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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 May 2011 (19.05.2011)

PCT

(10) International Publication Number
WO 2011/059997 A1

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2010/056109

(22) International Filing Date:
10 November 2010 (10.11.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/259,820 10 November 2009 (10.11.2009) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, WA 98005 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KENNEWICK, Mike** [US/US]; 13407 37th Place, NE, Bellevue, WA 98005 (US). **ARMSTRONG, Lynn, Elise** [US/US]; 19807 183rd Place, NE, Woodinville, WA 98077 (US).

(74) Agents: **ALI, Syed, Jafar** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

WO 2011/059997 A1

(54) Title: SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE

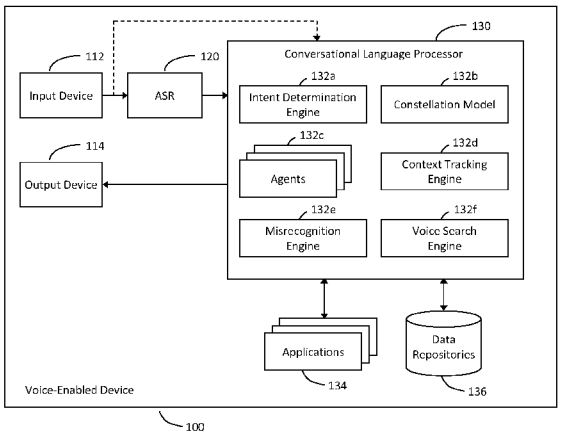

Figure 1

(57) Abstract: The system and method described herein may provide a natural language content dedication service in a voice services environment. In particular, providing the natural language content dedication service may generally include detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE

CROSS-REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of U.S. Provisional Patent Application Serial No. 61/259,820, entitled "System and Method for Providing a Natural Language Content Dedication Service," filed November 10, 2009, the contents of which are hereby incorporated by reference in their entirety.

FIELD OF THE INVENTION

[002] The invention generally relates to providing a natural language content dedication service in a voice services environment, and in particular, to detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

BACKGROUND OF THE INVENTION

[003] As technology has progressed in recent years, consumer electronic devices have emerged to become nearly ubiquitous in the everyday lives of many people. To meet the increasing demand that has resulted from growth in the functionality and mobility of mobile phones, navigation devices, embedded devices, and other such devices, many devices offer a wealth of features and functions in addition to core applications. Greater functionality also introduces trade-offs, however, including learning curves that often inhibit users from fully exploiting all of the capabilities of their electronic devices. For example, many existing electronic devices include complex human to machine interfaces that may not be particularly user-friendly, which can inhibit mass-market adoption for many technologies. Moreover, cumbersome interfaces often result in otherwise desirable features being difficult to find or use (e.g., because of menus that are complex or otherwise tedious to navigate). As such, many users tend not to use, or even know about, many of the potential capabilities of their devices.

[004] As such, the increased functionality of electronic devices often tends to be wasted, as market research suggests that many users only use only a fraction of the features or applications available on a given device. Moreover, in a society where wireless networking and broadband access are increasingly prevalent, consumers tend to naturally desire seamless mobile capabilities from their electronic devices. Thus, as consumer demand intensifies for simpler mechanisms to interact with electronic devices, cumbersome interfaces that prevent quick and focused interaction become an important concern. Nevertheless, the ever-growing demand for mechanisms to use technology in intuitive ways remains largely unfulfilled.

[005] One approach towards simplifying human to machine interactions in electronic devices has included the use of voice recognition software, which has the potential to enable users to exploit features that would otherwise be unfamiliar, unknown, or difficult to use. For example, a recent survey conducted by the Navteq Corporation, which provides data used in a variety of applications such as automotive navigation and web-based applications, demonstrates that voice recognition often ranks among the features most desired by consumers of electronic devices. Even so, existing voice user interfaces, when they actually work, still require significant learning on the part of the user.

[006] For example, many existing voice user interface only support requests formulated according to specific command-and-control sequences or syntaxes. Furthermore, many existing voice user interfaces cause user frustration or dissatisfaction because of inaccurate speech recognition. Similarly, by forcing a user to provide pre-established commands or keywords to communicate requests in ways that a system can understand, existing voice user interfaces do not effectively engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation towards a satisfactory goal (e.g., when users may be uncertain of particular needs, available information, device capabilities, etc.). As such, existing voice user interfaces tend to suffer from various drawbacks, including significant limitations on engaging users in a dialogue in a cooperative and conversational manner.

[007] Additionally, many existing voice user interfaces fall short in utilizing information distributed across different domains, devices, and applications in order to resolve natural language voice-based inputs. Thus, existing voice user interfaces suffer from

being constrained to a finite set of applications for which they have been designed, or to devices on which they reside. Although technological advancement has resulted in users often having several devices to suit their various needs, existing voice user interfaces do not adequately free users from device constraints. For example, users may be interested in services associated with different applications and devices, but existing voice user interfaces tend to restrict users from accessing the applications and devices as they see fit. Moreover, users typically can only practicably carry a finite number of devices at any given time, yet content or services associated with users' devices other than those currently being used may be desired in various circumstances.

[008] Accordingly, although users tend to have varying needs, where content or services associated with different devices may be desired in various contexts or environments, existing voice technologies tend to fall short in providing an integrated environment in which users can request content or services associated with virtually any device or network. As such, constraints on information availability and device interaction mechanisms in existing voice services environments tend to prevent users from experiencing technology in an intuitive, natural, and efficient way. For instance, when a user wishes to perform a given function using a given electronic device, but does not necessarily know how to go about performing the function, the user typically cannot engage in cooperative multi-modal interactions with the device to simply utter words in natural language to request the function.

[009] Furthermore, relatively simple functions can often be tedious to perform using electronic devices that do not have voice recognition capabilities. For example, purchasing new ring-tones for a mobile phone tends to be a relatively straightforward process, but users must typically navigate several menus and press many different buttons in order to complete the process. In another example, users often listen to music or interact with other media in mobile environments, such that interest in purchasing music, media, or other content may be fleeting or often occur on an impulse basis. Whereas existing human to machine interfaces that lack voice recognition capabilities typically fall short in providing mechanisms that can readily meet this demand, adding voice recognition to an electronic device can substantially simplify human to machine interaction in a manner that can meet user needs, improve experience, and satisfy potentially transient consumer interests. As such, interaction

with electronic devices could be made far more efficient if users were provided with the ability to use natural language in order to exploit buried or otherwise difficult to use functionality.

SUMMARY OF THE INVENTION

[010] According to one aspect of the invention, a system and method for providing a natural language content dedication service may generally operate in a voice services environment that includes one or more electronic devices that can receive multi-modal natural language device interactions. In particular, providing the natural language content dedication service may generally include detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

[011] According to one aspect of the invention, the natural language content dedication service may operate in a hybrid processing environment, which may generally include a plurality of multi-modal devices configured to cooperatively interpret and process natural language utterances included in the multi-modal device interactions. For example, a virtual router may receive messages that include encoded audio corresponding to natural language utterances contained in the multi-modal device interactions, which may be received at one or more of the plurality of multi-modal devices in the hybrid processing environment. For example, the virtual router may analyze the encoded audio to select a cleanest sample of the natural language utterances and communicate with one or more other devices in the hybrid processing environment to determine an intent of the multi-modal device interactions. The virtual router may then coordinate resolving the multi-modal device interactions based on the intent of the multi-modal device interactions.

[012] According to one aspect of the invention, a method for providing the natural language content dedication service may comprise detecting a multi-modal device interaction at an electronic device, wherein the multi-modal device interaction may include at least a natural language utterance. One or more messages containing information relating to the multi-modal device interaction may then be communicated to

the virtual router through a messaging interface. The electronic device may then receive one or more messages (e.g., from the virtual router through message interface), wherein the messages may contain information relating to an intent of the multi-modal device interaction. As such, the multi-modal device interaction may be resolved at the electronic device based on the information contained in the one or more messages received from the virtual router.

[013] According to one aspect of the invention, a system for providing the natural language content dedication service may generally include a voice-enabled client device that can communicate with a content dedication system through the messaging interface. The content dedication system include a voice-enabled server, which may be configured to communicate with the virtual router through another messaging interface, or the content dedication system may alternatively include the virtual router. In addition, the content dedication system may further include a billing system for processing transactions relating to the content dedication service. The natural language content dedication service may be provided on any suitable voice-enabled client device having a suitable combination of input and output devices that can receive and respond to multi-modal device interactions that include natural language utterances, and the input and output devices may be further arranged to receive and respond to any other suitable type of input and output.

[014] According to one aspect of the invention, operating the natural language content dedication service may generally include a user of the voice-enabled client device listening to music, watching video, or otherwise interacting with content and providing a multi-modal natural language request to engage in a transaction to dedicate the music, video, or other content. Furthermore, the voice-enabled client device may be included within the hybrid processing environment that includes the plurality of multi-modal devices, whereby the content dedication request may relate to content played on a different device from the voice-enabled client device, although the content dedication request may relate to any suitable content (i.e., the request need not necessarily relate to played content, as users may provide natural language to request content dedications for any suitable content, including a particular song or video that the user may be thinking about).

[015] According to one aspect of the invention, in response to the voice-enabled client device receiving a multi-modal interaction that includes a natural language utterance,

the voice-enabled client device may invoke an Automatic Speech Recognizer (ASR) to generate a preliminary interpretation of the utterance. The ASR may then provide the preliminary interpretation of the utterance to a conversational language processor, which may attempt to determine an intent for the multi-modal interaction. For example, to determine the intent for the multi-modal interaction, the conversational language processor may determine a most likely context for the interaction from the preliminary interpretation of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared knowledge, or any other suitable information for interpreting the multi-modal interaction. Thus, in response to the conversational language processor determining that the intent for the multi-modal interaction relates to a content dedication request, a content dedication application may be invoked to resolve the content dedication request.

[016] According to one aspect of the invention, to resolve the intent of the content dedication request, the conversational language processor may search one or more data repositories that contain content information to identify content matching criteria contained in the content dedication request. Moreover, the conversational language processor may further cooperate with other devices in the hybrid processing environment to search for or otherwise identify the content requested for dedication (e.g., in response to a local data repository not yielding adequate results, another device in the hybrid processing environment having a larger content data repository than the client device may be invoked.). The conversational language processor may then receive appropriate results identifying the content requested for dedication and present the results to the user through the output device (e.g., displaying information about the content, playing a sample clip of the content, displaying options to purchase the content, recommending similar content, etc.). The results presented through the output device may further include an option to confirm the content dedication, wherein the user may confirm the content dedication in a natural language utterance, a non-speech input, or any suitable combination thereof. The content dedication application may then be invoked to process the content dedication request.

[017] According to one aspect of the invention, to process the content dedication request, the content dedication application may capture a natural language utterance that contains the dedication to accompany the content. The user may then provide the

dedication utterance through the voice-enabled input device, and the dedication utterance may then be converted into an electronic signal that the content dedication application captures for the dedication. In addition, the content dedication application may prompt the user to provide any additional tags for the dedicated content (e.g., an image to insert as album art in the dedicated content, an utterance to insert or transcribe into metadata tags for the dedicated content, a non-speech or data input to insert in the metadata tags for the dedicated content, etc.). The content dedication application may further prompt the user to identify a recipient of the dedication, wherein the user may provide any suitable multi-modal input that includes information identifying the recipient of the content dedication. The content dedication application may then route the request to the content dedication system, which may process a transaction for the content dedication.

[018] According to one aspect of the invention, processing the transaction for the content dedication may include the content dedication system receiving encoded audio corresponding to the dedication utterance through the messaging interface. The content dedication system may then insert the encoded audio corresponding to the dedication utterance within the dedicated content, verbally annotate the dedicated content with the encoded audio, and/or transcribe the dedication utterance into a textual annotation for the dedicated content. Similarly, any utterances to insert into the metadata tags for the dedicated content may provide further verbal annotations for the dedicated content, and any such utterances may also be transcribed into text to provide further textual annotations for the dedicated content. The content dedication system may then invoke a content dedication application hosted on the voice-enabled server to process the request. In particular, the content dedication application hosted on the voice-enabled server may identify the content requested for dedication (e.g., if the content dedication application on the voice-enabled client device was unable to suitably identify the requested content, or the information communicated from the voice-enabled client device may identify the requested content if the content dedication application on the voice-enabled client device was able to suitably identify the requested content, etc.). In response to identifying the content to be dedicated, the content dedication system may then communicate with the billing system to process an appropriate transaction for the dedication request based on a selected purchase option for the dedication request.

