US 20060253272A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2006/0253272 A1

Gao et al. (43) Pub. Date: Nov. 9, 2006

(54) VOICE PROMPTS FOR USE IN SPEECH-TO-SPEECH TRANSLATION SYSTEM

(75) Inventors: **Yuqing Gao**, Mount Kisco, NY (US); **Liang Gu**, Mohegan Lake, NY (US); **Fu-Hua Liu**, Scarsdale, NY (US)

Correspondence Address:
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560 (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(21) Appl. No.: **11/123,287**

(22) Filed: **May 6, 2005**

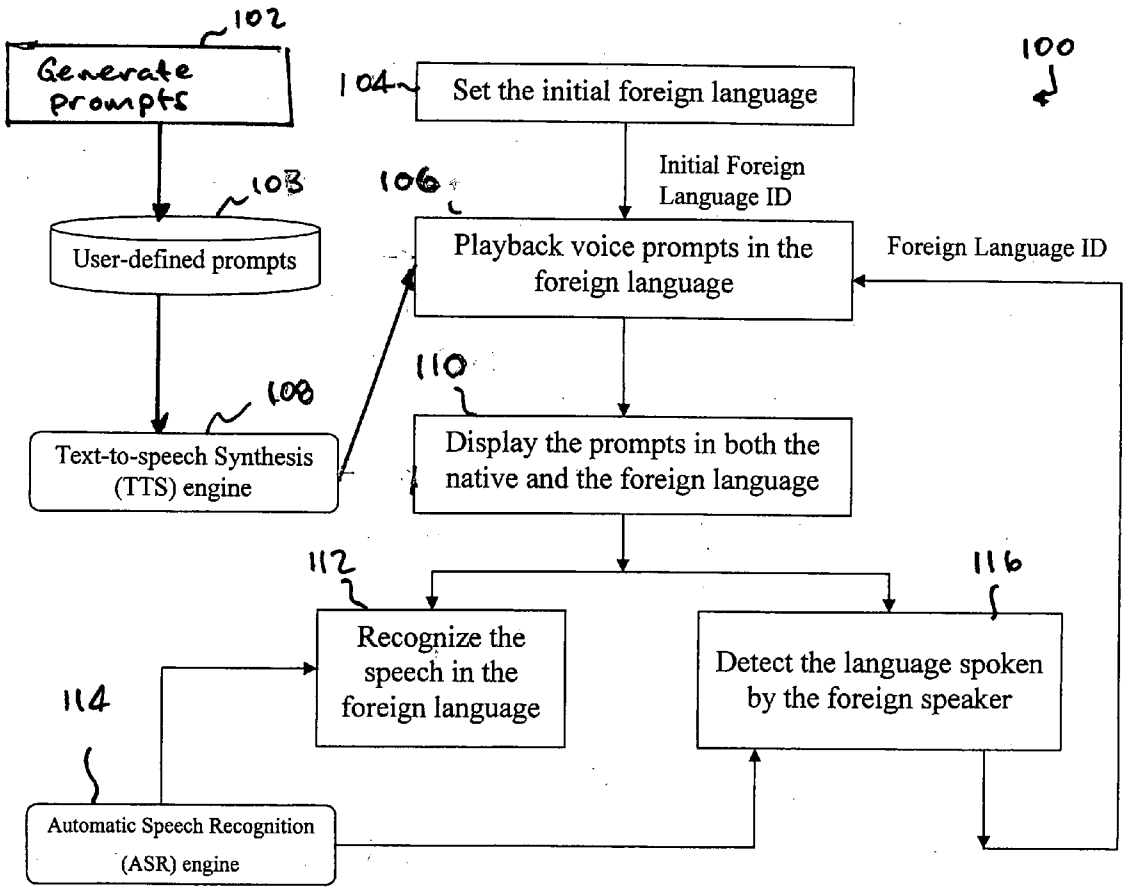
Publication Classification

(51) **Int. Cl.**
G06F 17/28 (2006.01)

(52) **U.S. Cl.** **704/2**

(57) **ABSTRACT**

Techniques for employing improved prompts in a speech-to-speech translation system are disclosed. By way of example, a technique for use in indicating a dialogue turn in an automated speech-to-speech translation system comprises the following steps/operations. One or more text-based scripts are obtained. The one or more text-based scripts are synthesizable into one or more voice prompts. At least one of the one or more voice prompts is synthesized for playback from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.



