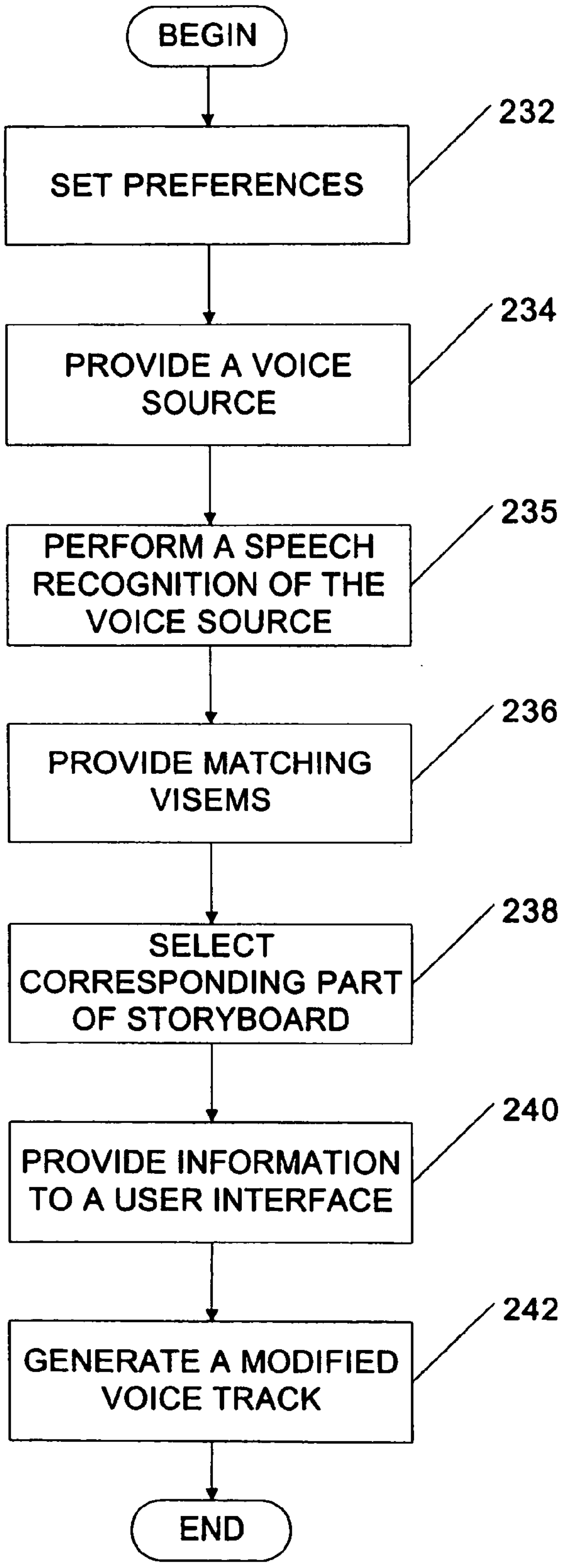

BEGIN
SET PREFERENCES
232
PROVIDE A VOICE SOURCE
234
PERFORM A SPEECH RECOGNITION OF THE VOICE SOURCE
235
PROVIDE MATCHING VISEMS
236
SELECT CORRESPONDING PART OF STORYBOARD
238
PROVIDE INFORMATION TO A USER INTERFACE
240
GENERATE A MODIFIED VOICE TRACK
242
END

FIG. 13

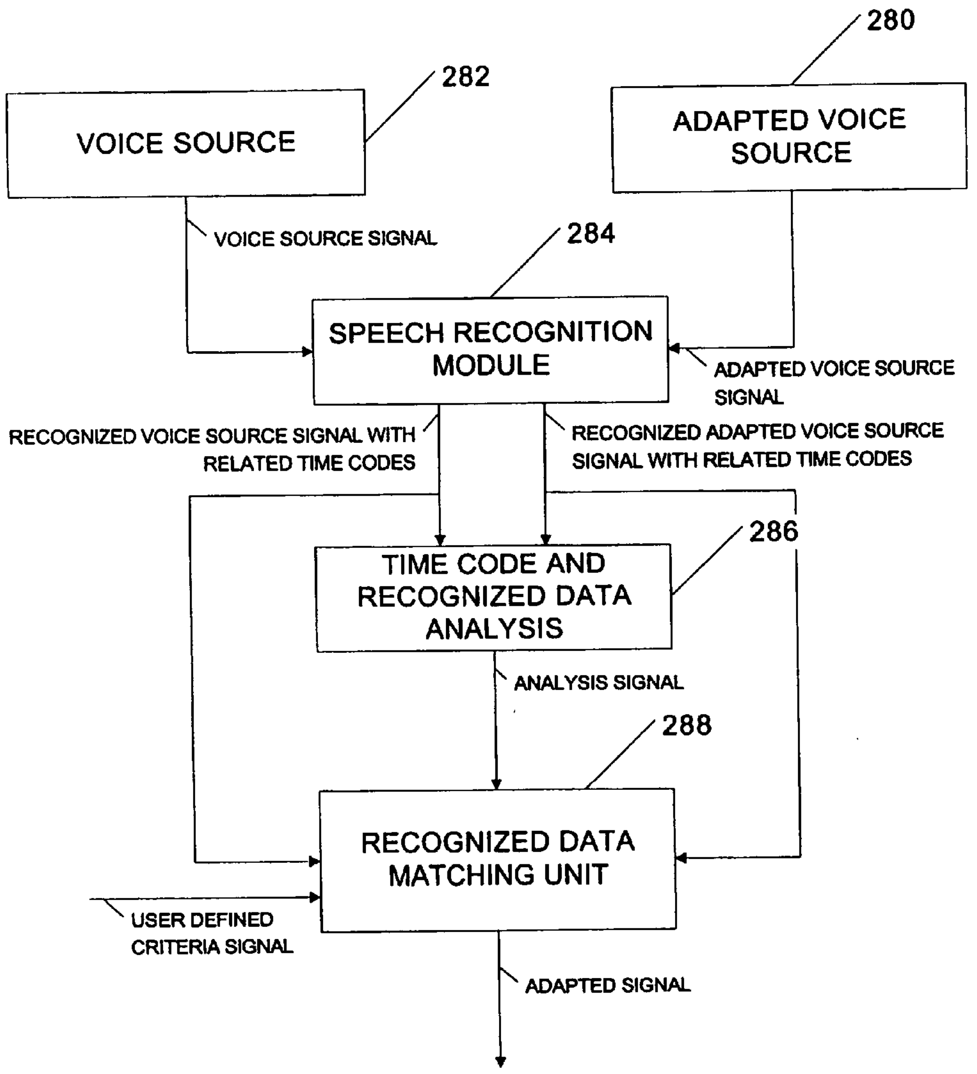

280
282
VOICE SOURCE
ADAPTED VOICE SOURCE
VOICE SOURCE SIGNAL
284
SPEECH RECOGNITION MODULE
ADAPTED VOICE SOURCE SIGNAL
RECOGNIZED VOICE SOURCE SIGNAL WITH RELATED TIME CODES
RECOGNIZED ADAPTED VOICE SOURCE SIGNAL WITH RELATED TIME CODES
286
TIME CODE AND RECOGNIZED DATA ANALYSIS
ANALYSIS SIGNAL
288
RECOGNIZED DATA MATCHING UNIT
USER DEFINED CRITERIA SIGNAL
ADAPTED SIGNAL