[019] According to one aspect of the invention, the content dedication system may support various purchase options to provide users with flexibility in requesting content dedications. For example, a buy-to-own purchase option may include the content dedication system purchasing full rights to the content from an appropriate content provider. The billing system may then charge the user of the voice-enabled client device an appropriate amount that encompasses the cost for purchasing the rights to the content from the content provider and a service charge for customizing the content with the dedication and any additional tags and subsequently delivering the customized content to the dedication recipient. The user may be charged in a similar manner under a pay-to-play purchase option, except that the rights purchased from the content provider may be limited (e.g., to a predetermined number of plays). Thus, the cost for purchasing the content from the content provider may be somewhat less under the pay-to-play purchase option, such that the billing system may charge the user somewhat less under the pay-to-play purchase option. Under a paid subscription purchase option, the user may pay a periodic service charge to the content dedication system that permits the user to make content dedications based on terms of the subscription (e.g., a predetermined number an unlimited number of content dedications may be made in a subscription period depending on the particular terms of the user's subscription). Furthermore, other purchase options may be suitably employed, as will be apparent.

[020] According to one aspect of the invention, a service provider associated with the content dedication system may negotiate agreements with content providers to determine the manner in which revenues for content transactions will be shared between the content dedication system and the content providers. For example, an agreement may permit a particular content provider to keep all of the revenue for transactions that include purchasing content from the content provider, while the service provider associated with the content dedication system agrees to recoup any such costs from users. In another example, an agreement may share revenue for content transactions between a content provider and the service provider. Thus, the agreements may generally include any suitable arrangements that define the manner in which content providers and the service provider agree to divide revenue for transactions in which the content dedication system purchases content from the content provider, while the service provider associated with the content dedication

system may be responsible for billing users for any natural language aspects of the content dedication service (e.g., for inserting dedication utterances, transcribing utterances into metadata tags, delivering the content to recipients, etc.).

[021] According to one aspect of the invention, in response to purchasing the requested content from an appropriate service provider and determining the appropriate billing options for the content dedication, the content dedication system may insert the natural language dedication utterance into the dedicated content, verbally annotate the dedicated content with the dedication utterance, or otherwise associate the content with the dedication utterance. Furthermore, the content dedication system may determine whether any additional tags have been specified for the dedicated content and insert such additional tags into the dedicated content, as appropriate (e.g., inserting an image or picture into metadata tags corresponding to album art for the dedicated content, transcribing natural language utterances, non-voice, and/or data inputs into text and inserting such text into the metadata tags for the dedicated content, etc.). The content dedication system may then send a content dedication message to the recipient of the dedication, wherein the message may include a link that the recipient can select to stream, download, or otherwise access the dedicated content, the dedication utterance, etc.). Thus, the content dedication message may generally notify the recipient that content has been dedicated to the recipient and provide various mechanisms for the recipient to access the content dedication, as will be apparent. Furthermore, if the purchase option selected for the content dedication includes the buy-to-own purchase option or another purchase option that confers full rights to the dedicated content, the recipient may then own the full rights to the dedicated content, otherwise the recipient may own rights with respect to the dedicated content based on whatever terms the selected purchase option provides.

[022] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[023] Figure 1 illustrates a block diagram of an exemplary voice-enabled device that can be used for hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[024] Figure 2 illustrates a block diagram of an exemplary system for hybrid processing in a natural language voice service environment, according to one aspect of the invention.

[025] Figure 3 illustrates a flow diagram of an exemplary method for initializing various devices that cooperate to perform hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[026] Figures 4-5 illustrate flow diagrams of exemplary methods for hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[027] Figure 6 illustrates a block diagram of an exemplary system for providing a natural language content dedication service, according to one aspect of the invention.

[028] Figures 7-8 illustrate flow diagrams of exemplary methods for providing a natural language content dedication service, according to one aspect of the invention.

DETAILED DESCRIPTION

[029] According to one aspect of the invention, Figure 1 illustrates a block diagram of an exemplary voice-enabled device 100 that can be used for hybrid processing in a natural language voice services environment. As will be apparent from the further description to be provided herein, the voice-enabled device 100 illustrated in Figure 1 may generally include an input device 112, or a combination of input devices 112, which may enable a user to interact with the voice-enabled device 100 in a multi-modal manner. In particular, the input devices 112 may generally include any suitable combination of at least one voice input device 112 (e.g., a microphone) and at least one non-voice input device 112 (e.g., a mouse, touch-screen display, wheel selector, etc.). As such, the input devices 112 may include any suitable combination of electronic devices having mechanisms for receiving both voice-based and non-voice-based inputs (e.g., a microphone coupled to one or more of a telematics device, personal navigation device, mobile phone, VoIP node, personal computer, media device, embedded device, server, or other electronic device).

[030] In one implementation, the voice-enabled device 100 may enable the user to engage in various multi-modal conversational interactions, which the voice-enabled device 100 may process in a free-form and cooperative manner to execute various tasks, resolve various queries, or otherwise resolve various natural language requests

included in the multi-modal interactions. For example, in one implementation, the voice-enabled device 100 may include various natural language processing components, including at least a voice-click module coupled to the one or more input devices 112, as described in further detail in co-pending U.S. Patent Application Serial No. 12/389,678, entitled "System and Method for Processing Multi-Modal Device Interactions in a Natural Language Voice Services Environment," filed February 20, 2009, the contents of which are hereby incorporated by reference in their entirety. Thus, as will be described in further detail herein, the one or more input devices 112 and the voice-click module may be collectively configured to process various multi-modal interactions between the user and the voice-enabled device 100.

[031] For example, in one implementation, the multi-modal interactions may include at least one natural language utterance, wherein the natural language utterance may be converted into an electronic signal. The electronic signal may then be provided to an Automatic Speech Recognizer (ASR) 120, which may also be referred to as a speech recognition engine 120 and/or a multi-pass speech recognition engine 120. In response to receiving the electronic signal corresponding to the utterance, the ASR 120 may generate one or more preliminary interpretations of the utterance and provide the preliminary interpretation to a conversational language processor 130. Additionally, in one implementation, the multi-modal interactions may include one or more non-voice interactions with the one or more input devices 112 (e.g., button pushes, multi-touch gestures, point of focus or attention focus selections, etc.). As such, the voice-click module may extract context from the non-voice interactions and provide the context to the conversational language processor 130 for use in generating an interpretation of the utterance (i.e., via the dashed line illustrated in Figure 1). As such, as described in greater detail below, the conversational language processor 130 may analyze the utterance and any accompanying non-voice interactions to determine an intent of the multi-modal interactions with the voice-enabled device 100.

[032] In one implementation, as noted above, the voice-enabled device 100 may include various natural language processing components that can support free-form utterances and/or other forms of non-voice device interactions, which may liberate the user from restrictions relating to the manner of formulating commands, queries, or other requests. As such, the user may provide the utterance to the voice input device 112 using any manner of speaking, and may further provide other non-voice

interactions to the non-voice input device 112 to request any content or service available through the voice-enabled device 100. For instance, in one implementation, in response to receiving the utterance at the voice input device 112, the utterance may be processed using techniques described in U.S. Patent Application Serial No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,398,209 on July 8, 2008, and co-pending U.S. Patent Application Serial No. 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the contents of which are hereby incorporated by reference in their entirety. In addition, the user may interact with one or more of the non-voice input devices 112 to provide buttons pushes, multi-touch gestures, point of focus or attention focus selections, or other non-voice device interactions, which may provide further context or other information relating to the natural language utterances and/or the requested content or service.

[033] In one implementation, the voice-enabled device 100 may be coupled to one or more additional systems that may be configured to cooperate with the voice-enabled device 100 to interpret or otherwise process the multi-modal interactions that include combinations of natural language utterances and/or non-voice device interactions. For example, as will be described in greater detail below in connection with Figure 2, the one or more additional systems may include one or more multi-modal voice-enabled devices having similar natural language processing capabilities to the voice-enabled device 100, one or more non-voice devices having data retrieval and/or task execution capabilities, and a virtual router that coordinates interaction among the voice-enabled device 100 and the additional systems. As such, the voice-enabled device 100 may include an interface to an integrated natural language voice services environment that includes a plurality of multi-modal devices, wherein the user may request content or services available through any of the multi-modal devices.

[034] For example, in one implementation, the conversational language processor 130 may include a constellation model 132b that provides knowledge relating to content, services, applications, intent determination capabilities, and other features available in the voice services environment, as described in co-pending U.S. Patent Application Serial No. 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May

27, 2008, the contents of which are hereby incorporated by reference in their entirety. As such, the voice-enabled device 100 may have access to shared knowledge relating to natural language processing capabilities, context, prior interactions, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models for the various systems and multi-modal devices, providing a cooperative environment for resolving the multi-modal interactions received at the voice-enabled device 100.

[035] In one implementation, the input devices 112 and the voice-click module coupled thereto may be configured to continually monitor for one or more multi-modal interactions received at the voice-enabled device 100. In particular, the input devices 112 and the voice-click module may continually monitor for one or more natural language utterances and/or one or more distinguishable non-voice device interactions, which may collectively provide the relevant context for retrieving content, executing tasks, invoking services or commands, or processing any other suitable requests. Thus, in response to detecting one or more multi-modal interactions, the input devices 112 and/or the voice-click module may signal the voice-enabled device 100 that an utterance and/or a non-voice interaction have been received. For example, in one implementation, the non-voice interaction may provide context for sharpening recognition, interpretation, and understanding of an accompanying utterance, and moreover, the utterance may provide further context for enhancing interpretation of the accompanying non-voice interaction. Accordingly, the utterance and the non-voice interaction may collectively provide relevant context that various natural language processing components may use to determine an intent of the multi-modal interaction that includes the utterance and the non-voice interaction.

[036] In one implementation, as noted above, processing the utterance included in the multi-modal interaction may be initiated at the ASR 120, wherein the ASR 120 may generate one or more preliminary interpretations of the utterance. In one implementation, to generate the preliminary interpretations of the utterance, the ASR 120 may be configured to recognize one or more syllables, words, phrases, or other acoustic characteristics from the utterance using one or more dynamic recognition grammars and/or acoustic models. For example, in one implementation, the ASR 120 may use the dynamic recognition grammars and/or the acoustic models to recognize a stream of phonemes from the utterance based on phonetic dictation techniques, as described in U.S. Patent Application Serial No. 11/513,269, entitled “Dynamic Speech

Sharpening," which issued as U.S. Patent No. 7,634,409 on December 15, 2009, the contents of which are hereby incorporated by reference in their entirety. In addition, the dynamic recognition grammars and/or the acoustic models may include unstressed central vowels (e.g., "schwa"), which may reduce a search space for recognizing the stream of phonemes for the utterance.

[037] Furthermore, in one implementation, the ASR 120 may be configured as a multi-pass speech recognition engine 120, as described in U.S. Patent Application Serial No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,640,160 on December 29, 2009, the contents of which are hereby incorporated by reference in their entirety. The multi-pass speech recognition 120 may be configured to initially invoke a primary speech recognition engine to generate a first transcription of the utterance, and further to optionally subsequently invoke one or more secondary speech recognition engines to generate one or more secondary transcriptions of the utterance. In one implementation, the first transcription may be generated using a large list dictation grammar, while the secondary transcriptions may be generated using virtual dictation grammars having decoy words for out-of-vocabulary words, reduced vocabularies derived from a conversation history, or other dynamic recognition grammars. For example, in one implementation, if a confidence level for the first transcription does not meet or exceed a threshold, the secondary speech recognition engines may be invoked to sharpen the interpretation of the primary speech recognition engine. It will be apparent, however, that the multi-pass speech recognition engine 120 may interpret the utterance using any suitable combination of techniques that results in a preliminary interpretation derived from a plurality of transcription passes for the utterance (e.g., the secondary speech recognition engines may be invoked regardless of the confidence level for the first transcription, or the primary speech recognition engine and/or the secondary speech recognition engines may employ recognition grammars that are identical or optimized for a particular interpretation context, etc.).

