The disclosure relates to systems, methods and apparatus to convert speech to text and vice versa. One apparatus comprises a vocoder, a speech to text conversion engine, a text to speech conversion engine, and a user interface. The vocoder is operable to convert speech signals into packets and convert packets into speech signals. The speech to text conversion engine is operable to convert speech to text. The text to speech conversion engine is operable to convert text to speech. The user interface is operable to receive a user selection of a mode from among a plurality of modes, wherein a first mode enables the speech to text conversion engine, a second mode enables the text to speech conversion engine, and a third mode enables the speech to text conversion engine and the text to speech conversion engine.
USER SELECTS MODE

NORMAL MODE

SECOND MODE

THIRD MODE

FOURTH MODE

USER LISTENS AND SPEAKS

USER LISTENS AND SENDS TEXT OR SYNTHESIZED SPEECH

USER CONVERTS INCOMING SPEECH TO TEXT AND SPEAKS

USER CONVERTS INCOMING SPEECH TO TEXT AND SENDS TEXT OR SYNTHESIZED SPEECH

FIG. 2
FIRST COMMUNICATION DEVICE 100A

INPUT SPEECH SIGNALS

SPEECH TO TEXT CONVERSION ENGINE 300

TEXT

ENCODED SPEECH PACKETS

ENCODER 302

OUTPUT SPEECH SIGNALS

USER INPUT INTERFACE, e.g., KEYPAD 330

DECODER 304

PACKETS

NETWORK 110A

VOICE SYNTHESIZER 434

TEXT TO SPEECH CONVERSION ENGINE 404

SPEECH SIGNALS

ENCODER 400

SPEECH TO TEXT CONVERSION ENGINE 402

CONTROLLER 444

TRANSCIEVER 114

TEXT

ENCODED SPEECH PACKETS

ENCODED TEXT PACKETS

FIG. 3

FIG. 4
VOICE AND TEXT COMMUNICATION SYSTEM, METHOD AND APPARATUS

TECHNICAL FIELD

[0001] The disclosure relates to communications and, more particularly, to a voice and text communication system, method and apparatus.

BACKGROUND

[0002] A cellular phone may include an audio capture device, such as a microphone and/or speech synthesizer, and an audio encoder to generate audio packets or frames. The phone may use communication protocol layers and modules to transmit packets across a wireless communication channel to a network or another communication device.

SUMMARY

[0003] One aspect relates to an apparatus comprising a vocoder, a speech to text conversion engine, a text to speech conversion engine, and a user interface. The vocoder is operable to convert speech signals into packets and convert packets into speech signals. The text to speech conversion engine is operable to convert text to speech. The text to speech conversion engine is operable to convert text to speech. The user interface is operable to receive a user selection of a mode from among a plurality of modes, wherein a first mode enables the speech to text conversion engine, a second mode enables the text to speech conversion engine, and a third mode enables the speech to text conversion engine.

[0004] Another aspect relates to an apparatus comprising: a vocoder operable to convert speech signals into packets and convert packets into speech signals; a speech to text conversion engine operable to convert speech to text; a selection unit operable to switch between first and second modes, wherein the first mode enables the vocoder, and a second mode enables the text to speech conversion engine; and a transceiver operable to wirelessly transmit encoded speech packets and text packets to a communication network.

[0005] Another aspect relates to a network apparatus comprising: a vocoder operable to convert packets into speech signals; a speech to text conversion engine operable to convert speech to text; a selection unit operable to switch between first and second modes, wherein the first mode enables the vocoder, and a second mode enables the text to speech conversion engine, and a transceiver operable to wirelessly transmit encoded speech packets and text packets to a communication network.

[0006] Another aspect relates to a method comprising: receiving encoded speech packets; converting the received encoded speech packets into speech signals; and receiving a user selection of a mode from among a plurality of modes, wherein a first mode enables speech to text conversion, a second mode enables text to speech conversion, and a third mode enables speech to text and text to speech conversion.

[0007] The details of one or more embodiments are set forth in the accompanying drawings and the description below.

DETAILED DESCRIPTION

[0012] Receiving a call on a mobile device in a meeting, airplane, train, theater, restaurant, church or other place may be disruptive to others. It may be much less disruptive if a user could select another mode on the mobile device to receive the call and/or respond to the call. In one mode, the device receives the call and converts speech/voice signals to text without requiring the caller on the other end to input text.