FIG. 14

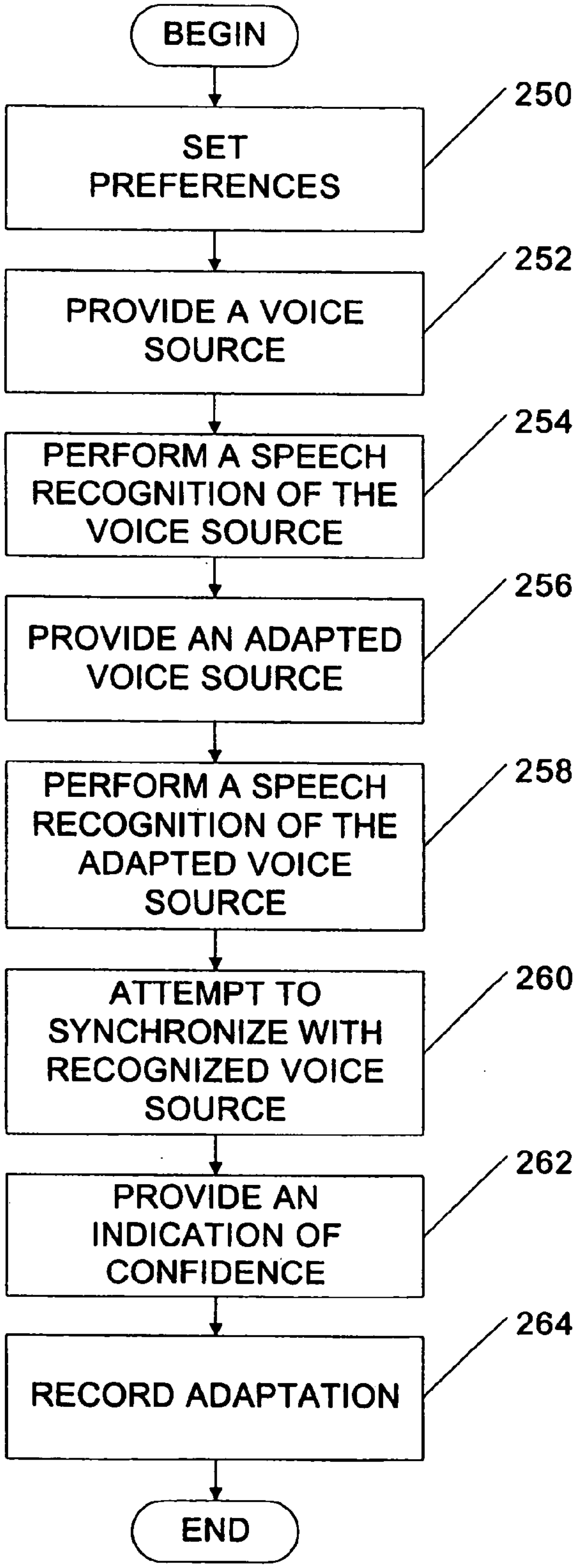

BEGIN
SET PREFERENCES
250
PROVIDE A VOICE SOURCE
252
PERFORM A SPEECH RECOGNITION OF THE VOICE SOURCE
254
PROVIDE AN ADAPTED VOICE SOURCE
256
PERFORM A SPEECH RECOGNITION OF THE ADAPTED VOICE SOURCE
258
ATTEMPT TO SYNCHRONIZE WITH RECOGNIZED VOICE SOURCE
260
PROVIDE AN INDICATION OF CONFIDENCE
262
RECORD ADAPTATION
264
END

FIG. 15

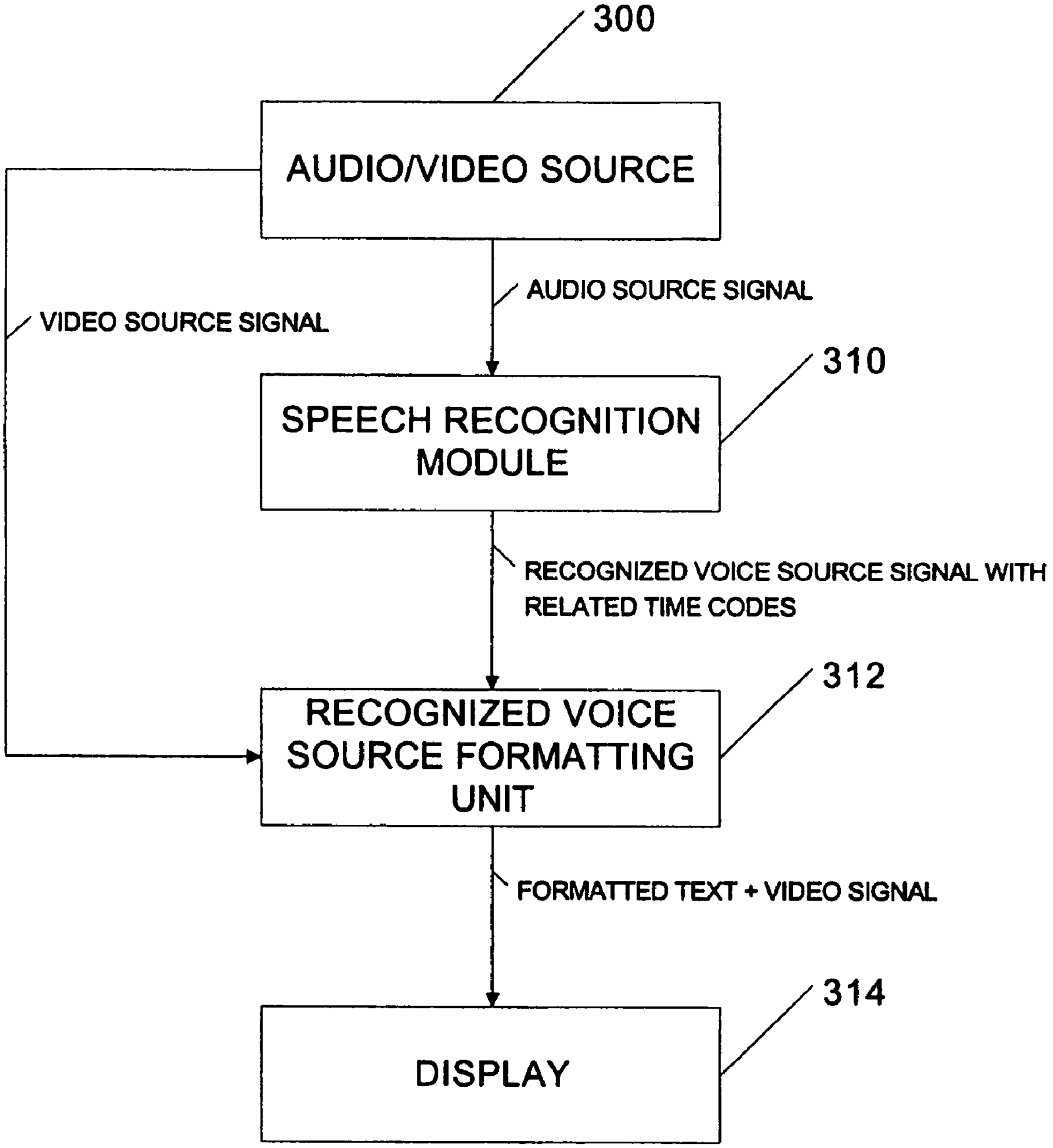

300
AUDIO/VIDEO SOURCE
AUDIO SOURCE SIGNAL
VIDEO SOURCE SIGNAL
310
SPEECH RECOGNITION MODULE
RECOGNIZED VOICE SOURCE SIGNAL WITH RELATED TIME CODES
312
RECOGNIZED VOICE SOURCE FORMATTING UNIT
FORMATTED TEXT + VIDEO SIGNAL
314
DISPLAY