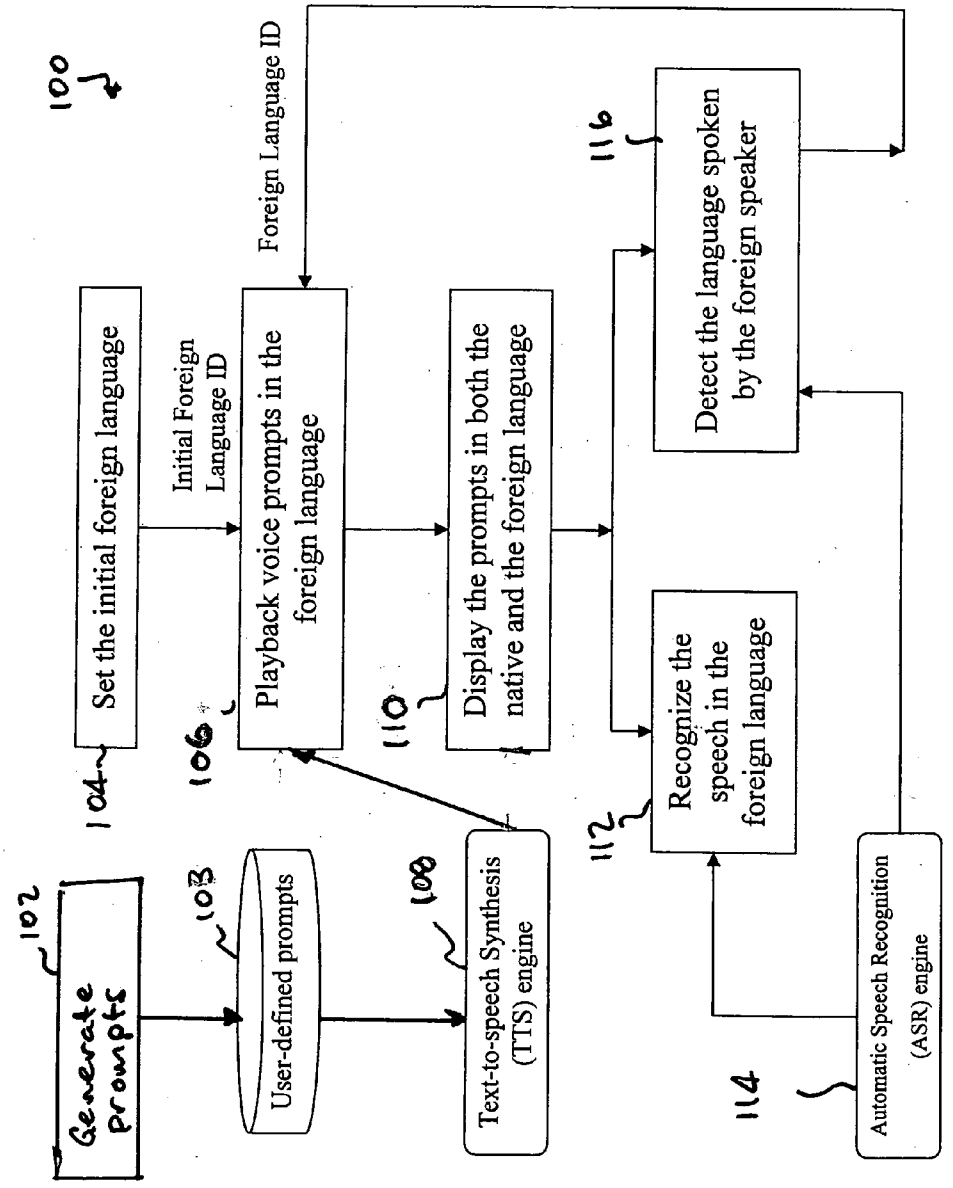

100
Generate Prompts
102
User-defined prompts
103
Text-to-speech Synthesis (TTS) engine
108
Automatic Speech Recognition (ASR) engine
114
Set the initial foreign language
104
Initial Foreign Language ID
Playback voice prompts in the foreign language
106
Foreign Language ID
Display the prompts in both the native and the foreign language
110
Recognize the speech in the foreign language
112
Detect the language spoken by the foreign speaker
116