[0013] FIG. 1 illustrates a system comprising a first communication device 100, a network 110, and a second communication device 120. The system may include other components. The system may use any type of wireless communication, such as Global System for Mobile Communications (GSM), code division multiple access (CDMA), CDMA2000, 1x EV-DO, Wideband CDMA (WCDMA), orthogonal frequency division multiple access (OFDMA), Bluetooth, WiFi, WiMax, etc.

[0014] The first communication device 100 comprises a voice coder (vocoder) 102 and a transceiver 104. The first communication device 100 may include other components in addition to or instead of the components shown in FIG. 1. The first communication device 100 may represent or be implemented in a landline (non-wireless) phone, a wireless communication device, a personal digital assistant (PDA), a handheld device, a laptop computer, a desktop computer, a digital camera, a digital recording device, a network-enabled digital television, a mobile phone, a cellular phone, a satellite telephone, a camera phone, a terrestrial-based radiotelephone, a direct two-way communication device (sometimes referred to as a "walkie-talkie"), a camcorder, etc.

[0015] The vocoder 102 may include an encoder to encode speech signals into packets and a decoder to decode packets into speech signals. The vocoder 102 may be any type of vocoder, such as an enhanced variable rate coder (EVRC), Adaptive Multi-Rate (AMR), Fourth Generation vocoder (4GV), etc. Vocoder ars are described in co-assigned U.S. Patent Nos. 6,397,175, 6,434,519, 6,438,518, 6,449,592, 6,456,964, 6,477,502, 6,584,438, 6,678,649, 6,691,084, 6,804,218, 6,947,888, which are hereby incorporated by reference.

[0016] The transceiver 104 may wirelessly transmit and receive packets containing encoded speech.

[0017] The network 110 may represent one or more base stations, base station controllers (BSCs), mobile switching centers (MSCs), etc. If the first device 100 is a landline phone, then network 110 may include components in a plain old telephone service (POTS) network. The network 110 comprises a vocoder 112 and a transceiver 114. The network 110 may include other components in addition to or instead of the components shown in FIG. 1.

[0018] The second communication device 120 may represent or be implemented in a wireless communication device, a personal digital assistant (PDA), a handheld device, a laptop computer, a desktop computer, a digital camera, a digital recording device, a network-enabled digital television, a mobile phone, a cellular phone, a satellite telephone, a camera
phone, a terrestrial-based radiotelephone, a direct two-way communication device (sometimes referred to as a “walkie-talkie”), a camcorder, etc.

[0019] The second communication device 120 comprises a transceiver 124, a speech and text unit 140, a speaker 142, a display 128, a user input interface 130, e.g., a keypad, and a microphone 146. The speech and text unit 140 comprises a vocoder 122, a speech to text conversion engine 126, a controller 144, a text to speech conversion engine 132, and a voice synthesizer 134. The speech and text unit 140 may include other components in addition to or instead of the components shown in FIG. 1.

[0020] One or more of the components or functions in the speech and text unit 140 may be integrated into a single module, unit, component, or software. For example, the speech to text conversion engine 126 may be combined with the vocoder 122. The text to speech conversion engine 132 may be combined with the vocoder 122, such that text is converted into encoded speech packets. The voice synthesizer 134 may be combined with the vocoder 122 and/or the text to speech conversion engine 132.

[0021] The speech to text conversion engine 126 may convert voice/speech to text. The text to speech conversion engine 132 may convert text to speech. The controller 144 may control operations and parameters of one or more components in the speech and text unit 140.

[0022] The device 120 may provide several modes of communication for a user to receive calls and/or respond to calls, as shown in the table below and in FIG. 2.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Listen</th>
<th>Speak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal mode</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Second mode</td>
<td>Yes</td>
<td>No - transmit text or synthesized speech</td>
</tr>
<tr>
<td>Third mode</td>
<td>No - convert incoming speech to text</td>
<td>Yes</td>
</tr>
<tr>
<td>Fourth mode</td>
<td>No - convert incoming speech to text</td>
<td>No - transmit text or synthesized speech</td>
</tr>
</tbody>
</table>

In a normal mode (blocks 202 and 210), the user of the second device 120 receives a call from the first device 100, listens to speech from the speaker 142, and speaks into the microphone 146.