[038] Accordingly, in one implementation, the dynamic recognition grammars used in the ASR 120 may be optimized for different languages, contexts, domains, memory constraints, and/or other suitable criteria. For example, in one implementation, the voice-enabled device 100 may include one or more applications 134 that provide content or services for a particular context or domain, such as a navigation application

134. As such, in response to the ASR 120 determining navigation as the most likely context for the utterance, the dynamic recognition grammars may be optimized for various physical, temporal, directional, or other geographical characteristics (e.g., as described in co-pending U.S. Patent Application Serial No. 11/954,064, entitled "System and Method for Providing a Natural Language Voice User Interface in an Integrated Voice Navigation Services Environment," filed December 11, 2007, the contents of which are hereby incorporated by reference in their entirety). In another example, an utterance containing the word "traffic" may be subject to different interpretations depending on whether the user intended a navigation context (i.e., traffic on roads), a music context (i.e., the 1960's rock band), or a movie context (i.e., the Steven Soderbergh film). Accordingly, the recognition grammars used in the ASR 120 may be dynamically adapted to optimize accurate recognition for any given utterance (e.g., in response to incorrectly interpreting an utterance to contain a particular word or phrase, the incorrect interpretation may be removed from the recognition grammar to prevent repeating the incorrect interpretation).

[039] In one implementation, in response to the ASR 120 generating the preliminary interpretations of the utterance included in the multi-modal interaction using one or more of the techniques described above, the ASR 120 may provide the preliminary interpretations to the conversational language processor 130. The conversational language processor 130 may generally include various natural language processing components, which may be configured to model human-to-human conversations or interactions. Thus, the conversational language processor 130 may invoke one or more of the natural language processing components to further analyze the preliminary interpretations of the utterance and any accompanying non-voice interactions to determine the intent of the multi-modal interactions received at the voice-enabled device 100.

[040] In one implementation, the conversational language processor 120 may invoke an intent determination engine 130a configured to determine the intent of the multi-modal interactions received at the voice-enabled device 100. In one implementation, the intent determination engine 130a may invoke a knowledge-enhanced speech recognition engine that provides long-term and short-term semantic knowledge for determining the intent, as described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language

Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety. For example, in one implementation, the semantic knowledge may be based on a personalized cognitive model derived from one or more prior interactions with the user, a general cognitive model derived from one or more prior interactions with various different users, and/or an environmental cognitive model derived from an environment associated with the user, the voice-enabled device 100, and/or the voice services environment (e.g., ambient noise characteristics, location sensitive information, etc.).

[041] Furthermore, the intent determination engine 132a may invoke a context tracking engine 132d to determine the context for the multi-modal interactions. For example, any context derived from the natural language utterance and/or the non-voice interactions in the multi-modal interactions may be pushed to a context stack associated with the context tracking engine 132d, wherein the context stack may include various entries that may be weighted or otherwise ranked according to one or more contexts identified from the cognitive models and the context for the current multi-modal interactions. As such, the context tracking engine 132d may determine one or more entries in the context stack that match information associated with the current multi-modal interactions to determine a most likely context for the current multi-modal interactions. The context tracking engine 132d may then provide the most likely context to the intent determination engine 132a, which may determine the intent of the multi-modal interactions in view of the most likely context.

[042] In addition, based on the most likely context, the intent determination engine 132a may reference the constellation model 132b to determine whether to invoke any of the various systems or multi-modal devices in the voice services environment. For example, as noted above, the constellation model 132b may provide intent determination capabilities, domain knowledge, semantic knowledge, cognitive models, and other information available through the various systems and multi-modal devices. As such, the intent determination engine 132a may reference the constellation model 132b to determine whether one or more of the other systems and/or multi-modal devices should be engaged to participate in determining the intent of the multi-modal interactions. For example, in response to the constellation model 132b indicating that one or more of the other systems and/or multi-modal devices have natural language processing capabilities optimized for the most likely context, the intent determination

engine 132a may forward information relating to the multi-modal interactions to such systems and/or multi-modal devices, which may then determine the intent of the multi-modal interactions and return the intent determination to the voice-enabled device 100.

[043] In one implementation, the conversational language processor 130 may be configured to engage the user in one or more cooperative conversations to resolve the intent or otherwise process the multi-modal interactions, as described in co-pending U.S. Patent Application Serial No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, the contents of which are hereby incorporated by reference in their entirety. In particular, the conversational language processor 130 may generally identify a conversational goal for the multi-modal interactions, wherein the conversational goal may be identifying from analyzing the utterance, the non-voice interactions, the most likely context, and/or the determined intent. As such, the conversational goal identified for the multi-modal interactions may generally control the cooperative conversation between the conversational language processor 130 and the user. For example, the conversational language processor 130 may generally engage the user in one or more query conversations, didactic conversations, and/or exploratory conversations to resolve or otherwise process the multi-modal interactions.

[044] In particular, the conversational language processor 130 may engage the user in a query conversation in response to identifying that the conversational goal relates to retrieving discrete information or performing a particular function. Thus, in a cooperative query conversation, the user may lead the conversation towards achieving the particular conversational goal, while the conversational language processor 130 may initiate one or more queries, tasks, commands, or other requests to achieve the goal and thereby support the user in the conversation. In response to ambiguity or uncertainty in the intent of the multi-modal interaction, the conversational language processor 130 may engage the user in a didactic conversation to resolve the ambiguity or uncertainty (e.g., where noise or malapropisms interfere with interpreting the utterance, multiple likely contexts cannot be disambiguated, etc.). As such, in a cooperative didactic conversation, the conversational language processor 130 may lead the conversation to clarify the intent of the multi-modal interaction (e.g., generating feedback provided through an output device 114), while the user may regulate the conversation and provide additional multi-modal interactions to clarify the

intent. In response to determining the intent of the multi-modal interactions with suitable confidence, with the intent indicating an ambiguous or uncertain goal, the conversational language processor 130 may engage the user in an exploratory conversation to resolve the goal. In a cooperative exploratory conversation, the conversational language processor 130 and the user may share leader and supporter roles, wherein the ambiguous or uncertain goal may be improvised or refined over a course of the conversation.

[045] Thus, the conversational language processor 130 may generally engage in one or more cooperative conversations to determine the intent and resolve a particular goal for the multi-modal interactions received at the voice-enabled device 100. The conversational language processor 130 may then initiate one or more queries, tasks, commands, or other requests in furtherance of the intent and the goal determined for the multi-modal interactions. For example, in one implementation, the conversational language processor 130 may invoke one or more agents 132c having capabilities for processing requests in a particular domain or application 134, a voice search engine 132f having capabilities for retrieving information requested in the multi-modal interactions (e.g., from one or more data repositories 136, networks, or other information sources coupled to the voice-enabled device 100), or one or more other systems or multi-modal devices having suitable processing capabilities for furthering the intent and the goal for the multi-modal interactions (e.g., as determined from the constellation model 132b).

[046] Additionally, in one implementation, the conversational language processor 130 may invoke an advertising application 134 in relation to the queries, tasks, commands, or other requests initiated to process the multi-modal interactions, wherein the advertising application 134 may be configured to select one or more advertisements that may be relevant to the intent and/or the goal for the multi-modal interactions, as described in co-pending U.S. Patent Application Serial No. 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, the contents of which are hereby incorporated by reference in their entirety.

[047] In one implementation, in response to receiving results from any suitable combination of queries, tasks, commands, or other requests processed for the multi-modal interactions, the conversational language processor 130 may format the results

for presentation to the user through the output device 114. For example, the results may be formatted into a natural language utterance that can be converted into an electronic signal and provided to the user through a speaker coupled to the output device 114, or the results may be visually presented on a display coupled to the output device 114, or in any other suitable manner (e.g., the results may indicate whether a particular task or command was successfully performed, or the results may include information retrieved in response to one or more queries, or the results may include a request to frame a subsequent multi-modal interaction if the results are ambiguous or otherwise incomplete, etc.).

[048] Furthermore, in one implementation, the conversational language processor 130 may include a misrecognition engine 132e configured to determine whether the conversational language processor 130 incorrectly determined the intent for the multi-modal interactions. In one implementation, the misrecognition engine 132e may determine that the conversational language processor 130 incorrectly determined the intent in response to one or more subsequent multi-modal interactions provided proximately in time to the prior multi-modal interactions, as described in U.S. Patent Application Serial No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," which issued as U.S. Patent No. 7,620,549 on November 17, 2009, the contents of which are hereby incorporated by reference in their entirety. For example, the misrecognition engine 132e may monitor for one or more subsequent multi-modal interactions that include a stop word, override a current request, or otherwise indicate an unrecognized or misrecognized event. The misrecognition engine 132e may then determine one or more tuning parameters for various components associated with the ASR 120 and/or the conversational language processor 130 to improve subsequent interpretations.

[049] Accordingly, as described in further detail above, the voice-enabled device 100 may generally include various natural language processing components and capabilities that may be used for hybrid processing in the natural language voice services environment. In particular, the voice-enabled device 100 may be configured to determine the intent for various multi-modal interactions that include any suitable combination of natural language utterances and/or non-voice interactions and process one or more queries, tasks, commands, or other requests based on the determined intent. Furthermore, as noted above and as will be described in greater detail below,

one or more other systems and/or multi-modal devices may participate in determining the intent and processing the queries, tasks, commands, or other requests for the multi-modal interactions to provide a hybrid processing methodology, wherein the voice-enabled device 100 and the various other systems and multi-modal devices may each perform partial processing to determine the intent and otherwise process the multi-modal interactions in a cooperative manner. For example, in one implementation, hybrid processing in the natural language voice services environment may include one or more techniques described in U.S. Provisional Patent Application Serial No. 61/259,827, entitled "System and Method for Hybrid Processing in a Natural Language Voice Services Environment," filed on November 10, 2009, the contents of which are hereby incorporated by reference in their entirety.

[050] According to one aspect of the invention, Figure 2 illustrates a block diagram of an exemplary system for hybrid processing in a natural language voice service environment. In particular, the system illustrated in Figure 2 may generally include a voice-enabled client device 210 similar to the voice-enabled device described above in relation to Figure 1. For example, the voice-enabled client device 210 may include any suitable combination of input and output devices 215a respectively arranged to receive natural language multi-modal interactions and provide responses to the natural language multi-modal interactions. In addition, the voice-enabled client device 210 may include an Automatic Speech Recognizer (ASR) 220a configured to generate one or more preliminary interpretations of natural language utterances received at the input device 215a, and further configured to provide the preliminary interpretations to a conversational language processor 230a.

[051] In one implementation, the conversational language processor 230a on the voice-enabled client device 210 may include one or more natural language processing components, which may be invoked to determine an intent for the multi-modal interactions received at the voice-enabled client device 210. The conversational language processor 230a may then initiate one or more queries, tasks, commands, or other requests to resolve the determined intent. For example, the conversational language processor 230a may invoke one or more applications 234a to process requests in a particular domain, query one or more data repositories 236a to retrieve information requested in the multi-modal interactions, or otherwise engage in one or more cooperative conversations with a user of the voice-enabled client device 210 to

resolve the determined intent. Furthermore, as noted above in connection with Figure 1, the voice-enabled client device 210 may also cooperate with one or more other systems or multi-modal devices having suitable processing capabilities for initiating queries, tasks, commands, or other requests to resolve the intent of the multi-modal interactions.

[052] In particular, to cooperate with the other systems or multi-modal devices in the hybrid processing environment, the voice-enabled client device 210 may use a messaging interface 250a to communicative with a virtual router 260, wherein the messaging interface 250a may generally include a light client (or thin client) that provides a mechanism for the voice-enabled client device 210 to transmit input to and receive output from the virtual router 260. In addition, the virtual router 260 may further include a messaging interface 250b providing a mechanism for communicating with one or more additional voice-enabled devices 270a-n, one or more non-voice devices 280a-n, and a voice-enabled server 240. Furthermore, although Figure 2 illustrates messaging interface 250a and messaging interface 250b as components that are distinct from the devices to which they are communicatively coupled, it will be apparent that such illustration is for ease of description only, as the messaging interfaces 250a-b may be provided as on-board components that execute on the various devices illustrated in Figure 2 to facilitate communication among the various devices in the hybrid processing environment.

[053] For example, in one implementation, the messaging interface 250a that executes on the voice-enabled client device 210 may transmit input from the voice-enabled client device 210 to the virtual router 260 within one or more XML messages, wherein the input may include encoded audio corresponding to natural language utterances, preliminary interpretations of the natural language utterances, data corresponding to multi-touch gestures, point of focus or attention focus selections, and/or other multi-modal interactions. In one implementation, the virtual router 260 may then further process the input using a conversational language processor 230c having capabilities for speech recognition, intent determination, adaptive misrecognition, and/or other natural language processing. Furthermore, the conversational language processor 230c may include knowledge relating to content, services, applications, natural language processing capabilities, and other features available through the various devices in the hybrid processing environment.