FIG. 1

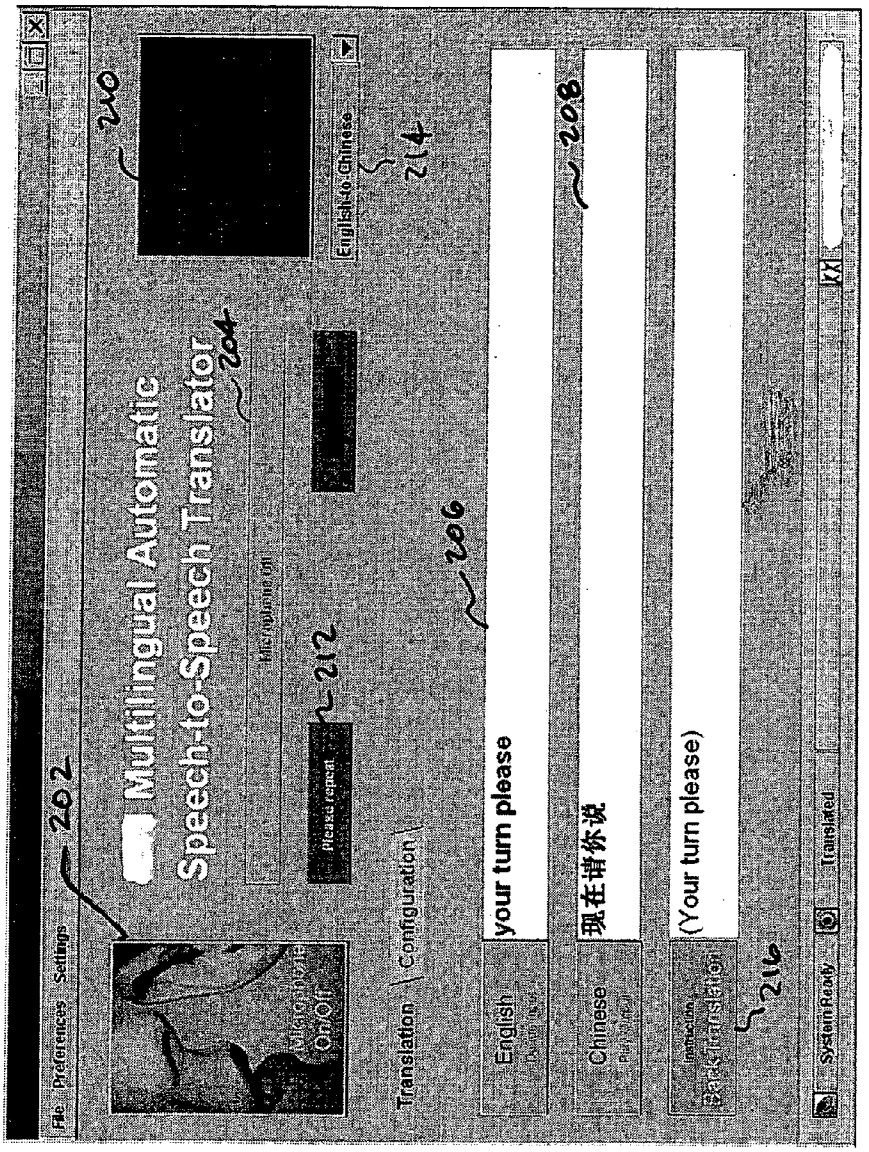

File Preferences Settings
Multilingual Automatic
Speech-to-Speech Translator
Translation
Configuration
English-to-Chinese
Please repeat
your turn please
现在请你说
(Your turn please)
English
Chinese
Translate
System Ready
Translated
202
204
206
208
210
212
214
216
220