FIG. 16

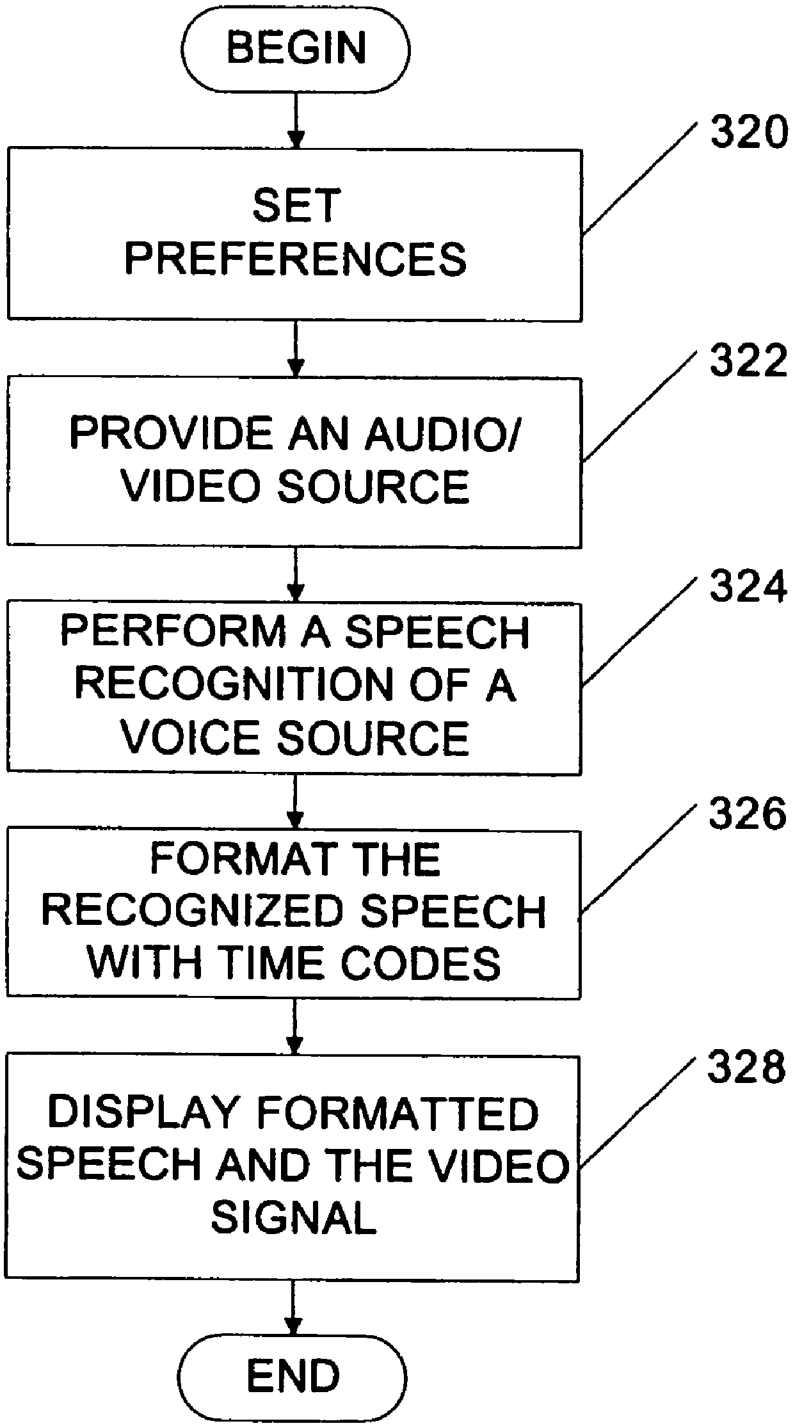

BEGIN
320
SET PREFERENCES
322
PROVIDE AN AUDIO/ VIDEO SOURCE
324
PERFORM A SPEECH RECOGNITION OF A VOICE SOURCE
326
FORMAT THE RECOGNIZED SPEECH WITH TIME CODES
328
DISPLAY FORMATTED SPEECH AND THE VIDEO SIGNAL
END

FIG. 17

# METHOD AND APPARATUS FOR PERFORMING AN AUDIOVISUAL WORK USING SYNCHRONIZED SPEECH RECOGNITION DATA

[0001] The present application is a continuation-in-part of U.S. application Ser. No. 10/067,131 filed on Sep. 12, 2001 designating the United States of America now pending, the specification of which is hereby incorporated by reference. The present application is also a continuation of PCT/CA02/01386 filed on Sep. 12, 2002, designating the United States, now pending and the specification of which is hereby incorporated by reference.

## FIELD OF THE INVENTION

[0002] This invention pertains to the field of what is commonly referred to as speech recognition. More precisely, this invention provides a method and an apparatus for performing an audiovisual work using synchronized recognition data.

## BACKGROUND OF THE INVENTION

[0003] The system for post-synchronization that is used throughout most of the world is based on what is called a "beep-and-wipe" system. In a recording studio, the actor is given earphones, through which the dialog is fed.

[0004] An audible beep is sent as a signal to signify the beginning of the line to be re-recorded. A visual indicator, called a wipe, is superimposed on the screen as a visual indication of when to begin and stop. A series of takes are recorded, sometimes as many as 24, and are given to the editor in order to verify by eye or by trying to match the sound waves of the original production take with that of the newly recorded ones. Most of the editing is, in the end, totally dependent on the experienced eye and ear of the human operators. The method used for film dubbing in the greater part of the world is the same, except in the United States where the voice of the translator is fed into one of the earphones while the other carries the mixed track of dialog from the original language. The norm for the recording of dialog using this method is between ten to twelve lines of text per hour of studio time.

[0005] The system used in France, Quebec, and South Africa consists in taking the film that is to be post-synchronized (or dubbed) and transferring it to either a three quarter inch or a half inch video tape. The video is fed from a VCR to a special machine, called a detection machine, that links a roll of white 35 mm leader film with the VCR so that they run synchronously with each other. A detection of the scene cuts, and all the lip movements and dialog is then performed of the original language. A highly skilled craftsperson, called a detector, then proceeds to write with a pencil, on the strip of white leader. The detector copies the original language of the film dialog, following the precise movements of the lips and matches them to the spoken word. During this process, a particular emphasis is laid on a precise matching of the labials and semi-labials. A calligrapher then runs a strip of clear 35 mm leader on top, that is matched sprocket to sprocket with the original white strip underneath. The two rolls are then run simultaneously on a small-geared table. After the rolls are locked, the calligrapher proceeds to copy the detection on the clear leader using a special pen and India ink. When this is completed, the calligraphied dialog is typed by a typist into a computer and copies of the text are printed for the director, the recording engineer, and the actors. The problems inherent with this system are that they are inefficient in their consumption of time and "man hours". Approximately 150 "man hours" are needed to complete all the operations for a "feature length film" (i.e. a film ranging from 90 to 100 minutes in running time). Since these operations are dependent upon a number of hands, they are open to errors and inaccuracies in the detection process and the calligraphy. After the recording sessions are completed, an editor works on the dialog tracks, adjusting the synchronization. When that is completed to everyone's satisfaction, a final mix of the tracks is done, and the script is re-conformed and is tabled for distribution.

[0006] The U.S. Pat. No. 5,732,184 teaches a system for the editing of video and audio sequences, and relates only to a system for editing video clips, or small portions of video, and sound clips based on short sections of sound waves displayed on a video screen. The cursor is able to display no more than three frames of video and sound at the same time in one direction or the other. The cursor then becomes an aid to identifying the material only.

[0007] Published GB Patent application GB 2,101,795 relates to dubbing translation of soundtracks on film. This invention depends upon an ability to provide histograms, or a digital representation, of the sound amplitude. Somewhat difficult for the actors, as it is like asking them to learn a whole new alphabet. The invention also suggests that recorded material can be electronically shaped to fit the lip movement in order to produce a more natural speech. Unfortunately, it is known, in light of the current technology, that any reshaping that is not minimal will only distort the sound and will not therefore provide a natural sound. Each section, or loop of film, requires that it is manually operated by a trained user.

[0008] In the French patent publication 2,765,354, a system is disclosed and allows dubbing into French from other languages. This invention is also used to match the new French dialog to the images. Unfortunately, the system disclosed is slow and time consuming, as it is not automatic and requires manual input. It provides a maximum of 6 usable lines on a timeline. Furthermore, it also does not allow any modifications to be made since the dialog has already been permanently encrusted on the picture. It requires the performers to learn a whole new language of symbols different from the symbols normally used in the standard manual form of operation.