[054] As such, in one implementation, the virtual router 260 may further communicate with the voice-enabled devices 270, the non-voice devices 280, and/or the voice-enabled server 240 through the messaging interface 250b to coordinate processing for the input received from the voice-enabled client device 210. For example, based on the knowledge relating to the features and capabilities of the various devices in the hybrid processing environment, the virtual router 260 may identify one or more of the devices that have suitable features and/or capabilities for resolving the intent of the input received from the voice-enabled client device 210. The virtual router 260 may then forward one or more components of the input to the identified devices through respective messaging interfaces 250b, wherein the identified devices may be invoked to perform any suitable processing for the components of the input forwarded from the virtual router 260. In one implementation, the identified devices may then return any results of the processing to the virtual router 260 through the respective messaging interfaces 250b, wherein the virtual router 260 may collate the results of the processing and return the results to the voice-enabled client device 210 through the messaging interface 250a.

[055] Accordingly, the virtual router 260 may communicate with any of the devices available in the hybrid processing environment through messaging interfaces 250a-b to coordinate cooperative hybrid processing for multi-modal interactions or other natural language inputs received from the voice-enabled client device 210. For example, in one implementation, the cooperative hybrid processing may be used to enhance performance in embedded processing architectures in which the voice-enabled client device 210 includes a constrained amount of resources (e.g., the voice-enabled client device 210 may be a mobile device having a limited amount of internal memory or other dedicated resources for natural language processing). As such, when the voice-enabled client device 210 has an embedded processing architecture, one or more components of the voice-enabled client device 210 may be configured to optimize efficiency of on-board natural language processing to reduce or eliminate bottlenecks, lengthy response times, or degradations in performance.

[056] For example, in one implementation, optimizing the efficiency of the on-board natural language processing may include configuring the ASR 220a to use a virtual dictation grammar having decoy words for out-of-vocabulary words, reduced vocabularies derived from a conversation history, or other dynamic recognition

grammars (e.g., grammars optimized for particular languages, contexts, domains, memory constraints, and/or other suitable criteria). In another example, the on-board applications 234a and/or data repositories 236a may be associated with an embedded application suite providing particular features and capabilities for the voice-enabled client device 210. For example, the voice-enabled client device 210 may be embedded within an automotive telematics system, a personal navigation device, a global positioning system, a mobile phone, or another device in which users often request location-based services. Thus, in such circumstances, the on-board applications 234a and the data repositories 236a in the embedded application suite may be optimized to provide certain location-based services that can be efficiently processed on-board (e.g., destination entry, navigation, map control, music search, hands-free dialing, etc.).

[057] Furthermore, although the components of the voice-enabled client device 210 may be optimized for efficiency in embedded architectures, a user may nonetheless request any suitable content, services, applications, and/or other features available in the hybrid processing environment, and the other devices in the hybrid processing environment may collectively provide natural language processing capabilities to supplement the embedded natural language processing capabilities for the voice-enabled client device 210. For example, the voice-enabled client device 210 may perform preliminary processing for a particular multi-modal interaction using the embedded natural language processing capabilities (e.g., the on-board ASR 220a may perform advanced virtual dictation to partially transcribe an utterance in the multi-modal interaction, the on-board conversational language processor 230a may determine a preliminary intent of the multi-modal interaction, etc.), wherein results of the preliminary processing may be provided to the virtual router 260 for further processing.

[058] In one implementation, the voice-enabled client device 210 may also communicate input corresponding to the multi-modal interaction to the virtual router 260 in response to determining that on-board capabilities cannot suitably interpret the interaction (e.g., if a confidence level for a partial transcription does not satisfy a particular threshold), or in response to determining that the interaction should be processed off-board (e.g., if a preliminary interpretation indicates that the interaction relates to a local search request requiring large computations to be performed on the voice-enabled server 240). As such, the virtual router 260 may capture the input

received from the voice-enabled client device 210 and coordinate further processing among the voice-enabled devices 270 and the voice-enabled server 240 that provide natural language processing capabilities in addition to the non-voice devices 280 that provide capabilities for retrieving data or executing tasks. Furthermore, in response to the virtual router 260 invoking one or more of the voice-enabled devices 270, the input provided to the voice-enabled devices 270 may be optimized to suit the processing requested from the invoked voice-enabled devices 270 (e.g., to avoid over-taxing processing resources, a particular voice-enabled device 270 may be provided a partial transcription or a preliminary interpretation and resolve the intent for a given context or domain).

[059] Alternatively, in response to the virtual router 260 invoking the voice-enabled server 240, the input provided to the voice-enabled devices 270 may further include encoded audio corresponding to natural language utterances and any other data associated with the multi-modal interaction. In particular, as shown in Figure 2, the voice-enabled server 240 may have a natural language processing architecture similar to the voice-enabled client device 210, except that the voice-enabled server 240 may include substantial processing resources that obviate constraints that the voice-enabled client device 210 may be subject to. Thus, when the voice-enabled server 240 cooperates in the hybrid processing for the multi-modal interaction, the encoded audio corresponding to the natural language utterances and the other data associated with the multi-modal interaction may be provided to the voice-enabled server 240 to maximize a likelihood of the voice-enabled server 240 correctly determining the intent of the multi-modal interaction (e.g., the ASR 220b may perform multi-pass speech recognition to generate an accurate transcription for the natural language utterance, the conversational language processor 230b may arbitrate among intent determinations performed in any number of different contexts or domains, etc.). Accordingly, in summary, the hybrid processing techniques performed in the environment illustrated in Figure 2 may generally include various different devices, which may or may not include natural language capabilities, cooperatively determining the intent of a particular multi-modal interaction and taking action to resolve the intent.

[060] Although the cooperative hybrid processing techniques described above have been particularly described in the context of an embedded processing architecture, such techniques are not necessarily limited to embedded processing architectures. In

particular, the same techniques may be applied in any suitable voice services environment having various devices that can cooperate to initiate queries, tasks, commands, or other requests to resolve the intent of multi-modal interactions. Furthermore, in one implementation, the voice-enabled client device 210 may include a suitable amount of memory or other resources that can be dedicated to natural language processing (e.g., the voice-enabled client device 210 may be a desktop computer or other device that can process natural language without substantially degraded performance). In such circumstances, one or more of the components of the voice-enabled client device 210 may be configured to optimize the on-board natural language processing in a manner that could otherwise cause bottlenecks, lengthy response times, or degradations in performance in an embedded architecture. For example, in one implementation, optimizing the on-board natural language processing may include configuring the ASR 220a to use a large list dictation grammar in addition to and/or instead of the virtual dictation grammar used in embedded processing architectures.

[061] Nonetheless, as will be described in greater detail below in connection with Figures 3-5, the cooperative hybrid processing techniques may be substantially similar regardless of whether the voice-enabled client device 210 has an embedded or non-embedded architecture. In particular, regardless of the architecture for the voice-enabled client device 210, cooperative hybrid processing may include the voice-enabled client device 210 optionally performing preliminary processing for a natural language multi-modal interaction and communicating input corresponding to the multi-modal interaction to the virtual router 260 for further processing through the messaging interface 250a. Alternatively (or additionally), the cooperative hybrid processing may include the virtual router 260 coordinating the further processing for the input among the various devices in the hybrid environment through messaging interface 250b, and subsequently returning any results of the processing to the voice-enabled client device 210 through messaging interface 250a.

[062] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for initializing various devices that cooperate to perform hybrid processing in a natural language voice services environment. In particular, as noted above, the hybrid processing environment may generally include communication among various different devices that may cooperatively process natural language

multi-modal interactions. For example, in one implementation, the various devices in the hybrid processing environment may include a virtual router having one or more messaging interfaces for communicating with one or more voice-enabled devices, one or more non-voice devices, and/or a voice-enabled server. As such, in one implementation, the method illustrated in Figure 3 may be used to initialize communication in the hybrid processing environment to enable subsequent cooperative processing for one or more natural language multi-modal interactions received at any particular device in the hybrid processing environment.

[063] In one implementation, the various devices in the hybrid processing environment may be configured to continually listen or otherwise monitor respective input devices to determine whether a natural language multi-modal interaction has occurred. As such, the method illustrated in Figure 3 may be used to calibrate, synchronize, or otherwise initialize the various devices that continually listen for the natural language multi-modal interactions. For example, as described above in connection with Figure 2, the virtual router, the voice-enabled devices, the non-voice devices, the voice-enabled server, and/or other devices in the hybrid processing environment may be configured to provide various different capabilities or services, wherein the initialization method illustrated in Figure 3 may be used to ensure that the hybrid processing environment obtains a suitable signal to process any particular natural language multi-modal interaction and appropriately invoke one or more of the devices to cooperatively process the natural language multi-modal interaction. Furthermore, the method illustrated in Figure 3 and described herein may be invoked to register the various devices in the hybrid processing environment, register new devices added to the hybrid processing environment, publish domains, services, intent determination capabilities, and/or other features supported on the registered devices, synchronize local timing for the registered devices, and/or initialize any other suitable aspect of the devices in the hybrid processing environment.

[064] In one implementation, initializing the various devices in the hybrid processing environment may include an operation 310, wherein a device listener may be established for each of the devices in the hybrid processing environment. The device listeners established in operation 310 may generally include any suitable combination of instructions, firmware, or other routines that can be executed on the various devices to determine capabilities, features, supported domains, or other information associated

with the devices. For example, in one implementation, the device listeners established in operation 310 may be configured to communicate with the respective devices using the Universal Plug and Play protocol designed for ancillary computer devices, although it will be apparent that any appropriate mechanism for communicating with the various devices may be suitably substituted.

[065] In response to establishing the device listeners for each device registered in the hybrid processing environment (or in response to establishing device listeners for any device newly registered in the hybrid processing environment), the device listeners may then be synchronized in an operation 320. In particular, each of the registered devices may have an internal clock or other timing mechanism that indicates local timing for an incoming natural language multi-modal interaction, wherein operation 320 may be used to synchronize the device listeners established in operation 310 according to the internal clocks or timing mechanisms for the respective devices. Thus, in one implementation, synchronizing the device listeners in operation 320 may include each device listener publishing information relating to the internal clock or local timing for the respective device. For example, the device listeners may publish the information relating to the internal clock or local timing to the virtual router, whereby the virtual router may subsequently coordinate cooperative hybrid processing for natural language multi-modal interactions received at one or more of the devices in the hybrid processing environment. It will be apparent, however, that the information relating to the internal clock or local timing for the various devices in the hybrid processing environment may be further published to the other voice-enabled devices, the other non-voice devices, the voice-enabled server, and/or any other suitable device that may participate in cooperative processing for natural language multi-modal interactions provided to the hybrid processing environment.

[066] In one implementation, in response to establishing and synchronizing the device listeners for the various devices registered in the hybrid processing environment, the device listeners may continually listen or otherwise monitor respective devices on the respective registered devices in an operation 330 to detect information relating to one or more natural language multi-modal interactions. For example, the device listeners may be configured to detect occurrences of the natural language multi-modal interactions in response to detecting an incoming natural language utterance, a point of focus or attention focus selection associated with an incoming natural language

utterance, and/or another interaction or sequence of interactions that relates to an incoming natural language multi-modal interaction. In addition, operation 330 may further include the appropriate device listeners capturing the natural language utterance and/or related non-voice device interactions that relate to the natural language utterance.

[067] In one implementation, the captured natural language utterance and related non-voice device interactions may then be analyzed in an operation 340 to manage subsequent cooperative processing in the hybrid processing environment. In one implementation, for example, operation 340 may determine whether one device listener or multiple device listeners captured information relating to the natural language multi-modal interaction detected in operation 330. In particular, as noted above, the hybrid processing environment may generally include various different devices that cooperate to process natural language multi-modal interactions, whereby the information relating to the natural language multi-modal interaction may be provided to one or a plurality of the devices in the hybrid processing environment. As such, operation 340 may determine whether one device listener or multiple device listeners captured the information relating to the natural language multi-modal interaction in order to determine whether the hybrid processing environment needs to synchronize signals among various device listeners that captured information relating to the multi-modal interaction.