[0023] FIG. 2 illustrates a method of using the second device 120 of FIG. 1. When the second device 120 receives a call from the first device 100, a user of the second device 120 can select one of the modes via the user interface 130 in block 200. Alternatively, the user may switch between modes in block 200 before the second device 120 receives a call from another device. For example, if the user of the second device 120 enters a meeting, airplane, train, theater, restaurant, church or other place where incoming calls may be disruptive to others, the user may switch from the normal mode to one of the other three modes.

[0024] In a second mode (blocks 204 and 212), the user of the second device 130 may listen to speech from the first device 100, such as using an ear piece, headset, or headphones, but not talk. Instead, the user of the second device 130 may type on the keypad 130 or use a writing stylus to enter handwritten text on the display 128. The display 128 or the text to speech conversion engine 132 may have a module that recognizes handwritten text and characters. The device 120 may (a) send the text to the first device 100 or (b) convert the text to speech with the text to speech conversion engine 132.

[0025] The voice synthesizer 134 may synthesize the speech to produce personalized speech signals to substantially match the user’s natural voice. The voice synthesizer 134 may include a memory that stores characteristics of the user’s voice, such as pitch. A voice synthesizer is described in co-assigned U.S. Patent No. 6,950,799, which is incorporated by reference. Another voice synthesizer is described in co-assigned U.S. patent application Ser. No. 11/398,364, which is incorporated by reference.

[0026] The vocoder 122 encodes the speech into packets. There may or may not be a short delay. In one configuration, other than a short time delay, communication with the second device 120 may appear seamless to the user of the first device 100. If the user of the second device 120 is in a meeting, the conversation may be more message-based than seamless.

[0027] In third and fourth modes (blocks 206, 208, 214 and 216), the device 120 receives a call, and the speech to text conversion engine 126 converts speech/voice signals to text for display on the display 128. In one configuration, the third and fourth modes may allow the user of the first device 100 to continue talking and not require the user of the first device 100 to switch to a text input mode. The speech to text conversion engine 126 may include a voice recognition module to recognize words and sounds to convert them to text.

[0028] In the third mode, the device 120 allows the user to speak into the microphone 146, which passes speech to the vocoder 122 to encode into packets.

[0029] In the fourth mode, the user of the second device 130 may type on the keypad 130 or use a writing stylus to enter handwritten text on the display 128. The device 120 may (a) send the text to the first device 100 or (b) convert the text to speech with the text to speech conversion engine 132. The voice synthesizer 134 may synthesize the speech to produce personalized speech signals to substantially match the user’s natural voice. The vocoder 122 encodes the speech into packets.

[0030] In the second and fourth modes, if the second device 120 is set to convert text to speech and synthesize speech, there may be a time delay between when the second device 120 accepts a call from the first device 100 and when the first device 100 receives speech packets. The second device 120 may be configured to play a pre-recorded message by the user to inform the first device 100 that the user of the second device 120 is in a meeting and will respond using text to speech conversion.

[0031] The second and fourth modes may provide one or more advantages, such as transmitting speech without background noise, no need or reduced need for echo cancellation, no need or reduced need for noise suppression, faster encoding, less processing, etc.

[0032] FIG. 1 shows an example where changes (new functions and/or elements) may be implemented in only the second communication device 120. To realize the new modes (second, third and fourth modes) of communication, the second communication device 120 has a vocoder 122, a speech-to-text engine 126, a text-to-speech engine 132, etc. With this device 120, the system can support the new modes without any changes in the network 110 and conventional phones 100 (landline, mobile phones, etc.). The device 120 may receive and send voice packets regardless of the mode selected by the user.
FIG. 3 illustrates another configuration 100A of the first communication device 100 of FIG. 1. In FIG. 3, the first communication device 100A comprises a speech to text conversion engine 300, an encoder 302, a transceiver 104, a decoder 304, and a user interface 330. The speech to text conversion engine 300 may convert voice/speech to text to be transmitted by the transceiver 104 to the network 110. The first communication device 100A of FIG. 3 may allow the second device 120 to be designed without a speech to text conversion engine 126. The first communication device 100A of FIG. 3 may save bandwidth by sending text instead of speech to the network 110. The user interface 330 may be operable to receive a user selection of a mode from among a plurality of modes, wherein a first mode enables the vocoder 302, 304, and a second mode enables the speech to text conversion engine 300.