[0009] The international publication WO98/101860 provides a fairly simple device that attempts to use a computerized calligraphy of the dialogs. Its primary market is actually the home-entertainment or classroom games market. This device allows the player to substitute their voice for the one on the screen, using a basic recording device.

[0010] The "beep-and-wipe" system (in ADR, or Automatic Dialog Replacement) that is currently used throughout the world, is a system that is learned by performers, who then must develop proficiency for it. Otherwise, it becomes rather tedious, frustrating, and time consuming. Actors must do it instinctively, i.e. they must learn to anticipate when to begin taking into account the fact that it takes the human brain 1/20th of a second to decode what the eyes have seen and then, the time it takes for the actor to respond to what he or she has just seen would put the synchronization out approximately 1½ frames. The amount of text that can be said by the actor is limited in terms of time because it is based on the individual actor's retentive powers. The actor who begins his line late realizes it, and tries to catch up by the end of the sentence, making it very difficult to edit. This means that many takes have to be recorded, causing the editor to spend large quantities of time piecing together the final take. The time required by, not only the actor but by the director, the studio engineer, the editor, plus the cost of the studio itself will only create a greater expense of both time and money. An expense that could be avoided.

**[0011]** Spot editing is the editing in the studio by the studio engineer, who tries to match or tailor the waveforms of the original dialog with the newly recorded one. While some spot editing can be done in studio by trying to match waveforms, the drawbacks to this are that it requires some training and knowledge in the ability to read the waveforms so as to be able to properly match them, and also if there is too much variation in the tailoring of the waveforms, it will ultimately cause a distortion in the sound.

**[0012]** The human factor is very important in the current post-synchronization methods used around the world. Operators must be highly trained. Experienced operators are therefore needed as such methods rely on the capacity of the operators to interact and to react with the system, therefore the quality of the post-synchronization performed may vary from time to time. Furthermore these methods are very time consuming, and therefore are very costly.

**[0013]** Accordingly, there is a need for a method and apparatus that will overcome the above-mentioned drawbacks.

## SUMMARY OF THE INVENTION

**[0014]** It is an object of the invention to provide a method and apparatus for achieving a synchronization of speech recognition data with time.

**[0015]** It is another object of the invention to provide a method and apparatus for achieving post-production synchronization for film and video that will enable an operator to anticipate dialog.

**[0016]** Yet another object of the invention is to provide a method and apparatus for achieving post-production synchronization for film and video without repeatedly moving backward in time.

**[0017]** It is another object of the invention to assist animation production.

**[0018]** It is another object of the invention to assist karaoke production.

**[0019]** Yet another object of the invention is to assist adaptation of an audiovisual work.

**[0020]** Yet another object of the invention is to assist closed-caption generation.

**[0021]** According to an aspect of the invention, there is provided a method for producing an audiovisual work, the method comprising the steps of providing an audio signal to a speech recognition module, performing a speech recognition of said audio signal, the speech recognition comprising an extracting of a plurality of basic units of recognized speech and related time codes, receiving the plurality of basic units of recognized speech and the related time codes from the speech recognition module, processing the received plurality of basic units to provide synchronization information for a production of said audiovisual work, and displaying on a user interface said synchronization information.

**[0022]** According to another aspect of the invention, there is provided a method for performing closed-captioning of an audio source, the method comprising the steps of providing an audio signal of an audio/video signal to a speech recognition module, performing a speech recognition of said audio/video signal, and incorporating text of said recognized speech of the audio signal as closed-captioning into a visual or non-visual portion of the audio/video signal in synchronization.

**[0023]** According to another aspect of the invention, there is provided an apparatus for producing an audiovisual work comprising a speech recognition module receiving an audio signal and providing a plurality of basic units of recognized speech and related time codes, means for processing the plurality of basic units to provide synchronization information for a production of said audiovisual work; and means for displaying on a user interface said synchronization information.

**[0024]** According to another aspect of the invention, there is provided an apparatus for performing closed-captioning of an audio source, the apparatus comprising a speech recognition module receiving an audio signal and providing recognized speech, incorporating means for incorporating text of said recognized speech of the audio signal as closed-captioning into a visual or non-visual portion of the audio/video signal in synchronization.

**[0025]** A "rythmo band" is a clear band of 35 mm which is written in India ink by a calligrapher and is projected in the recording studio by means of a specifically designed overhead projector and is run locked in synchronization with a projected image. The "rythmo band" comprises the script and the translated script in the case of film dubbing.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0026]** The invention will be better understood by way of the following description of the preferred embodiment, together with the accompanying drawings, in which:

**[0027]** **FIG. 1** is a flow chart of the preferred embodiment of the invention;

**[0028]** **FIG. 2** is a block diagram of one embodiment of the invention; the apparatus comprises a video source, a conformed text source, a phoneme to grapheme unit, a phoneme recognition module, a post-production sound recording synchguide, a new sound source, a project database;

**[0029]** **FIG. 3** is a flow chart of one embodiment of the invention;

**[0030]** **FIG. 4** is a flow chart which shows how a project environment is set;

**[0031]** **FIG. 5** is a flow chart which shows how the script is prepared;

**[0032]** **FIG. 6** is a flow chart which shows how the synchguide is prepared;

[0033] FIG. 7 is a flow chart which shows how the synchguide may be changed;

[0034] FIG. 8 is screenshot which shows the user interface in one embodiment of the invention;

[0035] FIG. 9 is a block diagram of another embodiment of the invention; in this embodiment, the apparatus comprises a video source, a conformed text source, a phoneme recognition module; a post-production sound recording synchguide, a new sound source, a project database;

[0036] FIG. 10 is a block diagram of another embodiment of the invention for assisting closed-caption generation; in this embodiment, the apparatus comprises an audio video source, a speech recognition module and a closed-caption editor;

[0037] FIG. 11 is a flow chart which shows how the assisting closed-caption generation is performed;

[0038] FIG. 12 is a block diagram of another embodiment of the invention for assisting animation production; the apparatus comprises a speech recognition module, a computerized animation assistant, a visem database, a storyboard database;

[0039] FIG. 13 is a flow chart which shows how the assisting animation production is performed;

[0040] FIG. 14 is a block diagram which shows another embodiment of the invention for assisting adaptation of an audiovisual work;

[0041] FIG. 15 is a flow chart which shows another embodiment of the invention for assisting adaptation of an audiovisual work;

[0042] FIG. 16 is a block diagram which shows another embodiment of the invention for assisting Karaoke generation; and

[0043] FIG. 17 is a flow chart which shows another embodiment of the invention for assisting Karaoke generation.

## DETAILED DESCRIPTION

[0044] In the present application, the word "synchguide" will be introduced and will relate to an extended version of the concept of a "rythmo band".

[0045] Now referring to FIG. 1, there is shown the preferred embodiment of the invention.

[0046] According to step 2, an audio signal is provided. The audio signal comprises at least speech data. The audio signal may further comprise Foley. The audio signal originates from an audio signal source.

[0047] According to step 4, speech recognition is performed using the audio signal provided by the audio signal source. In one embodiment of the invention, the speech recognition provides an element indicative of a speech source.

[0048] According to step 6, basic units of recognized speech are extracted with related time codes. In the preferred embodiment, the basic units are phonemes of the recognized speech. In the preferred embodiment of the invention, each phoneme is provided with two related time codes indicative of a beginning of the phoneme and of a end of the phoneme with respect to a common time origin.

[0049] According to step 8, the recognized phonemes and the related time codes are aligned with data in an user interface. The user interface comprises at least a time scale to which the recognized phonemes are aligned to. The user interface may comprise various other information depending on a targeted use.

[0050] Post-Production Assistant

[0051] Now referring to FIG. 2, there is shown one specific embodiment of the invention. This specific embodiment of the invention is intended to assist post-production operations.