[068] For example, a user interacting with the hybrid processing environment may view a web page presented on a non-voice display device and provide a natural language multi-modal interaction that requests more information about purchasing a product displayed on the web page. The user may then select text on the web page containing the product name using a mouse, keyboard, or other non-voice input device and provide a natural language utterance to a microphone or other voice-enabled device such as “Is this available on Amazon.com?” In this example, a device listener associated with the non-voice display device may detect the text selection for the product name in operation 330, and a device listener associated with the voice-enabled device may further detect the natural language utterance inquiring about the availability of the product in operation 330. Furthermore, in one implementation, the user may be within a suitable range of multiple voice-enabled devices, which may result in multiple device listeners capturing different signals corresponding to the

natural language utterance (e.g., the interaction may occur within range of a voice-enabled mobile phone, a voice-enabled telematics device, and/or other voice-enabled devices, depending on the arrangement and configuration of the various devices in the hybrid processing environment).

[069] Accordingly, as will be described in greater detail herein, a sequence of operations that synchronizes different signals relating to the multi-modal interaction received at the multiple device listeners may be initiated in response to operation 340 determining that multiple device listeners captured information relating to the natural language multi-modal interaction. On the other hand, in response to operation 340 determining that only one device listener captured information relating to the natural language multi-modal interaction, the natural language multi-modal interaction may be processed in an operation 390 without executing the sequence of operations that synchronizes different signals (i.e., the one device listener provides all of the input information relating to the multi-modal interaction, such that hybrid processing for the interaction may be initiated in operation 390 without synchronizing different input signals). However, in one implementation, the sequence of synchronization operations may also be initiated in response to one device listener capturing a natural language utterance and one or more non-voice interactions to align different signals relating to the natural language multi-modal interaction, as described in greater detail herein.

[070] As described above, each device listener that receives an input relating to the natural language multi-modal interaction detected in operation 330 may have an internal clock or other local timing mechanism. As such, in response to determining that one or more device listeners captured different signals relating to the natural language multi-modal interaction in operation 340, the sequence of synchronization operations for the different signals may be initiated in an operation 350. In particular, operation 350 may include the one or more device listeners determining local timing information for the respective signals based on the internal clock or other local timing mechanism associated with the respective device listeners, wherein the local timing information determined for the respective signals may then be synchronized.

[071] For example, in one implementation, synchronizing the local timing information for the respective signals may be initiated in an operation 360. In particular, operation 360 may generally include notifying each device listener that received an input relating to the multi-modal interaction of the local timing information determined for each

respective signal. For example, in one implementation, each device listener may provide local timing information for a respective signal to the virtual router, and the virtual router may then provide the local timing information for all of the signals to each device listener. As such, in one implementation, operation 360 may result in each device listener receiving a notification that includes local timing information for each of the different signals that relate to the natural language multi-modal interaction detected in operation 330. Alternatively (or additionally), the virtual router may collect the local timing information for each of the different signals from each of the device listeners and further synchronize the local timing information for the different signals to enable hybrid processing for the natural language multi-modal interaction.

[072] In one implementation, any particular natural language multi-modal interaction may include at least a natural language utterance, and may further include one or more additional device interactions relating to the natural language utterance. As noted above, the utterance may generally be received prior to, contemporaneously with, or subsequent to the additional device interactions. As such, the local timing information for the different signals may be synchronized in an operation 370 to enable hybrid processing for the natural language multi-modal interaction. In particular, operation 370 may include aligning the local timing information for one or more signals corresponding to the natural language utterance and/or one or more signals corresponding to any additional device interactions that relate to the natural language utterance. In addition, operation 370 may further include aligning the local timing information for the natural language utterance signals with the signals corresponding to the additional device interactions.

[073] Thus, in matching the utterance signals and the non-voice device interaction signals, any devices that participate in the hybrid processing for the natural language multi-modal interaction may be provided with voice components and/or non-voice components that have been aligned with one another. For example, in one implementation, operation 370 may be executed on the virtual router, which may then provide the aligned timing information to any other device that may be invoked in the hybrid processing. Alternatively (or additionally), one or more of the other devices that participate in the hybrid processing may locally align the timing information (e.g., in response to the virtual router invoking the voice-enabled server in the hybrid processing, resources associated with the voice-enabled server may be employed to

align the timing information and preserve communication bandwidth at the virtual router).

[074] Furthermore, in one implementation, the virtual router and/or other devices in the hybrid processing environment may analyze the signals corresponding to the natural language utterance in an operation 380 to select the cleanest sample for further processing. In particular, as noted above, the virtual router may include a messaging interface for receiving an encoded audio sample corresponding to the natural language utterance from one or more of the voice-enabled devices. For example, the audio sample received at the virtual router may include the natural language utterance encoded in the MPEG-1 Audio Layer 3 (MP3) format or another lossy format to preserve communication bandwidth in the hybrid processing environment. However, it will be apparent that the audio sample may alternatively (or additionally) be encoded using the Free Lossless Audio Codec (FLAC) format or another lossless format in response to the hybrid processing environment having sufficient communication bandwidth for processing lossless audio that may provide a better sample of the natural language utterance.

[075] Regardless of whether the audio sample has been encoded in a lossy or lossless format, the signal corresponding to the natural language utterance that provides the cleanest sample may be selected in operation 380. For example, one voice-enabled device may be in a noisy environment or otherwise associated with conditions that interfere with generating a clean audio sample, while another voice-enabled device may include a microphone array or be configured to employ techniques that maximize fidelity of encoded speech. As such, in response to multiple signals corresponding to the natural language utterance being received in operation 330, the cleanest signal may be selected in operation 380 and hybrid processing for the natural language utterance may then be initiated in an operation 390.

[076] Accordingly, the synchronization and initialization techniques illustrated in Figure 3 and described herein may ensure that the hybrid processing environment synchronizes each of the signals corresponding to the natural language multi-modal interaction and generates an input for further processing in operation 390 most likely to result in a correct intent determination. Furthermore, in synchronizing the signals and selecting the cleanest audio sample for the further processing in operation 390, the techniques illustrated in Figure 3 and described herein may ensure that none of the

devices in the hybrid processing environment take action on a natural language multi-modal interaction until the appropriate signals to be used in operation 390 have been identified. As such, hybrid processing for the natural language multi-modal interaction may be initiated in operation 390, as described in further detail herein.

[077] According to one aspect of the invention, Figure 4 illustrates a flow diagram of am exemplary method for performing hybrid processing at one or more client devices in a natural language voice services environment. In particular, as will be described in greater below with reference to Figure 5, the one or more client devices may perform the hybrid processing in cooperation with a virtual router through a messaging interface that communicatively couples the client devices and the virtual router. For example, in one implementation, the messaging interface may generally include a light client (or thin client) that provides a mechanism for the client devices to transmit input relating to a natural language multi-modal interaction to the virtual router, and that further provides a mechanism for the client devices to receive output relating to the natural language multi-modal interaction from the virtual router.

[078] For example, in one implementation, the hybrid processing at the client devices may be initiated in response to one or more of the client devices receiving a natural language multi-modal interaction in an operation 410. In particular, the natural language multi-modal interaction may generally include a natural language utterance received at a microphone or other voice-enabled input device coupled to the client device that received the natural language multi-modal interaction, and may further include one or more other additional input modalities that relate to the natural language utterance (e.g., text selections, button presses, multi-touch gestures, etc.). As such, the natural language multi-modal interaction received in operation 410 may include one or more queries, commands, or other requests provided to the client device, wherein the hybrid processing for the natural language multi-modal interaction may then be initiated in an operation 420.

[079] As described in greater detail above, the natural language voice services environment may generally include one or more voice-enabled client devices, one or more non-voice devices, a voice-enabled server, and a virtual router arranged to communicate with each of the voice-enabled client devices, the non-voice devices, and the voice-enabled server. In one implementation, the virtual router may therefore coordinate the hybrid processing for the natural language multi-modal interaction

among the voice-enabled client devices, the non-voice devices, and the voice-enabled server. As such, the hybrid processing techniques described herein may generally refer to the virtual router coordinating cooperative processing for the natural language multi-modal interaction in a manner that involves resolving an intent of the natural language multi-modal interaction in multiple stages.

[080] In particular, as described above in connection with Figure 3, the various devices that cooperate to perform the hybrid processing may be initialized to enable the cooperative processing for the natural language multi-modal interaction. As such, in one implementation, in response to initializing the various devices, each of the client devices that received an input relating to the natural language multi-modal interaction may perform initial processing for the respective input in an operation 420. For example, in one implementation, a client device that received the natural language utterance included in the multi-modal interaction may perform initial processing in operation 420 that includes encoding an audio sample corresponding to the utterance, partially or completely transcribing the utterance, determining a preliminary intent for the utterance, or performing any other suitable preliminary processing for the utterance. In addition, the initial processing in operation 420 may also be performed at a client device that received one or more of the additional input modalities relating to the utterance. For example, the initial processing performed in operation 420 for the additional input modalities may include identifying selected text, selected points of focus or attention focus, or generating any other suitable data that can be used to further interpret the utterance. In one implementation, an operation 430 may then include determining whether the hybrid processing environment has been configured to automatically route inputs relating to the natural language multi-modal interaction to the virtual router.

[081] For example, in one implementation, operation 430 may determine that automatic routing has been configured to occur in response to multiple client devices receiving the natural language utterance included in the multi-modal interaction in operation 410. In this example, the initial processing performed in operation 420 may include the multiple client devices encoding respective audio samples corresponding to the utterance, wherein messages that include the encoded audio samples may then be sent to the virtual router in an operation 460. The virtual router may then select one of the encoded audio samples that provides a cleanest signal and coordinate subsequent

hybrid processing for the natural language multi-modal interaction, as will be described in greater detail below with reference to Figure 5. In another example, operation 430 may determine that automatic routing has been configured to occur in response to the initial processing resulting in a determination that the multi-modal interaction relates to a request that may be best suited for processing on the voice-enabled server (e.g., the request may relate to a location-based search query or another command or task that requires resources managed on the voice-enabled server, content, applications, domains, or other information that resides on one or more devices other than the client device that received the request, etc.). However, it will be apparent that the hybrid processing environment may be configured for automatic routing in response to other conditions and/or regardless of whether any attendant conditions exist, as appropriate.

[082] In one implementation, in response to the virtual router coordinating the hybrid processing for the natural language multi-modal interaction, the virtual router may provide results of the hybrid processing to the client device in an operation 470. For example, the results provided to the client device in operation 470 may include a final intent determination for the natural language multi-modal interaction, information requested in the interaction, data generated in response to executing a command or task requested in the interaction, and/or other results that enable the client device to complete processing for the natural language request in operation 480. For example, in one implementation, operation 480 may include the client device executing a query, command, task, or other request based on the final intent determination returned from the virtual router, presenting the requested information returned from the virtual router, confirming that the requested command or task has been executed, and/or performing any additional processing to resolve the natural language request.

[083] Referring back to operation 430, in response to determining that the conditions that trigger automatic routing have not been satisfied or that automatic router has otherwise not been configured, the client device may further process the natural language multi-modal interaction in an operation 440. In one implementation, the further processing in operation 440 may include the client device attempting to determine an intent for the natural language multi-modal interaction using local natural language processing capabilities. For example, the client device may merge any non-voice input modalities included in the multi-modal interaction a transcription for the utterance included in the multi-modal interaction. The conversational language

processor on the client device may then determine the intent for the multi-modal interaction utilizing local information relating to context, domains, shared knowledge, criteria values, or other information. The client device may then generate one or more interpretations for the utterance to determine the intent for the multi-modal interaction (e.g., identifying a conversation type, one or more requests contained in the interactions, etc.).

[084] In one implementation, operation 440 may further include determining a confidence level for the intent determination generated on the client device (e.g., the confidence level may be derived in response to whether the client devices includes a multi-pass speech recognition engine, whether the utterance contained any ambiguous words or phrases, whether the intent differs from one context to another, etc.). In one implementation, an operation 450 may then determine whether or not to invoke off-board processing depending on the confidence level determined in operation 440. For example, operation 450 may generally include determining whether the intent determined in operation 440 satisfies a particular threshold value that indicates an acceptable confidence level for taking action on the determined intent. As such, in response to the confidence level for the intent determination satisfying the threshold value, operation 450 may determine to not invoke off-board processing. In particular, the confidence level satisfying the threshold value may indicate that the client device has sufficient information to take action on the determined intent, whereby the client device may then process one or more queries, commands, tasks, or other requests to resolve the multi-modal interaction in operation 480.