FIG. 4 illustrates another configuration 110A of the network 110 of FIG. 1. In FIG. 4, the network 110A comprises a voice coder/decoder 400, a transceiver 114 and a speech to text conversion engine 402. In another configuration, the network 110A may further comprise a text to speech conversion engine 404, a voice synthesizer 402 and a controller 444. The vocoder 400 decodes speech packets to provide speech signals. The speech to text conversion engine 402 may convert voice/speech to text to be transmitted by the transceiver 114 to the second device 120. The network 110A of FIG. 4 may allow the second device 120 to be designed without a speech to text conversion engine 126 or allow the speech to text conversion engine 126 to be deactivated. The network 110A of FIG. 4 may save bandwidth by sending text instead of speech to the second device 120.

The network 110A in FIG. 4 may acquire knowledge of a configuration, situation or preference of the receiving device 120. If the network 110A realizes that the receiving device 120 will not benefit from receiving voice packets (e.g., sensing a user preference or place of the call, for example, an extremely noisy environment and it is difficult to listen to received speech), then the network 110A will transform voice packets to text packets. Even if the receiving device 120 has the ability to change voice packets to text packets (using a speech-to-text engine 126), it can be a waste of bandwidth and device power to do this transformation (from voice to text) if the user is in a text-receiving mode (a meeting, or silent communication in general).

Thus, the network 110A in FIG. 4 may be used in a system where changes (new features and/or elements) are implemented only in the network 110A, i.e., no changes in communication devices or handsets. The network 110A may take care of changing voice packets into text and vice versa where the mobile handsets do not have speech to text conversion units; or if the mobile handsets do have speech to text conversion units, the handsets prefer not to do the conversion or cannot do the conversion due to a lack of computational resources, battery power, etc.

For example, the first device 100 in FIG. 1 can send/receive voice packets (i.e., first mode), while the second device 120 sends/receives text (i.e., fourth mode). The second device 120 may not have unit 140 (or just have a vocoder 122) or have unit 140 deactivated. To allow the second device 120 to operate in the fourth mode, the network 110A in FIG. 4 will change the first device’s voice packets into text packets (using the speech-to-text engine 402) to send to the second device 120 and will change text packets from the second device 120 to voice packets (using the text-to-speech engine 404) to send to the first device 100.

If the second device 120 does not have the unit 140, the second device 120 can signal (in-band for example) a desired mode to the network 110A and thus ask the network 110A to convert between speech and text, i.e., do the functions of unit 140.

Personalized speech synthesis may be done in the network 110A. As described above, the unit 140 in FIG. 1 has a voice synthesizer 134 to change the output of the text-to-speech engine 132 to personalized speech (the user’s voice). In a system with the network 110A of FIG. 4, to produce voice packets that carry a voice signature of the user of the second device 120, the second device 120 may send stored voice packets (at the beginning of using second or fourth modes) that have the spectral parameters and pitch information of the user to the network 110A. These few transmitted voice packets (preceding the text packets) can be used by the network 110A to produce personalized voice packets.

An example of transmitting packets for second or fourth modes from the second device 120 to the network 110A is described. The second device 120 transmits to the network 110A at the beginning of using these “text modes” (second or fourth modes) user pre-stored voice packets (N packets) plus a mode of operation (1, 2, 3, or 4; request to do the conversion). The second device 120 may then send text packets.

A combination of the two configurations (FIG. 1 and FIG. 4) is also possible. When using one of these modes, the network 110A will enable the text/speech conversion after sensing (e.g., receiving a request via signaling) the capability of the receiving device 120, which does the conversion or lets the network 110A or receiving device 100A does the conversion.

One or more components and features described above may be implemented in a push to talk (PTT) or push to read communication device. A PTT device allows a user to push a button on the device and talk, while the device converts speech to text and transmits text packets to a network or directly to another communication device. PTT communication is “message based,” rather than continuous, such as a standard voice call. A time period over which a user holds down the PTT button on the device may nicely frame the message that is then converted to text, etc.

The device 120 may have a dedicated memory for storing instructions and data, as well as dedicated hardware, software, firmware, or combinations thereof. If implemented in software, the techniques may be embodied as instructions on a computer-readable medium such as random access memory (RAM), read-only memory (ROM), non-volatile random access memory (NVRAM), electrically erasable programmable read-only memory (EEPROM), FLASH memory, or the like. The instructions cause one or more processors to perform certain aspects of the functionality described in this disclosure.