[0052] In this specific embodiment, the invention comprises a video source 10, a display 12, a conformed text source 14, a phoneme to grapheme unit 16, a phoneme recognition module 18, a word to phoneme database 20, a phoneme to grapheme database 21, a post-production sound recording synchguide 22, a video destination 24, a new sound source 26 and a project database 28.

[0053] The video source 10 provides a video stream to the display and to the post-production sound recording synchguide 22. The video source 10 also provides an audio source to the post-production sound editor 22 and to the phoneme recognition module 18. The video source 10 provides time codes to the conformed text source 14. In this embodiment of the invention, the time codes provided have a common time origin.

[0054] The conformed text source 14 is controlled by the post-production sound recording synchguide 22. The conformed text source 14 provides conformed text and time codes to the phoneme recognition module 18 and to the phoneme to grapheme unit 16. The phoneme recognition module 18 is a standard voice recognition module that provides phonemes as well as time codes. Someone skilled in the Art of voice recognition will have sufficient information to select an appropriate phoneme recognition module.

[0055] The phoneme to grapheme unit 16 receives the phonemes and the time codes from the phoneme recognition module 18. The phoneme recognition module 18 is controlled by the post-production sound recording synchguide 22. In the preferred embodiment of the invention, each phoneme provided by phoneme recognition module 18 has two time codes. One of the two time codes is dedicated to the beginning of the phoneme; the other of the two time codes is dedicated to the end of the phoneme.

[0056] The phoneme to grapheme unit 16 provides at least the graphemes with the time codes to the post-production sound recording synchguide 22. Two time codes are dedicated for each grapheme in the preferred embodiment of the invention. The first time code of the two time codes is dedicated to the beginning of the grapheme, while the second time code of the two time codes is dedicated to the end of the grapheme.

[0057] The post-production sound recording synchguide 22 receives the graphemes with the time codes from the phoneme to grapheme unit 16, a new audio source from the new audio source 26 and provides the results to the audio destination 24. The post-production sound recording synchguide 22 is connected to the project database 28.

[0058] The phoneme to grapheme unit **16** is connected to the word to phoneme database **20** and to the phoneme to grapheme database **21**.

[0059] The word to phoneme database **20** comprises a list of words with their corresponding phonemes.

[0060] The phoneme to grapheme database **21** comprises a list of words where the phonemes of each word are mapped to the corresponding graphemes.

[0061] The project database **28** comprises data about the project as explained below.

[0062] Now referring to **FIG. 3**, there is shown another embodiment of the invention. According to step **30** of **FIG. 3**, the project environment is set. The script is then prepared according to step **32** of **FIG. 3**. The synchguide is prepared according to step **34** of **FIG. 3**. Then, according to step **36**, the synchguide is modified. According to step **38**, information related to the project is generated.

[0063] Now referring to **FIG. 4**, there is shown how the project environment is setup. According to step **40**, the global parameters for the project are entered. The global parameters comprise, but are not limited to, the name of the project, the type of project, the identity as well as the access login and password of the persons allowed to work on the project, etc.

[0064] According to step **42**, the project parameters are entered. The project parameters comprise, and are not limited to, the name and the location of the video source, the name and the location of the audio source. In the preferred embodiment of the invention, the global parameters and the project parameters are provided to the post-production sound recording synchguide **22** and stored in the project database **28**.

[0065] According to step **44**, the project is planned. The step of planning the project comprises the step of assigning a time schedule to the persons allowed to work on the project.

[0066] Now referring to **FIG. 5**, there is shown how the script is prepared in the preferred embodiment of the invention. According to step **48**, the script is conformed. According to step **50**, the script is formatted. According to step **52**, a part of the formatted script is selected. The part of the formatted script is selected from the conformed text source **14** using the control of the post-production sound recording synchguide **22**.

[0067] Now referring to **FIG. 6**, there is shown how the synchguide is prepared.

[0068] According to step **58**, the script is provided to the phoneme recognition module **18**. According to step **60**, phonemes are generated by the phoneme recognition module **18** using at least the provided script and time codes. According to step **62**, graphemes are generated using the phoneme to grapheme unit **16**, the word to phoneme database **20** and the phoneme to grapheme database **21**. In the preferred embodiment, graphemes are generated with their related time codes. More precisely, the phoneme to grapheme unit **16** receives a word from the conformed text source **14**; the phonemes of the words provided by the conformed text source **14** are then found using the word to phoneme database **21**. The phoneme to grapheme unit **16** also receives the phonemes as well as time codes from the phoneme recognition module **18**. A match is then performed between the phonemes provided by the phoneme recognition module **18** and the phoneme found using the word to phoneme database **21**. The phoneme to grapheme unit then provides then the graphemes using the phoneme to grapheme database **21**, together with the word and the matched phonemes.

[0069] According to step **63**, the graphemes are placed on the synchguide. In this embodiment of the invention, the graphemes are placed on the synchguide using the post-production sound recording synchguide **22**. The synchguide is displayed using the display **12**.

[0070] According to step **64**, a check is performed on the synchguide in order to confirm that the original synchguide is correct. If the synchguide is not correct, i.e. for instance small grammatical errors are detected; the text may be amended according to step **66**. If the synchguide is correct and according to step **68**, labials as well as other miscellaneous information is provided.

[0071] Now referring to **FIG. 7**, there is shown how the synchguide may be modified.

[0072] According to step **70**, the user may provide a new text. The new text is provided to the conformed text source **14**. According to step **72**, a new sound source may be provided using the new sound source **26**. According to step **74**, the new sound source is aligned with the new text. This step is performed by generating the phonemes related to the new text source and their related time codes and then performing the phoneme to grapheme conversion using the phoneme to grapheme unit **16** together with the word to phoneme database **20** and the phoneme to grapheme database **21**. Using the time codes generated by the phoneme to grapheme unit **16**, the new sound source is aligned with the new text.

[0073] According to step **76**, at least one part of the new synchguide is then aligned with the old synchguide. The alignment is performed in the preferred embodiment of the invention using the time codes.

[0074] In another embodiment of the invention, the new synchguide is saved in the project database **28**.

[0075] According to step **78** of **FIG. 7**, the new synchguide is provided to the user.

[0076] In another embodiment of the invention, the persons allowed to work on the project may work via a remote location. The post-production sound recording synchguide **22** may be connected in this embodiment of the invention to a post-production sound recording synchguide server. Each allowed person may then access the post-production sound recording synchguide server remotely through a Local Area Network (LAN) or through a Wide Area Network (WAN).

[0077] Now referring to **FIG. 8**, there is shown a screen shot of the user interface in this embodiment of the invention.

[0078] In this embodiment of the invention, the user interface comprises a menu, a guide track **90**, a symbol menu **94**, a loop/preview box **96**, a zoom window **99**, a navigation window **100** and a script window **102**.

[0079] The guide track **90** enables the user to visualize the universal guide track. The universal guide track comprises a list of all the current actors on the scene as well as all the sound effects that are not performed by an actor.

[**0080**] In one embodiment of the invention, identity of the actors is detected using the database of the project **28**. For each actor the corresponding dialog is provided. The dialog is synchronized with time and displayed in a manner that allows an easy post-synchronization. In the preferred embodiment, the dialog is synchronized with time using the time codes provided with the graphemes.

[**0081**] The graphemes are placed with a letter length that corresponds to the phonemes length in order to provide an easy post-synchronization. The graphemes may be placed with a letter length that corresponds to the phonemes length using a time dependant character set or using for instance a technology such as the "Truetype" technology. In another embodiment, different colors may be used to show the temporal properties of the graphemes.

[**0082**] Preferably, elements such as breaths, efforts, presence and exclamations are placed on the universal guide track using special expository symbols. In another embodiment, dedicated colors are used to present a special effect. In this embodiment, Foleys are placed on a Foley track.

[**0083**] A moveable synchronizer bar enables the allowed users to view with precision the current position. A user may use the moveable synchronizer bar to move to a precise position.