[085] Alternatively, in response to the confidence level for the intent determination failing to satisfy the threshold value, operation 450 may invoke off-board processing, which may include sending one or more messages to the virtual router in operation 460. The one or more messages may cause the virtual router to invoke additional hybrid processing for the multi-modal interaction in a similar manner as noted above, and as will be described in greater detail herein with reference to Figure 5.

[086] According to one aspect of the invention, Figure 5 illustrates a flow diagram of am exemplary method for performing hybrid processing at a virtual router in a natural language voice services environment. In particular, as the virtual router may coordinate the hybrid processing for natural language multi-modal interactions received at one or more client devices. In one implementation, in an operation 510,

the virtual router may receive one or more messages relating to a natural language multi-modal interaction received at one or more of the client devices in the voice services environment. For example, the virtual router may include a messaging interface that communicatively couples the virtual router to the client devices and a voice-enabled server, wherein the messaging interface may generally include a light client (or thin client) that provides a mechanism for the virtual router to receive input from one or more the client devices and/or the voice-enabled server, and further to transmit output to one or more the client devices and/or the voice-enabled server. The messages received in operation 510 may generally include any suitable processing results for the multi-modal interactions, whereby the virtual router may coordinate hybrid processing in a manner that includes multiple processing stages that may occur at the virtual router, one or more of the client devices, the voice-enabled server, or any suitable combination thereof.

[087] In one implementation, the virtual router may analyze the messages received in operation 510 to determine whether to invoke the hybrid processing in a peer-to-peer mode. For example, one or more of the messages may include a preliminary intent determination that the virtual router can use to determine whether to invoke one or more of the client devices, the voice-enabled server, or various combinations thereof in order to execute one or more of the multiple processing stages for the multi-modal interaction. In another example, one or more of the messages may include an encoded audio sample that the virtual router forwards to one or more of the various devices in the hybrid processing environment. As such, in one implementation, the virtual router may analyze the messages received in operation 510 to determine whether or not to invoke the voice-enabled server to process the multi-modal interaction (e.g., the messages may include a preliminary intent determination that indicates that the multi-modal interaction includes a location-based request that requires resources residing on the server).

[088] In response to the virtual router determining to invoke the voice-enabled server, the virtual router may forward the messages to the server in an operation 530. In particular, the messages forwarded to the server may generally include the encoded audio corresponding to the natural language utterance and any additional information relating to other input modalities relevant to the utterance. For example, as described in greater detail above with reference to Figure 2, the voice-enabled server may

include various natural language processing components that can suitably determine the intent of the multi-modal interaction, whereby the messages sent to the voice-enabled server may include the encoded audio in order to permit the voice-enabled server to determine the intent independently of any preliminary processing on the client devices that may be inaccurate or incomplete. In response to the voice-enabled server processing the messages received from the virtual router, results of the processing may then be returned to the virtual router in an operation 570. For example, the results may include the intent determination for the natural language multi-modal interaction, results of any queries, commands, tasks, or other requests performed in response to the determined intent, or any other suitable results, as will be apparent.

[089] Alternatively, in response to the virtual router determining to invoke the peer-to-peer mode in operation 520, the virtual router may coordinate the hybrid processing among one or more the client devices, the voice-enabled server, or any suitable combination thereof. For example, in one implementation, the virtual router may determine a context for the natural language multi-modal interaction in an operation 540 and select one or more peer devices based on the determined context in an operation 550. For example, one or more of the client devices may be configured to provide content and/or services in the determined context, whereby the virtual router may forward one or more messages to such devices in an operation 560 in order to request such content and/or services. In another example, the multi-modal interaction may include a compound request that relates to multiple contexts supported on different devices, whereby the virtual router may forward messages to each such device in operation 560 in order to request appropriate content and/or services in the different contexts.

[090] In still another example, the interaction may include a request to be processed on the voice-enabled server, yet the request may require content and/or services that reside on one or more of the client devices (e.g., a location-based query relating to an entry in an address book on one or more of the client devices). As such, the virtual router may generally forward various messages to the selected peer devices in operation 560 to manage the multiple stages in the hybrid processing techniques described herein. For example, the virtual router may send messages to one or more voice-enabled client devices that have intent determination capabilities in a particular context, one or more non-voice client devices that have access to content, services,

and/or other resources needed to process the multi-modal interaction, or any appropriate combination thereof. The virtual router may therefore send messages to the client devices and/or the voice-enabled server in operation 560 and receive responsive messages from the client devices and/or the voice-enabled server in operation 570 in any appropriate manner (e.g., in parallel, sequentially, iteratively, etc.). The virtual router may then collate the results received in the responsive messages in operation 580 and return the results to one or more of the client devices for any final processing and/or presentation of the results.

[091] According to one aspect of the invention, Figure 6 illustrates a block diagram of an exemplary system for providing a natural language content dedication service. In particular, as shown in Figure 6, the system for providing the natural language content dedication service may generally include a voice-enabled client device 610 that can communicate with a content dedication system 665 through a messaging interface 665. In one implementation, the voice-enabled client device 610 may communicate with the content dedication system 665 in a similar manner as described above with reference to Figures 2 through 5. For example, the content dedication system 665 may include a virtual router 660 and a voice-enabled server 640 that can service multi-modal natural language requests provided to the voice-enabled client device 610 in a similar manner as described above, and may further include a billing system 638 that may be used to process transactions relating to the content dedication service. Further, although Figure 6 illustrates the content dedication system 665 as including the virtual router 660, the voice-enabled server 640, and the billing system 638 within one component, such illustration will be understood to be for ease of description only, and that the virtual router 660, the voice-enabled server 640, and the billing system 638 may in fact be arranged within any number of components that can suitably communicate with one another to process multi-modal natural language requests relating to the content dedication service (e.g., the virtual router 660 may communicate with the voice-enabled server 640 through another messaging interface distinct from the messaging interface 650 for communicating with the voice-enabled client device 610, as shown in the exemplary system illustrated in Figure 2).

[092] In one implementation, the system shown in Figure 6 may provide the natural language content dedication service to any suitable voice-enabled client device 610 having a suitable combination of input and output devices 615a that can receive

natural language multi-modal interactions and provide responses to the natural language multi-modal interactions, wherein the input and output devices 615a may be further arranged to receive any other suitable type of input and provide any other suitable type of output. For example, in one implementation, the voice-enabled client device 610 may comprise a mobile phone that includes a keypad input device 615a, a touch screen input device 615a, or other input mechanisms 615a in addition to any input microphones or other suitable input devices 615a that can receive voice signals. As such, in this example, the mobile phone may further include an output display device 615a in addition to any output microphones or other suitable output devices 615a that can output audible signals. Thus, a user of the voice-enabled client device 610 may be listening to music, watching video, or otherwise interacting with content through the input and output devices 615a and provide a multi-modal natural language request to engage in a transaction to dedicate the music, video, or other content, as will be described in greater detail below. Furthermore, in one implementation, the voice-enabled client device 610 may be included within a hybrid processing environment that may include a plurality of different devices, whereby the content dedication request may relate to content played on a different device from the voice-enabled client device 610, although the content dedication request may relate to any suitable content (i.e., the request need not necessarily relate to played content, as users may provide natural language to request content dedications for any suitable content).

[093] Thus, in one implementation, in response to the voice-enabled client device 710 receiving a multi-modal interaction that includes a natural language utterance, the voice-enabled client device 610 may invoke an Automatic Speech Recognizer (ASR) 620a to generate one or more preliminary interpretations of the utterance. The ASR 620a may then provide the preliminary interpretations to a conversational language processor 630a, which may attempt to determine an intent for the multi-modal interaction. In one implementation, to determine the intent for the multi-modal interaction, the conversational language processor 630a may determine a most likely context for the interaction from the preliminary interpretations of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared knowledge, or any other suitable information for interpreting the multi-modal

interaction. Thus, in response to the conversational language processor 630a determining that the intent for the multi-modal interaction relates to a content dedication request, a content dedication application 634a may be invoked to resolve the content dedication request.

[094] For example, in one implementation, an initial multi-modal interaction may include the utterance "Find 'Superstylin'' by Groove Armada." In response to the initial multi-modal interaction, the ASR 620a may generate a preliminary interpretation that includes the words "Find" and "Superstylin'" and the phrase "Groove Armada." The ASR 620a may then provide the preliminary interpretation to the conversational language processor 630a, which may determine that the word "Find" indicates that the most likely intent of the interaction includes a search, while the word "Superstylin'" and the phrase "Groove Armada" provide criteria for the search. Furthermore, in response to determining the most likely intent of the interaction, the conversational language processor 630a may establish a music context for the interaction and attempt to resolve the search request. For example, the conversational language processor 630a may search one or more data repositories 636a that contain music information to identify music having a song title of "Superstylin'" and an artist name of "Groove Armada," and the conversational language processor 630a may further cooperate with other devices in the hybrid processing environment to search for the song (e.g., in response to the local data repositories 636a not yielding adequate results, another device in the environment having a larger music data repository than the client device 610 may be invoked). The conversational language processor 630a may then receive appropriate results for the search and present the results to the user through the output device 615a (e.g., displaying information about the song, playing a sample audio clip of the song, displaying options to purchase the song, recommending similar songs or similar artists, etc.).

[095] Continuing with the above example, a subsequent multi-modal interaction may include the utterance "Share this with my wife," "Dedicate it to Charlene," "That's the one, I want to pass along to some friends," or another suitable utterance that reflects a request to dedicate the content. Alternatively (or additionally), in response to the intent of the initial interaction including a request to search for content, the conversational language processor 630a may invoke the content dedication application 634a, which may present an option to dedicate the content through the output device 615a together

with the results of the search. As such, the request to dedicate the content may also be provided in a non-speech input, such as a button press or touch screen selection of the option to dedicate the content. Accordingly, in response to detecting a content dedication request (e.g., through the ASR 620a and the conversational language processor 630a processing a suitable utterance, the input device 615a receiving a suitable non-speech input, or any suitable combination thereof), the content dedication application 634a may be invoked to process the content dedication request.

[096] In one implementation, to process the content dedication request, the content dedication application 634a may capture a natural language utterance that contains the dedication. For example, the content dedication application 634a may provide a prompt through the output device 615a that instructs the user to provide the dedication utterance (e.g., a visual or audible prompt instructing the user to begin speaking, to speak after an audible beep, etc.). The user may then provide the dedication utterance through the voice-enabled input device 615a, and the dedication utterance may then be converted into an electronic signal that the content dedication application 634a captures for the dedication (e.g., "Dear Charlene, I was listening to this song and I thought of you. Enjoy!"). In addition, the content dedication application 634a may further prompt the user to provide any additional tags for the dedicated content (e.g., an image or a picture to be inserted as album art for the dedicated content, a natural language utterance that includes information to be inserted as voice-tags for the dedicated content, a non-speech input or data input that includes information to insert in tags for the dedicated content, etc.). The content dedication application 634a may then prompt the user to identify a recipient 690 of the dedication, wherein the user may provide any suitable multi-modal input that includes an e-mail address, a telephone number, an address book entry, or other information identifying the recipient 690 of the dedication.

[097] In one implementation, the content dedication application 634a may then route the request, including the dedication utterance, the additional tags (if any), and the information identifying the recipient 690 to the content dedication system 665 through the messaging interface 650. For example, the dedication utterance may be converted into encoded audio that can be communicated through the messaging interface 650, whereby the content dedication system 665 can insert the encoded audio corresponding to the dedication utterance within the dedicated content and/or verbally

annotate the dedicated content with the encoded audio. Alternatively (or additionally), the dedication utterance may be interpreted and parsed to transcribe one or more words or phrases from the dedication utterance, wherein the transcribed words or phrases may provide a textual annotation for the dedicated content (e.g., the textual annotation may be inserted within metadata for the dedicated content, such as an ID3 Comments tag). Similarly, any utterances to be inserted as voice-tags for the dedicated content may provide further verbal annotations for the dedicated content, or such utterances may be transcribed to provide further textual annotations for the dedicated content. In one implementation, verbal annotations, textual annotations, and other types of annotations may be created and associated with the dedicated content using techniques described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety.