FIG. 2

200

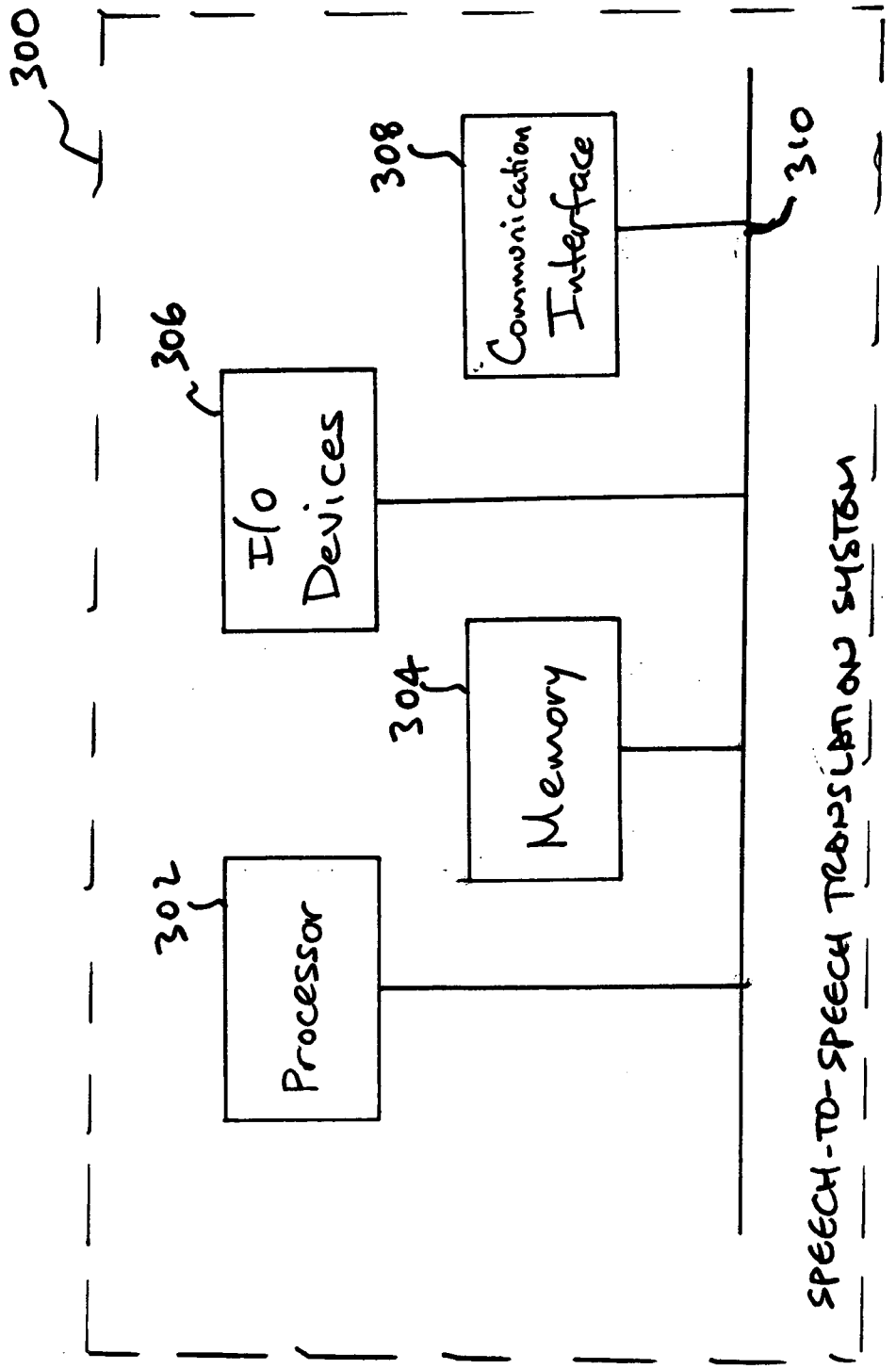

300
308
Communication Interface
310
306
I/O Devices
304
Memory
302
Processor
SPEECH-TO-SPEECH TRANSLATION SYSTEM

FIG. 3

VOICE PROMPTS FOR USE IN SPEECH-TO-SPEECH TRANSLATION SYSTEM

[0001] This invention was made with Government support under Contract No.: N66001-99-2-8916 awarded by DARPA BABYLON. The Government has certain rights in this invention.

FIELD OF THE INVENTION

[0002] This present invention generally relates to speech processing techniques and, more particularly, to techniques for employing voice prompts in a speech-to-speech translation system.

BACKGROUND OF THE INVENTION

[0003] Multilingual speech-to-speech language translation systems have been developed to facilitate communication between people that do not share a common language. One example of such a system is the speech-to-speech translation system developed by Carnegie Mellon University (Pittsburgh, Pa.).

[0004] A speech-to-speech translation system allows a user who has been trained with the system (hereinafter "system user") to communicate with another person who speaks another language (hereinafter "foreign language speaker" or just "foreign speaker") and is most often not familiar with the system, by providing speech-to-speech translation service between the two parties.

[0005] Since conventional speech-to-speech translation systems can handle only one speaker at a time, the two speakers need to take turns during the communication. Therefore, the indication (or prompt) of the switch of turns becomes a very important issue in order to ensure a smooth speech translation multilingual conversation.

[0006] Various prompts to indicate the switch of turns exist in conventional speech-to-speech translation systems. The most widely adopted prompt uses audio sound effects such as a beep sound. The sound effects can be language dependent so that a specific sound represents a specific language. The drawback of this approach is that both the system user and the foreign language speaker need to be trained to be familiar with the meaning of these sound effects. For a frequent system user, this brings additional inconvenience, as he or she must remember the meaning of sound effects for each language supported by the system. For a foreign speaker who is not familiar with or has never used this kind of system before, this function is not easily usable for them since the system user cannot explain the function to the foreign speaker because of the language barrier. The foreign speaker needs to guess the meanings of these sounds, often with great frustration and, consequently, with great dissatisfaction.