[**0084**] The navigation window **100** enables a user to navigate through the different parts of the project. The navigation window **100** comprises a display that allows a user to find out his relative position. The navigation window **100** also comprises a display that allows a user to change the current scene. The navigation window **100** also provides a zoom in/out tool. The navigation window **100** also provides a tool that enables speed control and an indication of the frame reference.

[**0085**] The script window **102** enables a user to have access to the conformed text. The text currently spoken on the screen is highlighted. A user may edit the text in the conformed text window.

[**0086**] The zoom window **99**, allows a user to view the lip movements with a greater precision.

[**0087**] In a first alternative embodiment, the invention does not comprise the conformed text source **14**. In this embodiment, the phoneme recognition module **18** may provide the phonemes with a great efficiency, as the conformed text source is not available.

[**0088**] Now referring to **FIG. 9**, there is shown another alternative embodiment. In this embodiment, the system comprises a video source **10**, a display **12**, a conformed text source **14**, a phoneme recognition module **18**, a post-production sound recording synchguide **22**, an audio destination **24**, a new sound source **26** and a project database **28**.

[**0089**] The video source **10** provides an audio source to the phoneme recognition module **18** and to the post-production sound recording synchguide **22**. The video source **10** further provides time codes to the conformed text source, to the phoneme recognition module **18** and to the post-production sound recording synchguide **22**. The video source **10** provides the video source to the display **12** and to the post-production sound recording synchguide **22**.

[**0090**] The conformed text source **14** provides the conformed text to the phoneme recognition module **18**. In this embodiment, the phoneme recognition module **18** provides the phonemes with the related time codes to the post-production sound recording synchguide **22**. The phoneme recognition module **18** and the conformed text source **14** are controlled by the post-production sound recording synchguide **22**. The phoneme recognition module **18** is of the same type than the one described in the first embodiment of the invention.

[**0091**] In this embodiment, the post-production sound recording synchguide **22** provides the phonemes with their related time codes on the synchguide which is displayed by the display. More precisely, the post-production sound recording synchguide **22** provides a user interface where the phonemes are placed together with an indication of the current temporal location. The user has therefore an indication of when a sound begins and when a sound ends. It will be appreciated that this embodiment is simpler than the first embodiment but it greatly improves the prior art of "beep and wipe" systems. An alternative to this embodiment is to not include the conformed text source **14**

[**0092**] Closed-Captioning Assistant

[**0093**] Closed-captioning may be divided in two different types.

[**0094**] A first type of closed-captioning, also named "offline-captioning", is when there is sufficient lead time between completion of a program, series or film and its transmission. In such a case, a file of caption can be prepared. Each caption will be assigned its own unique time code cue which references back to the original master tape.

[**0095**] Someone skilled in the art will appreciate that it takes an experienced steno-captioner about 9 hours to close-caption a 22 minute, half an hour program. It takes the same experienced person, depending on the complexity and degree of difficulty, about 5 times that to do a 100-minute feature film or 45 hours. Someone skilled in the art will appreciate that inserting the time codes alone can take up to 20% of the total time to prepare the file of caption.

[**0096**] A second type of closed-captioning, also named "online-captioning", is when there is not sufficient lead time between completion of a program, series or film and its transmission. This is the case for news, current affairs programs or live broadcasts. In such cases, programs must be captioned online by a live real time steno-captioner. Unfortunately, someone skilled in the art will appreciate that there is never sufficient time to insert proper time codes, and often source materials do not contain the continuous time codes on it to trigger these cues. Consequently, if the steno-captioners are able to prepare their caption scripts just prior to transmission, it is still necessary for one of the steno-captioners to manually cue out each caption one by one. It will be appreciated that this is an inefficient use of highly skilled and valuable personnel.

[**0097**] According to the invention, and as explained below, it is possible to automatically cue online air captioning device and again free up person or persons to prepare following broadcast material.

[0098] Now referring to **FIG. 10**, there is shown an embodiment of the invention for assisting closed-captioning.

[0099] In this embodiment, an audio video source **200** provides an audio signal to a speech recognition module **202**. The speech recognition module **202** provides recognized words and related time codes to a closed-caption editor **204**. The closed-caption editor **204** further receives a video signal from the audio video source **200**.

[0100] Now referring to **FIG. 11**, there is shown how the embodiment described in **FIG. 10** operates.

[0101] According to step **206**, a closed-captioning operator sets its preferences. The preferences comprise at least user preferences related to a user interface. The preferences may comprise additional preferences.

[0102] According to step **208**, an audio video source signal is provided by the audio video source **200**.

[0103] According to step **210**, speech recognition of the audio video source signal is performed by the voice recognition module **202**. The speech recognition module **202** outputs recognized words and related time codes and provides the recognized words and the related time codes to the closed-caption editor **204**.

[0104] According to step **212**, results from the speech recognition are analyzed according to user preferences provided at step **206**. For instance, if a rate of recognized word reaches a predetermined level, an indication to that effect is provided to a closed-captioning operator.

[0105] According to step **214**, the recognized words and the related time codes are provided to the closed-captioning operator. It will be appreciated by someone skilled in the art that the invention reduces the time required to perform the closed-caption generation by at least automating the matching of time code cues to captions which have been prepared by an operator transcribing the audio material. Furthermore, it will be appreciated that the time required to perform the closed-caption generation can be reduced even further by performing the automatic voice recognition. According to the invention, a 100-minute film could be transcribed and time coded, automatically within 15 to 20 minutes, leaving the closed-captioning operator to verify manually what had previously been accomplished automatically.

[0106] It will be therefore appreciated that the saving of time generated by this advantageous embodiment will be easily translated into the time banking of valuable personnel.

[0107] In the preferred embodiment of the invention, the speech recognition performed according to step **210** is performed on a married track that contains all the dialogues, the music, effects and background or ambient noises.

[0108] In this embodiment, the closed-captioning operator is able to quickly visualize the recognized words on the user interface. The closed-captioning operator is then able to amend the recognized words to correct a defect or to insert a symbol related to an event. In the case of a large rate of recognized word, the closed-captioning operator may then simply decide to rewrite new closed-captions which comply with a suitable rate.

[0109] As the recognized words are provided with related time codes, the closed-captioning operator is able to control precisely a location of an element to insert or to amend.

[0110] It will be appreciated that closed-captioning remains synchronized with time as recognized words are provided with related time code.

[0111] The closed-captioning operator may then, in one embodiment, confirm an amendment or a recognized word by pressing a predetermined key. Upon confirmation, related data is then inserted into a Vertical Blanking Interval as known by someone skilled in the art. In an alternative embodiment, an automatic cue is performed.

[0112] In a simpler embodiment, no feedback is provided by the closed-captioning operator. The recognized words are directly inserted with related time codes in the Vertical Blanking Interval.

[0113] In another embodiment, one word at a time can be confirmed by the closed-captioning operator. In another embodiment, more than one word at a time can be confirmed.

[0114] Alternatively, the closed-captioning operator may provide words to the closed-caption editor **204**. The words provided may be provided via steno data that are then translated into words according to the art. In such a case, an open-captioning operator may also provide abbreviations instead of words to the closed-caption editor **204**. The closed-caption editor **204** may in return translate the provided abbreviation into a corresponding word, enabling a saving of time. In such embodiment, the closed-caption editor **204** further comprises a look-up database and a steno data to word translation unit. The look-up database comprises a relation between an abbreviation and a word.

[0115] In an alternative embodiment, the speech recognition module **202** comprises a word database. The word database may be amended by the user. Such word database enables a user to introduce new words or specify a correct orthography of a word. Such word database is therefore of great advantage for close captioning.

[0116] The speech recognition module **202** may also comprise an orthography module which highlights words that are poorly detected or for which orthography does not seem correct.

[0117] Animation Assistant

[0118] Now referring to **FIG. 12**, there is shown an embodiment of the invention for assisting animation creation.

[0119] In this embodiment, a voice source **216** provides a voice source signal to a speech recognition module **220**. A script source **218** provides a script signal to the speech recognition module **220**. The voice source signal provided by the voice source **216** is generated by an actor according to a script read. The voice source signal may be provided to the speech recognition module **220** in accordance with various data formats.