[098] In one implementation, in response to the content dedication system 665 receiving the content dedication request, the dedication utterance, and any additional tags from the content dedication application 634a, the content dedication system 665 may invoke a similar content dedication application 634b on the voice-enabled server 640 to process the request. In particular, the content dedication application 634b on the voice-enabled server 640 may identify the content requested for dedication and initiate a transaction for the content requested for dedication. For example, in one implementation, if the content dedication application 634a on the voice-enabled client device 610 was able to identify the requested content, the content dedication request received at the content dedication system 665 may include an identification of the content to be dedicated. Alternatively, because the content dedication system 665 can cooperate in resolving the multi-modal interactions involved in the dedication request, the content dedication system 665 may use shared knowledge relating to the dedication request to identify the content to be dedicated. For example, the content dedication system 665 may invoke one or more local natural language processing components (e.g., ASR 620b, conversational language processor 630b or 630c, etc.), search one or more local data repositories 636b, interact with one or more content providers 680 over a network 670, pull data from a satellite radio system that played the requested content at the voice-enabled client device 610 or another device in the

hybrid processing environment, or otherwise consult available resources that can be used to identify the content to be dedicated.

[099] In one implementation, in response to identifying the content to be dedicated, the content dedication system 665 may communicate with a billing system 638 to identify one or more purchase options for the dedication request and process an appropriate transaction for the dedication request. In particular, the content dedication system 665 may generally support various different purchase options to provide users with flexibility in the manner of requesting content dedications, including a buy-to-own purchase option, a pay-to-play purchase option, a paid subscription purchase option, or other appropriate options. In one implementation, the purchase options may be modeled on techniques for providing natural language services and subscriptions described in U.S. Patent Application Serial No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,398,209 on July 8, 2008, and co-pending U.S. Patent Application Serial No. 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the contents of which are hereby incorporated by reference in their entirety. For example, in the buy-to-own purchase option, the content dedication system 665 may purchase full rights to the content from an appropriate content provider 680, wherein the billing system 638 may then charge the user of the voice-enabled client device 610 a particular amount that encompasses the cost for purchasing the content from the content provider 680 plus a service charge for dedicating the content, tagging the dedicated content, delivering the dedicated content to the recipient 690, or any other appropriate services rendered for the content dedication request. The billing system 638 may charge the user in a similar manner under the pay-to-play purchase option, except that the rights purchased from the content provider 680 may be limited to a single play, such that the cost for purchasing the content from the content provider 680 may be somewhat less under the pay-to-play purchase option.

[0100] Under the paid subscription purchase option, however, the user may pay a periodic service charge to the content dedication system 665 that permits the user to make a predetermined number of content dedications or an unlimited number of content dedications in a subscription period, or the subscription purchase option may permit the user to make content dedications in any other suitable manner (e.g.,

different subscription levels having different content dedication options may be offered, such that the user may select a subscription level that meets the user's particular needs). For example, under the paid subscription purchase option, the billing system 638 may only charge the user any costs for obtaining the rights to the content, which may be purchased under either the buy-to-own option or the pay-to-play option, or alternatively the user may be charged nothing if the user already owns the dedication content. In another example, a first subscription level may cost a first amount to permit the user to make a particular number of content dedications in a subscription period, while a second subscription level may cost a higher amount to permit the user to make an unlimited number of content dedications in the subscription period, while still other subscription levels having different terms may be offered, as will be apparent.

[0101] Furthermore, in one implementation, a service provider associated with the content dedication system 665 may negotiate an agreement with the content provider 680 to determine the manner in which revenues for content transactions will be shared between the content dedication system 665 and the content provider 680. For example, the agreement may provide that the content provider 680 may keep all of the revenue for transactions that include purchasing content from the content provider 680 and that the service provider associated with the content dedication system 665 may recoup costs for such transactions from users. In another example, the agreement may provide that the content provider 680 and the service provider associated with the content dedication system 665 may share the revenue for the transactions that include purchasing content from the content provider 680. Thus, the agreement may generally include any suitable arrangement that defines the manner in which the content provider 680 and the service provider associated with the content dedication system 665 manage the revenue for the transactions that include purchasing content from the content provider 680, while the service provider associated with the content dedication system 665 may manage billing users for the natural language aspects of the content dedication service according to the techniques described in further detail above.

[0102] According to one aspect of the invention, Figure 7 illustrates a flow diagram of an exemplary method for providing a natural language content dedication service. In particular, the method for providing a natural language content dedication service, as shown in Figure 7, may be performed on a voice-enabled client device that can communicate with a content dedication system through a messaging interface, wherein

the voice-enabled client device may communicate with the content dedication system in a similar manner as described above with reference to Figures 2 through 6. For example, the content dedication system may include a virtual router and a voice-enabled server that can service multi-modal natural language requests provided to the voice-enabled client device, and may further include a billing system for processing transactions relating to the content dedication service.

[0103] In one implementation, the method shown in Figure 7 may be used to provide the natural language content dedication service to any suitable voice-enabled client device having a suitable combination of input and output devices that can receive natural language multi-modal interactions and provide responses to the natural language multi-modal interactions, wherein the input and output devices may be further arranged to receive any other suitable type of input and provide any other suitable type of output. For example, in one implementation, the voice-enabled client device may comprise a mobile phone that includes a keypad input device, a touch screen input device, or other input mechanisms in addition to any input microphones or other suitable input devices that can receive voice signals, and may further include an output display device in addition to any output microphones or other suitable output devices that can output audible signals. Thus, a user of the voice-enabled client device may be listening to music, watching video, or otherwise interacting with content through the input and output devices, wherein an operation 710 may include the voice-enabled client device receiving a multi-modal natural language interaction to engage in a transaction to dedicate the music, video, or other content.

[0104] Thus, in one implementation, in response to the voice-enabled client device receiving the multi-modal interaction in operation 710 that includes a natural language utterance, the voice-enabled client device may invoke an Automatic Speech Recognizer (ASR) to generate one or more preliminary interpretations of the utterance. The ASR may then provide the preliminary interpretations to a conversational language processor, which may attempt to determine an intent for the multi-modal interaction. In one implementation, to determine the intent for the multi-modal interaction, the conversational language processor may determine a most likely context for the interaction from the preliminary interpretations of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared

knowledge, or any other suitable information for interpreting the multi-modal interaction. Thus, in one implementation, an operation 720 may include the conversational language processor detecting a content dedication request in response to determining that the intent for the multi-modal interaction relates to a content dedication request. The conversational language processor may then invoke a content dedication application to resolve the content dedication request.

[0105] For example, in one implementation, the multi-modal interaction received in operation 720 may include the utterance "Buy that song and send it to Michael." In response to the initial multi-modal interaction, the ASR may generate a preliminary interpretation that includes words and/or phrases such as "Buy," "that song," "send it," and "to Michael." The ASR may then provide the preliminary interpretation to the conversational language processor, which may determine that the word and/or phrase combination of "Buy" and "send it" indicates that the most likely intent of the interaction includes a content dedication request, while the word and/or phrase combination of "that song" and "to Michael" provide criteria for the intended content and recipient for the dedication. Furthermore, in response to determining the most likely intent of the interaction, the conversational language processor may establish a device context, a dedication context, a content or music context, an address book context, or other suitable contexts in an attempt to resolve the request.

[0106] For example, the device context may enable the conversational language processor to retrieve data from the voice-enabled client device or another suitable device that provides the user's intended meaning for the phrase "that song" (e.g., the user may be referring to a song playing on the user's satellite radio device, such that the conversational language processor can identify the song that was playing when the device interaction was received in operation 710). Furthermore, the address book context may enable the conversational language processor to identify "Michael," the intended recipient of the dedication request. The conversational language processor may then receive appropriate results for resolving the intent of the request and present the results to the user through the output device (e.g., displaying information requesting that the user confirm that the content and recipient was correctly identified, playing a sample audio clip of the song, displaying options to purchase the song, recommending similar songs or similar artists, etc.). Accordingly, in response to detecting the content dedication request in operation 720 and identifying the relevant

criteria identifying the content to be dedicated and the intended recipient, the content dedication application may be invoked to process the content dedication request.

[0107] For example, in one implementation, processing the content dedication request may include the content dedication application capturing a natural language utterance that contains the dedication for the requested content in an operation 730. For example, the content dedication application may provide a prompt through the output device that instructs the user to provide the dedication utterance (e.g., a visual or audible prompt instructing the user to begin speaking, to speak after an audible beep, etc.). The user may then provide the dedication utterance through the voice-enabled input device, and the dedication utterance may then be converted into an electronic signal that the content dedication application captures in operation 730. In addition, an operation 740 may include the content dedication application further prompting the user to provide any additional tags for the dedicated content (e.g., an image or a picture to be inserted as album art for the dedicated content, one or more natural language utterances to insert as voice-tags for the dedicated content, one or more natural language utterances to be transcribed into text that can be inserted in tags for the dedicated content, a non-speech input or data input that include information to insert in tags for the dedicated content, etc.).

[0108] Thus, in response to determining that the user has provided additional information to insert in tags for the dedication content in operation 740, the content dedication application may then capture the tags in an operation 750. For example, operation 750 may include the content dedication application capturing an image or picture that the user identifies for album art to be inserted in the dedicated content, capturing any natural language utterances to insert as voice-tags for the dedicated content, communicating with the ASR and/or conversational language processor to transcribe any natural language utterances to be inserted as text within tags for the dedicated content, capturing any non-speech inputs or data inputs that include information to insert as text within the tags for the dedicated content, or otherwise capturing information that relates to the additional tags.

[0109] In one implementation, the content dedication application may then process the content dedication request in an operation 760, which may include prompting the user to identify the recipient of the dedication. For example, the content dedication application may request information identifying the recipient of the dedication in

response to the user not having already identified the recipient, in response to the user identifying the recipient in a manner that includes ambiguity or other criteria that cannot be resolved without further information, to distinguish among different contact information known for the recipient, or in response to other appropriate circumstances. Thus, the user may provide any suitable multi-modal input that includes an e-mail address, a telephone number, an address book entry, or other criteria that can be used to uniquely identify the information for contacting the recipient of the dedication. Furthermore, in one implementation, processing the content dedication request in operation 760 may further include the content dedication application routing the request, including the dedication utterance, the additional tags (if any), and the information for contacting the recipient to the content dedication system through the messaging interface, wherein the content dedication system may then process a transaction for the content dedication request, as will be described in greater detail below.

[0110] According to one aspect of the invention, Figure 8 illustrates a flow diagram of an exemplary method for providing a natural language content dedication service. In particular, the method for providing a natural language content dedication service, as shown in Figure 8, may include an operation 810 in which a content dedication system may receive a natural language content dedication request through a messaging interface. For example, in one implementation, the content dedication system may receive the natural language content dedication request from a content dedication application that executes on a voice-enabled client device. In one implementation, the natural language content dedication request received in operation 810 may generally include a natural language dedication utterance, information to be inserted within one or more tags for content to be dedicated, and information identifying a recipient of the dedicated content.

[0111] For example, in one implementation, the dedication utterance received in operation 810 may include encoded audio that the content dedication system can insert within the dedicated content or use to verbally annotate the dedicated content. Alternatively (or additionally), the content dedication system may interpret and parse the dedication utterance to transcribe one or more words or phrases from the dedication utterance, wherein the transcribed words or phrases may provide a textual annotation for the dedicated content (e.g., the textual annotation may be inserted

within metadata for the dedicated content, such as an ID3 Comments tag). Similarly, any utterances to be inserted as voice-tags for the dedicated content may provide further verbal annotations for the dedicated content, or such utterances may be transcribed to provide further textual annotations for the dedicated content. In one implementation, verbal annotations, textual annotations, and other types of annotations may be created and associated with the dedicated content using techniques described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety.

[0112] In one implementation, in response to the content dedication system receiving the content dedication request, the dedication utterance, and any additional tags from the content dedication application in operation 810, the content dedication system may invoke a content dedication application on a voice-enabled server to process the content dedication request. In particular, an operation 820 may include the content dedication application on the voice-enabled server identifying the content requested for dedication. In one implementation, the request received from the content dedication application in operation 810 may include information identifying the requested content and/or information that the content dedication system can use to identify the requested content. For example, the request may include a multi-modal natural language input that includes a natural language utterance and/or a non-voice input, wherein the content dedication system may use shared knowledge relating to the dedication request to identify the content to be dedicated.

[0113] Thus, operation 820 may include the content dedication system invoking one or more natural language processing components (e.g., an ASR, a conversational language processor, etc.), searching one or more data repositories, interacting with one or more content providers, pulling data from a satellite radio system that played the requested content at the voice-enabled client device or another device in a hybrid processing environment, or otherwise consulting available resources that can be used to identify the content to be dedicated. For example, in one implementation, the hybrid processing environment may include a device having an application that can identify played content (e.g., a Shazam® listening device that a user can hold near a speaker to identify content playing through the speaker). Thus, operation 820 may generally

include the content dedication system communicating with any suitable device, system, application, or other resource to identify the content requested for dedication.