[0007] Another solution is to use visual prompts. The system user can point a microphone associated with the system to himself or herself when he or she starts to talk and point the microphone to the foreign speaker to indicate for the foreign speaker to start to talk. Other visual indications or gestures may be used to indict the switch of the turn. However, visual prompts are only helpful in face-to-face speech translation conversations and are useless for other scenarios such as automatic speech translation through call centers. Additionally, in some situations such as emergency medical care, patients speaking another language may keep their eyes closed due to their medical conditions so that the above-described visual prompts may be completely useless. Furthermore, these visual indications may still be confusing without verbal explanations.

SUMMARY OF THE INVENTION

[0008] Principles of the present invention provide techniques for employing improved prompts in a speech-to-speech translation system.

[0009] By way of example, in a first aspect of the invention, a technique for use in indicating a dialogue turn in an automated speech-to-speech translation system comprises the following steps/operations. One or more text-based scripts are obtained. The one or more text-based scripts are synthesizable into one or more voice prompts. At least one of the one or more voice prompts is synthesized for playback from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.

[0010] The technique may also comprise detecting a language spoken by a speaker interacting with the speech-to-speech translation system such that a voice prompt in the detected language is synthesized for playback to the speaker. An initial voice prompt may be synthesized for playback in a default language until the actual language of the speaker is detected.

[0011] The technique may also comprise one or more of displaying the at least one voice prompt synthesized for playback, recognizing speech uttered by the speaker interacting with the speech-to-speech translation system, and recognizing speech uttered by a system user of the speech-to-speech translation system. At least a portion of the speech uttered by the speaker or the system user may be translated from one language to another language. At least a portion of the translated speech may be displayed.

[0012] In a second aspect of the invention, a technique for providing an interface for use in an automated speech-to-speech translation system, the translation system being operated by a system user and interacted with by a speaker, comprises the following steps/operations. The system user enables a microphone of the translation system via the interface. At least one previously-generated voice prompt is output to the speaker, the at least one voice prompt comprising an audible message in a language understandable to the speaker, the audible message indicating a turn in a dialogue between the system user and the speaker. The speaker, once prompted, utters speech into the microphone, the uttered speech being translated by the translation system.

[0013] In a third aspect of the invention, an interface for use in an automated speech-to-speech translation system, the translation system being operated by a system user and interacted with by a speaker, comprises a first field for use by the system user to enable a microphone of the translation system, a second field for use by the system user for at least one of displaying speech uttered by the system user and displaying translated speech uttered by the speaker, and a third field for use by the speaker for at least one of displaying

speech uttered by the speaker and displaying translated speech uttered by the system user, wherein the translation system outputs at least one previously-generated voice prompt to the speaker, the at least one voice prompt comprising an audible message in a language understandable to the speaker, the audible message indicating a turn in a dialogue between the system user and the speaker, and the speaker, once prompted, uttering speech into the microphone, the uttered speech being translated by the translation system. The interface may comprise a fourth field for use by the system user to enable a microphone of the translation system such that speech uttered by the system user is captured by the translation system.

[0014] In a fourth aspect of the invention, an article of manufacture for use in indicating a dialogue turn in an automated speech-to-speech translation system, comprises a machine readable medium containing one or more programs which when executed implement the steps of obtaining one or more text-based scripts, the one or more text-based scripts being synthesizable into one or more voice prompts, and synthesizing for playback at least one of the one or more voice prompts from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.

[0015] Accordingly, principles of the invention provide a prompt solution for use in a speech-to-speech translation system that can sufficiently indicate both the switch of dialogue turns and the specific source language for the next turn.

[0016] These and other objects, features and advantages of the present invention will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] **FIG. 1** is a block/flow diagram illustrating a speech-to-speech translation system employing language detection-based multilingual voice prompts, according to an embodiment of the invention;

[0018] **FIG. 2** is a diagram illustrating a speech-to-speech translation system user interface, according to an embodiment of the invention; and

[0019] **FIG. 3** is a diagram illustrating a computing system in accordance with which one or more components/steps of a speech-to-speech translation system may be implemented, according to an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0020] As will be illustratively explained herein, principles of the invention introduce language-dependent voice prompts during machine-mediated automatic speech-to-speech translation.

[0021] It is to be understood that while principles of the invention are described as translating from one language to another language, the term "language" can also broadly include a dialect or derivation of a language. That is, language translation may also include translation from one dialect to another dialect.

[0022] It is also to be understood that the "system user" is one trained on (or at least operationally familiar with) the speech-to-speech translation system. The system user may also be considered a system operator. The "foreign language speaker" or just "foreign speaker" is one not familiar with or trained on the system. In one example application, the speech-to-speech translation system may be used to allow a customer service representative (i.e., system user) of some business to communicate with a customer (i.e., foreign language speaker), when the two individuals speak different languages.