[0120] The speech recognition module **220** provides recognized words and time codes to a computerized animation assistant **224**. A visem database **222** provides a visem signal to the computerized animation assistant **224**. A story board database **228** provides a story board signal to the computerized animation assistant **224**.

[0121] An adjusted voice track signal is provided by the computerized animation assistant **224** to an adjusted voice track database **230**.

[0122] Now referring to **FIG. 13**, there is shown how this embodiment operates.

[0123] According to step **232**, an animation assistant provides its preferences. The preferences comprise information related to a user interface of the computerized animation assistant **224**.

[0124] According to step **234**, a voice source **216** is provided. An actor provides a recording according to a script.

[0125] In one embodiment the voice source signal comprises a plurality of recordings originating from various actors.

[0126] According to step **235**, speech recognition is performed using at least the voice source signal provided by the voice source **216** and using the speech recognition module **220**. Recognized words and related time codes are provided to the computerized animation assistant **224**. In the preferred embodiment, the phonemes and the related time codes are also provided to the computerized animation assistant **224**.

[0127] According to step **236**, visems are provided to the computerized animation assistant **224** by the visem database **222** in response to a request performed by the computerized animation assistant **224**. The request comprises at least the phonemes provided by the speech recognition module **220**.

[0128] According to step **238**, the story board database **228** provides a story board signal to the computerized animation assistant **224** in response to a story board request. The story board request comprises at least the recognized words provided by the speech recognition module **220** to the computerized animation assistant **224**.

[0129] The story board signal provided relates to at least one part of the story board related to the recognized words provided.

[0130] According to step **240**, animation information is provided to a user interface using the computerized animation assistant **224**. The animation information provided to the user interface is intended to enable an easier and quicker creation of animation.

[0131] More precisely, the animation information comprises a sequence of recognized words with related part of the storyboard; the animation information further comprises related visem for each of the recognized words in a frame in order to facilitate the work of the draftsman. It will be appreciated by someone skilled in the art that such a tool enables the draftsman to precisely locate where an animation drawing must be made and further provides the draftsman with a clear indication of how a drawing should be made according to the visems provided. In this embodiment, the computerized animation assistant provides an adjusted voice track signal to the adjusted voice track database **230**.

[0132] Adaptation Assistant

[0133] Now referring to **FIG. 14**, there is shown another embodiment of the invention. In this embodiment an adaptation of an audiovisual work is performed.

[0134] A voice source **282** provides a voice source signal to a speech recognition module **284**. An adapted voice source **280** provides an adapted voice source signal to the speech recognition module **284**. The adapted voice source **280** is preferably provided by a user adapting the voice source.

[0135] The speech recognition module **284** performs a speech recognition of the voice source signal and of the adapted voice source signal.

[0136] The speech recognition module **284** provides recognized voice source signal with voice source related time codes to a recognized data analysis unit **286**. The speech recognition module **284** further provides recognized adapted voice source signal with adapted voice source related time codes to the recognized data analysis unit **286**.

[0137] The recognized data analysis unit **286** performs an analysis of the received recognized adapted voice source signal with the adapted voice source related time codes and the recognized voice source signal with the voice source related time codes. The recognized data analysis unit **286** provides an analysis result signal to a recognized data matching unit **288**.

[0138] Preferably, the analysis result provides an indication on whether it is possible to match the adapted voice source signal and the voice source signal using the voice source related time codes and the adapted voice source related time codes.

[0139] More precisely, the recognized data analysis unit **286** operates by trying to match the adapted voice source signal and the voice source signal using phonemes. In another embodiment, the match is performed using visems.

[0140] The recognized data matching unit **288** receives a user defined criteria signal. The user defined criteria signal provides an indication of a level of adaptation synchronization quality required by a user.

[0141] The recognized data matching unit **288** provides an adapted voice source signal.

[0142] Now referring to **FIG. 15**, there is shown a flow chart which shows how an adaptation is performed.

[0143] According to step **250**, a user sets its preferences. The setting of the preferences comprises a providing of a user defined criteria signal to the recognized data matching unit **288**. The user defined criteria signal is indicative of a level of adaptation synchronization required by the user.

[0144] According to step **252**, the voice source signal, originating from the voice source **282**, is provided to the speech recognition module **284**.

[0145] According to step **254**, a speech recognition of the voice source signal is performed by the speech recognition module **284**.

[0146] According to step **256**, an adapted voice source is provided by an operator. As explained previously, and preferably, the adapted voice source is created by adapting the voice source signal provided by the voice source **282**. The speech recognition module outputs the recognized voice source signal and the related recognized voice source time codes.

[0147] According to step **258**, a speech recognition of the adapted voice source signal is performed by the speech recognition module **284**. The speech recognition module outputs the recognized adapted voice source signal and the related recognized adapted voice source time codes.

[**0148**] According to step **260**, an attempt is made to match the recognized adapted voice source signal and the recognized voice source signal. The attempt is made by using the related recognized voice source time codes and the related recognized adapted voice source time codes. The user defined criteria signal is also used to assist adaptation synchronization.

[**0149**] According to step **262**, an indication of confidence is provided to the user. The indication of confidence provides an indication of an amount of time codes matched between the related recognized adapted voice source time codes and the related recognized voice source time codes.

[**0150**] According to step **264**, a recording of the result of the adaptation is performed.

[**0151**] In one implementation of this embodiment, a user may provide a minimum amount of time codes to be matched. In such implementation, the recording of the result of the adaptation may be cancelled if the minimum amount of time codes to be matched is not met.

[**0152**] It will be appreciated that in this embodiment, the invention may be advantageously used for assisting adaptation by attempting to match time codes.

[**0153**] Karaoke

[**0154**] Karaoke is a form of entertainment that originated in Japan twenty years ago and which means "empty orchestra". It is an abbreviation of Karappo Okesutura—Kara translates to empty and Oke translates to orchestra.

[**0155**] Methods currently in use today will either underline the words as they come up musically, in different colors sometimes, or they will uncover the lyrics as they pass in time to the music.

[**0156**] Now referring to **FIG. 16**, there is shown an embodiment of the invention for assisting Karaoke generation.

[**0157**] An audio/video source **300** provides an audio signal to a speech recognition module **310**. The speech recognition module **310** performs a speech recognition of the audio signal provided and generates recognized voice source signal with related time codes. The recognized voice source signal with related time codes are then received by the recognized voice source formatting unit **312**. The voice source formatting unit **312** also receives a video source signal from the audio/video source **300** and a music source signal. The recognized voice source formatting unit **312** generates a combined video signal comprising formatted text, the music signal and at least one part of the video source signal provided and provides the combined video signal to a display **314**. More precisely, the recognized voice source formatting unit **312** provides a formatted text synchronized with the video source signal and with the music using the recognized voice source signal and the related time codes. A marker is used to locate exactly a current temporal location on the formatted text with respect to music played.

[**0158**] In an alternative embodiment of the invention, the speech recognition module **310** provides the music without lyrics and respective time codes.

[**0159**] Preferably, the formatted text is then displayed on the display **314**, in a precise manner, using a time dependant character set and an horizontal font or it can be combined with computer generated animation. Alternatively, computer generated animation may be used to enhance the display, in order to have an entertaining display as well.

[**0160**] Now referring to **FIG. 17**, there is shown how assisting Karaoke is performed.

[**0161**] According to step **320**, a user sets his preferences.

[**0162**] According to step **322**, an audio/video source **300** is provided. The audio/video source **300** may be provided using a plurality of medium. The audio/video source **300** comprises an audio source signal and a video source signal.

[**0163**] According to step **324**, a speech recognition of the audio source signal is performed by the speech recognition module **310**.

[**0164**] According to step **326**, recognized speech and time codes, originating from the speech recognition module **310**, are used to generate the combined video signal. As explained above, the combined video signal comprises formatted text, the music signal and at least one part of the video source signal provided.