The techniques described in this disclosure may be implemented within a general purpose microprocessor, digital signal processor (DSP), application specific integrated circuit (ASIC), field programmable gate array (FPGA), or other equivalent logic devices. For example, the speech and text unit 140 and associated components and modules, may be implemented as parts of an encoding process, or coding/decoding (CODEC) process, running on a digital signal pro-
cessor (DSP) or other processing device. Accordingly, components described as modules may form programmable features of such a process, or a separate process.

[0045] The speech and text unit 140 may have a dedicated memory for storing instructions and data, as well as dedicated hardware, software, firmware, or combinations thereof. If implemented in software, the techniques may be embodied as instructions executable by one or more processors. The instructions may be stored on a computer-readable medium such as random access memory (RAM), read-only memory (ROM), non-volatile random access memory (NVRAM), electrically erasable programmable read-only memory (EEPROM), FLASH memory, magnetic or optical data storage device, or the like. The instructions cause one or more processors to perform certain aspects of the functionality described in this disclosure.

[0046] Various embodiments have been described. These and other embodiments are within the scope of the following claims.

1. An apparatus comprising:
a vocoder operable to convert speech signals into packets and convert packets into speech signals;
a speech to text conversion engine operable to convert speech to text;
a text to speech conversion engine operable to convert text to speech; and

2. The apparatus of claim 1, further comprising a display to display text from the speech to text conversion engine.

3. The apparatus of claim 1, further comprising a keypad to receive input text from a user.

4. The apparatus of claim 1, wherein the user interface is operable to receive a user selection of a mode after the apparatus receives a call from another apparatus.

5. A method comprising:
a vocoder operable to convert speech signals into packets and convert packets into speech signals;
a speech to text conversion engine operable to convert speech to text;
a user interface operable to receive a user selection of a mode from among a plurality of modes, wherein a first mode enables the vocoder, and a second mode enables the speech to text conversion engine; and

6. The apparatus of claim 1, further comprising a voice synthesizer to synthesize a user's voice.

7. The apparatus of claim 1, further comprising a transceiver operable to wirelessly transmit encoded speech packets and text packets to a communication network.

8. An apparatus comprising:
a vocoder operable to convert speech signals into packets and convert packets into speech signals;
a speech to text conversion engine operable to convert speech to text;
a user interface operable to receive a user selection of a mode from among a plurality of modes, wherein a first mode enables the vocoder, and a second mode enables the speech to text conversion engine; and

9. The apparatus of claim 8, further comprising a display to display text from the speech to text conversion engine.

10. The apparatus of claim 8, further comprising a keypad to receive input text from a user.

11. The apparatus of claim 8, wherein the user interface is operable to receive a user selection of a mode before the apparatus receives a call from another apparatus.

12. The apparatus of claim 8, wherein the user interface is operable to receive a user selection of a mode after the apparatus receives a call from another apparatus.

13. A network apparatus comprising:
a vocoder operable to convert packets into speech signals;
a speech to text conversion engine operable to convert speech to text;
a selection unit operable to switch between first and second modes, wherein the first mode enables the vocoder, and a second mode enables the vocoder and the speech to text conversion engine; and

14. The network apparatus of claim 13, further comprising a display to text conversion engine operable to convert text to speech, wherein the selection unit is operable to switch to a third mode when the vocoder and both conversion engines are enabled.

15. The network apparatus of claim 14, further comprising a voice synthesizer operable to synthesize a user's voice from text converted to speech.

16. The network apparatus of claim 15, wherein the voice synthesizer is operable to receive and store voice characteristics of a user's voice.

17. The network apparatus of claim 13, further comprising a controller operable to receive a request from a communication device to convert speech to text.

18. The network apparatus of claim 13, further comprising a controller operable to receive a request from a communication device to convert text to speech.

19. A method comprising:
receiving encoded speech packets;
converting the received encoded speech packets into speech signals; and
receiving a user selection of a mode from among a plurality of modes, wherein a first mode enables speech to text conversion, a second mode enables text to speech conversion, and a third mode enables speech to text and text to speech conversion.

20. The method of claim 19, further comprising receiving a user selection for a mode before receiving an incoming call.

21. The method of claim 19, further comprising receiving a user selection for a mode after receiving an incoming call.

* * * * *