[0023] A voice prompt solution is provided in accordance with principles of the invention that can verbally indicate the switch of the dialogue turns in the language of the foreign speaker by using an automatic language detection algorithm. Such a voice prompt solution is provided with a highly friendly user interface. The voice prompts comprise concise, natural and configurable voice instructions in the foreign language generated by text-to-speech synthesis (TTS) techniques. In a multilingual speech-to-speech translation system with more than two languages involved, the foreign language is determined based on the language detection result of the foreign speaker's speech during one or more previous turns, and with a default foreign language for the first dialogue turn. Therefore, no language selection is required and the system user only needs to click one button to activate the voice prompt for all the foreign language speakers. The user interface of the speech-to-speech translation system is hence very simple and highly convenient.

[0024] An illustrative embodiment of a methodology and system for implementing multilingual voice prompts in a speech-to-speech translation system will now be described.

[0025] Referring initially to **FIG. 1**, a block/flow diagram depicts a speech-to-speech translation system **100** employing language detection-based multilingual voice prompts, according to an embodiment of the invention.

[0026] As shown, in step **102**, voice prompts are generated in each desired foreign language and stored as respective script or text files (i.e., text-based scripts). A script file storage unit **103** may be used to store the generated prompts. For example, for a Chinese-to-English speech-to-speech translation, the voice prompt "your turn please" is generated in Chinese text as a script file and stored in script file storage unit **103**. Any number of voice prompts with various audible messages can be generated and stored in such a manner. Such voice prompts are easily generated and reconfigured since a system user can design preferred prompts by modifying existing prompt script files.

[0027] In step **104**, an initial or default foreign language is set. This initial language could be a foreign language prevalent in the geographic area of system use, e.g., Chinese or Spanish. The voice prompts in this default language are used at the beginning of the speech-translated system-mediated dialogues.

[0028] In step **106**, a voice prompt is generated (synthesized) via a text-to-speech synthesis (TTS) engine **108** from a prompt script file and audibly presented (played back) to

the foreign speaker via an output speaker associated with the system. The synthesized speech associated with the voice prompt may be generated from the text of the corresponding script file in a manner well known in the art of speech processing. Thus, the well-known automated TTS techniques are not described herein in detail, and the invention is not limited to any particular TTS technique. It is also to be understood that an initial foreign language identifier (ID) can be used to instruct the system as to which foreign language voice prompt to initially select for playback.

[0029] In step **110**, a text-form message of the played voice prompt is displayed on the system user interface in both the native language of the system and the foreign language as a visual feedback for the system user and the foreign language speaker. An illustrative system user interface will be described below in the context of **FIG. 2**.

[0030] Once prompted (e.g., "your turn please") that it is his or her turn, the foreign speaker will then speak into a microphone of the system. During each turn of the foreign speaker, the speech is recognized (step **112**) via an automatic speech recognition (ASR) system **114**. Based on the actual speech and/or the recognized speech, in step **116**, a language identification algorithm detects the language the foreign speaker is speaking in. It is to be understood that speech may be automatically recognized and the foreign language detected in manners well known in the art of speech processing. Thus, the well-known automated ASR techniques and automated language detection techniques are not described herein in detail, and the invention is not limited to any particular ASR or language detection techniques.

[0031] As also shown in step **116**, the language detection algorithm used generates an identifier that identifies the language detected by the algorithm. This is provided back to step **106** and replaces the default language identifier. Accordingly, before the dialogue turn switches back again to the foreign speaker, a voice prompt is played to the foreign speaker using the foreign language detected in the previous dialogue turn.

[0032] Referring now to **FIG. 2**, an illustrative speech-to-speech translation system user interface **200**, according to an embodiment of the invention, is shown. It is to be understood that control of the various buttons (displayed icons on the screen associated with the system) is exercised by the system user, i.e., the person trained to use the system. It is also to be understood that the various buttons, bars and textboxes described below are predefined functional fields within the screen area of the system user interface. Also, any TTS or ASR operations described below are respectively performed by TTS engine **108** and ASR engine **114** described above in the context of **FIG. 1**.

[0033] The system user presses (clicks on) button **202** of the system interface (also referred to as the graphical user interface or GUI) to turn on the system microphone. Voice volume bar **204** appears in the upper-middle part of the GUI page.

[0034] An audio prompt (such as a "beep") is played to indicate that the microphone is now on. The system user speaks native language (e.g., English) into the microphone. The recognized speech (recognized via ASR engine **114**) is shown in the first textbox **206**.