[**0165**] According to step **328**, formatted text and at least one part of the video signal is displayed on the display **314**. Music without the lyrics is also provided.

[**0166**] Someone skilled in the art will therefore appreciate that this embodiment is of great advantage as it provides a synchronized formatted text with respect to music.

[**0167**] Musical Guide Track

[**0168**] As someone skilled in the art will appreciate, it is of great advantage to be able to know the exact location of music in an audiovisual work.

[**0169**] Thus, in one embodiment of the invention, a composer/conductor may wish to insert notes or any indications that may be required to further create or amend music. The insertion of notes or any indications are performed according to a specific insertion scheme and are further detected in an audiovisual work.

[**0170**] Upon detection according to the insertion scheme, the notes and the indications are provided together with related time codes to a display. The notes and the indications, provided together with related time codes, are then used to further amend music in the audiovisual work.

[**0171**] It will therefore be appreciated that such embodiment, allows a total focusing on what is on the screen and enables the composer/conductor to incorporate more musical passages and visual images in the session so as to provide a greater fluidity of the music being recorded. This can be further appreciated by a reduction in studio recording time and music editing time.

**1**. A method for producing an audiovisual work, the method comprising the steps of:

providing an audio signal to a speech recognition module;

performing a speech recognition of said audio signal, the speech recognition comprising an extracting of a plurality of basic units of recognized speech and related time codes;

receiving the plurality of basic units of recognized speech and the related time codes from the speech recognition module;

processing the received plurality of basic units to provide synchronization information for a production of said audiovisual work; and

displaying on a user interface said synchronization information.

**2**. The method as claimed in claim 1, wherein the production comprises post-production audio synchronization, said synchronization information comprises a graphic representation of a sound to be performed at each point in time over a span of time during said audiovisual work, and said interface controls said graphic representation over said span while facilitating synchronized recording of said sound in order to perform post-production.

**3**. The method as claimed in claim 2, wherein the basic units of recognized speech are phonemes.

**4**. The method as claimed in claim 2, further comprising the step of converting the basic units of recognized speech received with the time codes into words and words related time codes.

**5**. The method as claimed in claim 2, further comprising the step of converting the basic units of recognized speech received with the time codes into graphemes and graphemes related time codes, the graphemes being processed to provide synchronization information.

**6**. The method as claimed in claim 5, further comprising the step of providing a conformed text source, further wherein the synchronization information provided to the user comprises an indication of a temporal location with respect to the audio signal.

**7**. The method as claimed in claim 5, further comprising the step of providing a script of at least one part of the audio signal, further wherein the synchronization information provided to the user comprises an indication of a temporal location with respect to the script provided.

**8**. The method as claimed in claim 5, wherein the displaying on a user interface of said synchronization information, comprises the displaying of the graphemes using a horizontally sizeable font.

**9**. The method as claimed in claim 5, further comprising the step of detecting a Foley in the audio signal using a Foley detection unit, the detecting comprising the providing of an indication of the Foley and a related Foley time code.

**10**. The method as claimed in claim 5, further comprising the step of amending at least one part of the audio signal and audio signal related time codes using at least the graphemes and the synchronization information.

**11**. The method as claimed in claim 4, further comprising the providing of a plurality of words in accordance with the provided audio signal, the providing being performed by an operator.

**12**. The method as claimed in claim 11, further comprising the step of amending a recognized word in accordance with the plurality of words provided by the operator.

**13**. The method as claimed in claim 12, further comprising the step of creating a composite signal comprising at least the amended word, a video signal related to the audio source and the audio source.

**14**. The method as claimed in claim 1, wherein the displaying on a user interface of said synchronization information is used to produce animation.

**15**. The method as claimed in claim 14, wherein for blocks of continuous spoken word, said synchronization information provides essential visem information for each sequential frame to be drawn by an animator.

**16**. The method as claimed in claim 15, further comprising the step of providing a storyboard database, further comprising the step of converting the basic units of recognized speech received with the time codes into words and words related time codes, the processing of the plurality of words and the words related time codes providing an indication of a current temporal location of the audio signal with respect to the storyboard.

**17**. The method as claimed in claim 16, wherein the basic units of recognized speech are phonemes, further comprising the step of providing a plurality of visems for each of the plurality of words, using a visem database and using the phonemes.

**18**. The method as claimed in claim 17, further comprising the step of outputting an adjusted voice track comprising the audio signal, at least one part of the storyboard and the plurality of visems.

**19**. The method as claimed in claim 1, wherein the production comprises adaptation assisting, the adaptation assisting comprises a graphic representation of the plurality of basic units of recognized speech, the related time codes and a plurality of adapted basic units provided by a user, and said interface providing a visual indication of a matching of the plurality of adapted basic units with the plurality of basic speech units, the matching enabling synchronized adaptation of said audio signal.

**20**. The method as claimed in claim 19, wherein the plurality of adapted basic units is provided by performing a speech recognition of an adapted voice source.

**21**. The method as claimed in claim 20, wherein the speech recognition of the adapted voice source further provides related adapted time codes, further wherein the step of adapting the audio signal using said synchronization information and the plurality of adapted basic units is performed by attempting to match at least one of the plurality of basic units with at least one of the plurality of adapted basic units using the related time codes and the related adapted time codes.

**22**. A method for performing closed-captioning of an audio source, the method comprising the steps of:

providing an audio signal of an audio/video signal to a speech recognition module;

performing a speech recognition of said audio/video signal, and

incorporating text of said recognized speech of the audio signal as closed-captioning into a visual or non-visual portion of the audio/video signal in synchronization.

**23**. The method as claimed in claim 21 further comprising the step of providing an indication of an amount of successful replacement of the plurality of basic units of recognized speech of the audio signal by the plurality of basic units of recognized speech of the adapted audio signal.

**24**. The method as claimed in claim 23, further comprising the step of providing a minimum amount required of successful replacement of the plurality of basic units of recognized speech of the audio signal by the plurality of basic units of recognized speech of the adapted audio signal, the method further comprising the step of canceling the providing of the at least one replaced plurality of basic units with related replaced time codes if the at least one replaced plurality of basic units is lower than the minimum amount required of successful replacement.

**25**. The method as claimed in claim 1, wherein the audio signal comprises a plurality of voices originating from a plurality of actors, further comprising the step of assigning each of the plurality of basic units and the related time codes to a related actor of the plurality of actors.

**26**. The method as claimed in claim 1, wherein the production comprises closed-captioning production of the audio source, said closed-captioning comprises a graphic representation of the recognized plurality of basic units, the method further comprising the incorporating of at least one of the plurality of basic units as closed-captioning in a visual or non-visual portion of the audio/video portion of the audio/video signal in synchronization.

**27**. The method as claimed in claim 26, further comprising the step of amending at least one part of the plurality of basic units.

**28**. The method as claimed in claim 1, further comprising the step of converting the basic units of recognized speech received with the time codes into words and words related time codes, further comprising the step of creating a database comprising a word and related basic units.

**29**. The method as claimed in claim 28, further comprising the step of amending a word of said database, wherein phonemes of the word and the amended word are substantially the same.

**30**. The method as claimed in claim 1, further comprising the step of converting the basic units of recognized speech received with the time codes into words and words related time codes, further comprising the step of amending at least one word.

**31**. The method as claimed in claim 30, further comprises the step of providing a visual indication of a word to amend.

**32**. The method as claimed in claim 1, wherein the audio signal comprises lyrics that are sung, further wherein the production of said audiovisual work comprises a karaoke generation using said audio signal, said karaoke generation comprises a graphic representation of lyrics to be sung at each point in time over a span of time during said audiovisual work using the plurality of basic units of recognized speech provided and related time codes, together with an index representation of a current temporal position with respect to the graphic representation of the lyrics to be sung.

**33**. The method as claimed in claim 2, further comprising the step of detecting at least one note encoded in the audio signal according to an encoding scheme, further comprising the providing of the detected at least one note on said graphic representation.

* * * * *