[0035] After the user finishes his/her speech and all the speech has been recognized, button **202** is pressed again to turn off the microphone. Voice volume bar **204** indicates that the microphone is off. The recognized message is then translated into the foreign language (e.g., Chinese) using a foreign language translation engine (not shown) and displayed in second textbox **208**. It is to be understood that the recognized message may be automatically translated into another language in a manner well known in the art of language translation. Thus, the well-known automated translation techniques are not described herein in detail, and the invention is not limited to any particular language translation techniques. The translated sentence is further played back to the foreign language speaker using TTS techniques (TTS engine **108**).

[0036] The system user presses button **210** to turn on the microphone (which may be the same microphone used by the system user or a different microphone) and let the foreign speaker speak. A language-dependent voice prompt is played to indicate (in the foreign language speech) that the microphone is now on and ready for speech-to-speech translation. Such a voice prompt may be generated and presented as explained above in the context of **FIG. 1**. The language of the voice prompt is determined based on the language detection algorithm, as also described above.

[0037] In one embodiment, after the language-ID-based voice prompt is played, an audio prompt (such as a beep sound) may also be played to further notify the foreign speaker that the microphone is on and he or she can start to talk. In other words, the voice prompt solution of the invention can be combined with the conventional audio prompt solution to achieve even higher user satisfaction.

[0038] The foreign speaker then speaks into the microphone. His or her speech is recognized and displayed in textbox **208**. After the foreign speaker finishes his or her speech and all the speech has been recognized, button **210** is pressed again to turn off the microphone. The recognized message is then translated back into the native language (e.g., English) and displayed in textbox **206**. The translated sentence is further played back in the native language speech using TTS techniques.

[0039] The above steps are considered as one turn of the speech-translation system-mediated dialogue. The native language user (system user) and the foreign language speaker will repeat these steps to communicate with each other until all information has been successfully exchanged.

[0040] Also shown in system interface **200** is a button **214** which serves as a short-cut button to playback a voice prompt that says "please repeat" in the detected foreign language. This may be used if the system user or the system itself does not understand what the foreign language speaker has said. Also, a pull-down menu **214** enables the system user to manually select the languages to be used in translation operations (e.g., English-to-Chinese, as shown). Further, button **216** functions as an "instruction" button. When pressed, an instructional voice message is played in the detected foreign language to enable the foreign speaker to get familiar with the system functions and therefore enable a smooth system-mediated speech-to-speech translation.

[0041] Referring finally to **FIG. 3**, a computing system in accordance with which one or more components/steps of a speech-to-speech translation system (e.g., components and methodologies described in the context of **FIGS. 1 and 2**)

may be implemented, according to an embodiment of the present invention, is shown. It is to be understood that the individual components/steps may be implemented on one such computer system or on more than one such computer system. In the case of an implementation on a distributed computing system, the individual computer systems and/or devices may be connected via a suitable network, e.g., the Internet or World Wide Web. However, the system may be realized via private or local networks. The invention is not limited to any particular network.

[0042] Thus, the computing system shown in **FIG. 3** represents an illustrative computing system architecture for, among other things, a TTS engine, an ASR engine, a language detector, a language translator, and/or combinations thereof, within which one or more of the steps of the voice prompt-based speech-to-speech translation techniques of the invention may be executed.

[0043] As shown, the computer system **300** implementing a speech-to-speech translation system may comprise a processor **302**, a memory **304**, I/O devices **306**, and a communication interface **308**, coupled via a computer bus **310** or alternate connection arrangement.

[0044] It is to be appreciated that the term "processor" as used herein is intended to include any processing device, such as, for example, one that includes a CPU and/or other processing circuitry. It is also to be understood that the term "processor" may refer to more than one processing device and that various elements associated with a processing device may be shared by other processing devices.

[0045] The term "memory" as used herein is intended to include memory associated with a processor or CPU, such as, for example, RAM, ROM, a fixed memory device (e.g., hard drive), a removable memory device (e.g., diskette), flash memory, etc.

[0046] In addition, the phrase "input/output devices" or "I/O devices" as used herein is intended to include, for example, one or more input devices (e.g., keyboard, mouse, microphone, etc.) for entering data to the processing unit, and/or one or more output devices (e.g., speaker, display, etc.) for presenting results associated with the processing unit. Thus, I/O devices **306** collectively represent, among other things, the one or more microphones, output speaker, and screen display referred to above. The system interface (GUI) in **FIG. 2** is displayable in accordance with such a screen display.

[0047] Still further, the phrase "communication interface" as used herein is intended to include, for example, one or more transceivers to permit the computer system to communicate with another computer system via an appropriate communications protocol. That is, if the translation system is distributed (one or more components of the system remotely located from one or more other components), communication interface **308** permits all the components to communicate.

[0048] Accordingly, software components including instructions or code for performing the methodologies described herein may be stored in one or more of the associated memory devices (e.g., ROM, fixed or removable memory) and, when ready to be utilized, loaded in part or in whole (e.g., into RAM) and executed by a CPU.

[0049] Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be made by one skilled in the art without departing from the scope or spirit of the invention.

What is claimed is:

1. A method for use in indicating a dialogue turn in an automated speech-to-speech translation system, comprising the steps of:

obtaining one or more text-based scripts, the one or more text-based scripts being synthesizable into one or more voice prompts; and

synthesizing for playback at least one of the one or more voice prompts from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.

2. The method of claim 1, further comprising the step of detecting a language spoken by a speaker interacting with the speech-to-speech translation system such that a voice prompt in the detected language is synthesized for playback to the speaker.

3. The method of claim 2, wherein an initial voice prompt is synthesized for playback in a default language until the actual language of the speaker is detected.

4. The method of claim 1, further comprising the step of displaying the at least one voice prompt synthesized for playback.

5. The method of claim 1, further comprising the step of recognizing speech uttered by the speaker interacting with the speech-to-speech translation system.

6. The method of claim 5, further comprising the step of recognizing speech uttered by a system user of the speech-to-speech translation system.

7. The method of claim 6, wherein at least a portion of the speech uttered by the speaker or the system user is translated from one language to another language.

8. The method of claim 7, wherein at least a portion of the translated speech is displayed.

9. A method of providing an interface for use in an automated speech-to-speech translation system, the translation system being operated by a system user and interacted with by a speaker, the method comprising the steps of:

the system user enabling a microphone of the translation system via the interface;

outputting at least one previously-generated voice prompt to the speaker, the at least one voice prompt comprising an audible message in a language understandable to the speaker, the audible message indicating a turn in a dialogue between the system user and the speaker; and

the speaker, once prompted, uttering speech into the microphone, the uttered speech being translated by the translation system.

10. The method of claim 9, further comprising the step of displaying text in a first field of the interface representing speech uttered by the system user.

11. The method of claim 10, further comprising the step of displaying text in a second field of the interface representing speech uttered by the speaker.

12. Apparatus for use in indicating a dialogue turn in an automated speech-to-speech translation system, comprising:

a memory; and

at least one processor coupled to the memory and operative to: (i) obtain one or more text-based scripts, the one or more text-based scripts being synthesizable into one or more voice prompts, and (ii) synthesize for playback at least one of the one or more voice prompts from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.

13. The apparatus of claim 12, wherein the at least one processor is further operative to detect a language spoken by a speaker interacting with the speech-to-speech translation system such that a voice prompt in the detected language is synthesized for playback to the speaker.

14. The apparatus of claim 13, wherein an initial voice prompt is synthesized for playback in a default language until the actual language of the speaker is detected.

15. The apparatus of claim 12, wherein the at least one processor is further operative to display the at least one voice prompt synthesized for playback.

16. The apparatus of claim 12, wherein the at least one processor is further operative to recognize speech uttered by the speaker interacting with the speech-to-speech translation system.

17. The apparatus of claim 16, wherein the at least one processor is further operative to recognize speech uttered by a system user of the speech-to-speech translation system.

18. The apparatus of claim 17, wherein at least a portion of the speech uttered by the speaker or the system user is translated from one language to another language.

19. The apparatus of claim 18, wherein at least a portion of the translated speech is displayed.

20. An interface for use in an automated speech-to-speech translation system, the translation system being operated by a system user and interacted with by a speaker, the interface comprising:

a first field for use by the system user to enable a microphone of the translation system;

a second field for use by the system user for at least one of displaying speech uttered by the system user and displaying translated speech uttered by the speaker; and

a third field for use by the speaker for at least one of displaying speech uttered by the speaker and displaying translated speech uttered by the system user;

wherein the translation system outputs at least one previously-generated voice prompt to the speaker, the at least one voice prompt comprising an audible message in a language understandable to the speaker, the audible message indicating a turn in a dialogue between the system user and the speaker, and the speaker, once prompted, uttering speech into the microphone, the uttered speech being translated by the translation system.

21. The interface of claim 20, further comprising a fourth field for use by the system user to enable a microphone of the translation system such that speech uttered by the system user is captured by the translation system.

22. An article of manufacture for use in indicating a dialogue turn in an automated speech-to-speech translation system, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

obtaining one or more text-based scripts, the one or more text-based scripts being synthesizable into one or more voice prompts; and

synthesizing for playback at least one of the one or more voice prompts from at least one of the one or more text-based scripts, the at least one synthesized voice prompt comprising an audible message in a language understandable to a speaker interacting with the speech-to-speech translation system, the audible message indicating a dialogue turn in the automated speech-to-speech translation system.

* * * * *