[0114] In one implementation, in response to identifying the content to be dedicated, an operation 830 may include the content dedication system identifying one or more purchase options for the dedication request. In one implementation, operation 830 may include the content dedication system communicating with a billing system that supports various purchase options to provide users with flexibility in the manner of requesting content dedications. For example, the billing system may support content dedication purchase options that include a buy-to-own purchase option, a pay-to-play purchase option, a paid subscription purchase option, or other appropriate options. In response to identifying the purchase option for the content dedication request, an operation 840 may include the content dedication system processing a transaction for the content identified in operation 820.

[0115] For example, in response to operation 830 indicating that the user has selected the buy-to-own purchase option, the transaction processed in operation 840 may include purchasing full rights to the requested content from an appropriate content provider, wherein the billing system may then charge the user of the voice-enabled client device an appropriate amount that includes costs for purchasing the content from the content provider, adding the natural language utterance dedicating the content, tagging the dedicated content, delivering the dedicated content to the recipient, or any other appropriate services rendered for the content dedication request. Alternatively, in response to operation 830 indicating that the user has selected the pay-to-play purchase option, the transaction processed in operation 840 may include purchasing rights from the content provider that permits the purchased content to be played a predetermined number of times, wherein the billing system may charge the user in a similar manner as described above, except that the cost for purchasing limited rights to the content under the pay-to-play purchase option may be somewhat less than the cost for purchasing ownership rights to the content, as in the buy-to-own purchase option.

[0116] In another alternative implementation, in response to determining that the requested content dedication includes a selection of the paid subscription purchase option, the user may pay a periodic service charge to the content dedication system that permits the user to make a predetermined number of content dedications in a

subscription period, an unlimited number of content dedications in the subscription period, or otherwise make content dedications in accordance with terms defined in a subscription (e.g., different subscription levels having different content dedication options may be offered, such as a subscription level that only permits utterance dedications, another subscription level that further permits interpreting and parsing utterance dedications, etc.). Thus, the content transaction may be processed in operation 840 according to the purchase options identified in operation 830, and the content dedication system may then further process the dedication request to customize the dedicated content according to criteria provided in the request previously received in operation 810.

[0117] For example, in one implementation, an operation 850 may include the content dedication system inserting the natural language dedication utterance into the dedicated content, verbally annotating the dedicated content with the dedication utterance, or otherwise associating the dedicated content with the dedication utterance. Furthermore, an operation 860 may include the content dedication system determining whether any additional tags have been specified for the dedicated content. For example, as noted above, the dedication request may identify an image or picture to insert as album art for the dedicated content, one or more natural language utterances, non-voice, and/or data inputs to be transcribed into text that can be inserted in tags for the dedicated content, or any other suitable information that can be inserted within or associated with metadata for the dedicated content. Thus, in response to determining that information for any additional tags has been provided, an operation 870 may include inserting such additional tags into the dedicated content.

[0118] In one implementation, in response to having purchased the content requested for dedication, associating the dedication utterance with the dedicated content, and associating the additional tags (if any) with the dedicated content, the content dedication system may send a content dedication message to the recipient of the dedication in an operation 880. For example, the content dedication message may include a Short Message Service (SMS) text message, an electronic mail message, an automated telephone call managed by a text-to-speech engine, or any other appropriate message that can be appropriately delivered to the recipient (e.g., the message may include a link that the recipient can select to stream, download, or otherwise access the dedicated content, the dedication utterance, etc.). Thus, the

content dedication message may generally notify the recipient that content has been dedicated to the recipient and provide various mechanisms for the recipient to access the content dedication, as will be apparent.

[0119] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include various mechanisms for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, or other storage media, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, or other transmission media. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience, and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[0120] Accordingly, aspects and implementations of the invention may be described herein as including a particular feature, structure, or characteristic, but it will be apparent that every aspect or implementation may or may not necessarily include the particular feature, structure, or characteristic. In addition, when a particular feature, structure, or characteristic has been described in connection with a given aspect or implementation, it will be understood that such feature, structure, or characteristic may be included in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the preceding description without departing from the scope or spirit of the invention, and the specification and drawings should therefore be regarded as exemplary only, with the scope of the invention determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A system for providing a natural language content dedication service, comprising:

an electronic device configured to:

receive a multi-modal interaction that includes a request to dedicate content, wherein the multi-modal interaction includes a natural language utterance relating to the request to dedicate the content; and

capture an utterance dedicating the content for a recipient; and

a content dedication system in communication with the electronic device through a message interface, wherein the content dedication system is configured to:

identify the content to dedicate for the recipient based on an interpretation of the natural language utterance included in the multi-modal interaction;

process a transaction for the identified content to dedicate for the recipient; and

send a content dedication message to the recipient, wherein the content dedication message includes information for the recipient to access the content, and wherein the content dedication message further includes information for the recipient to access the utterance dedicating the content for the recipient.

2. The system of claim 1, wherein the content dedication system is further configured to receive one or more messages from the electronic device through the messaging interface, wherein one or more messages contain information relating to the multi-modal interaction that includes the request to dedicate the content, and wherein the one or more messages further contain information corresponding to the utterance dedicating the content for the recipient.

3. The system of claim 2, wherein the electronic device is further configured to receive information corresponding to one or more tags dedicating the content for the recipient.

4. The system of claim 3, wherein the content dedication system is further configured to:

insert the utterance dedicating the content for the recipient within the content; and

insert the tags dedicating the content for the recipient within metadata for the content.

5. The system of claim 1, wherein the tags dedicating the content for the recipient include one or more of an image or a picture to insert within album art metadata for the content.

6. The system of claim 1, wherein the content dedication system is further configured to annotate the content with the utterance dedicating the content for the recipient.

7. The system of claim 1, wherein the content dedication system is further configured to purchase the identified content from a content provider to process the transaction for the identified content.

8. The system of claim 7, wherein the content dedication system is further configured to:

identify a purchase option for the request to dedicate the content, wherein the purchase option includes one or more of a buy-to-own purchase option, a pay-to-pay purchase option, or a paid subscription purchase option; and

bill a user of the electronic device for the natural content dedication service based on identified purchase option.

9. The system of claim 1, wherein the content dedication message sent to the recipient includes an electronic link for the recipient to access the content and the utterance dedicating the content for the recipient.

10. The system of claim 1, wherein the content includes one or more of a song or a video.

11. A method for providing a natural language content dedication service, comprising:

receiving, at an electronic device, a multi-modal interaction that includes a request to dedicate content, wherein the multi-modal interaction includes a natural language utterance relating to the request to dedicate the content;

capturing an utterance dedicating the content for a recipient;

identifying the content to dedicate for the recipient based on an interpretation of the natural language utterance included in the multi-modal interaction;

processing, at a content dedication system, a transaction for the identified content to dedicate for the recipient; and

sending a content dedication message from the content dedication system to the recipient, wherein the content dedication message includes information for the recipient to access the content, and wherein the content dedication message further includes information for the recipient to access the utterance dedicating the content for the recipient.

12. The method of claim 11, further comprising receiving, at the content dedication system, one or more messages from the electronic device through a messaging interface, wherein one or more messages contain information relating to the multi-modal interaction that includes the request to dedicate the content, and wherein the one or more messages further contain information corresponding to the utterance dedicating the content for the recipient.

13. The method of claim 12, further comprising receiving, at the electronic device, information corresponding to one or more tags dedicating the content for the recipient.

14. The method of claim 13, further comprising:

inserting the utterance dedicating the content for the recipient within the content; and

inserting the tags dedicating the content for the recipient within metadata for the content.

15. The method of claim 11, wherein the tags dedicating the content for the recipient include one or more of an image or a picture to insert within album art metadata for the content.

16. The method of claim 11, further comprising annotating the content with the utterance dedicating the content for the recipient.

17. The method of claim 11, further comprising purchasing, by the content dedication system, the identified content from a content provider to process the transaction for the identified content.

18. The method of claim 17, further comprising:

identifying, at the content dedication system, a purchase option for the request to dedicate the content, wherein the purchase option includes one or more of a buy-to-own purchase option, a pay-to-pay purchase option, or a paid subscription purchase option; and

billing, by the content dedication system, a user of the electronic device for the natural content dedication service based on identified purchase option.

19. The method of claim 11, wherein the content dedication message sent to the recipient includes an electronic link for the recipient to access the content and the utterance dedicating the content for the recipient.

20. The method of claim 11, wherein the content includes one or more of a song or a video.

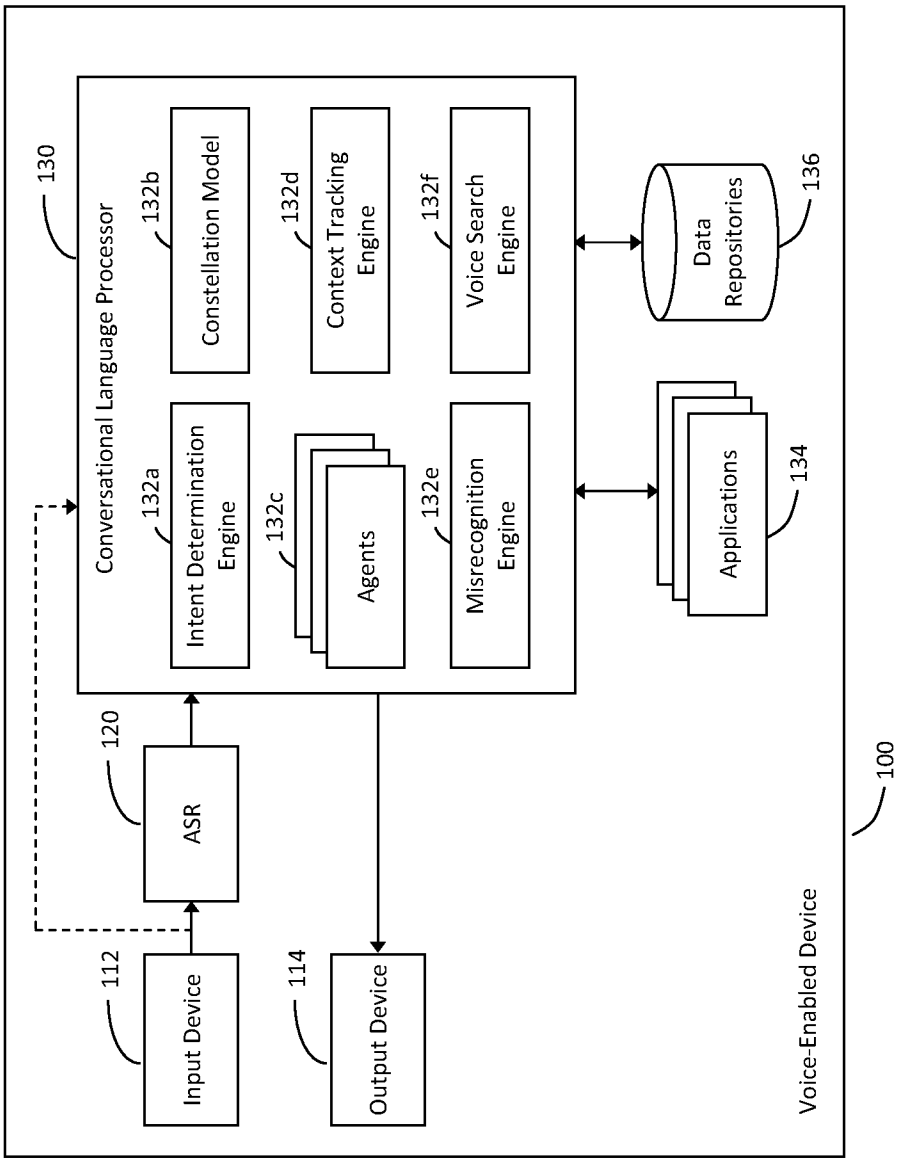
Voice-Enabled Device
100
Input Device
112
Output Device
114
ASR
120
Conversational Language Processor
130
Intent Determination Engine
132a
Constellation Model
132b
Agents
132c
Context Tracking Engine
132d
Misrecognition Engine
132e
Voice Search Engine
132f
Applications
134
Data Repositories
136

Figure 1

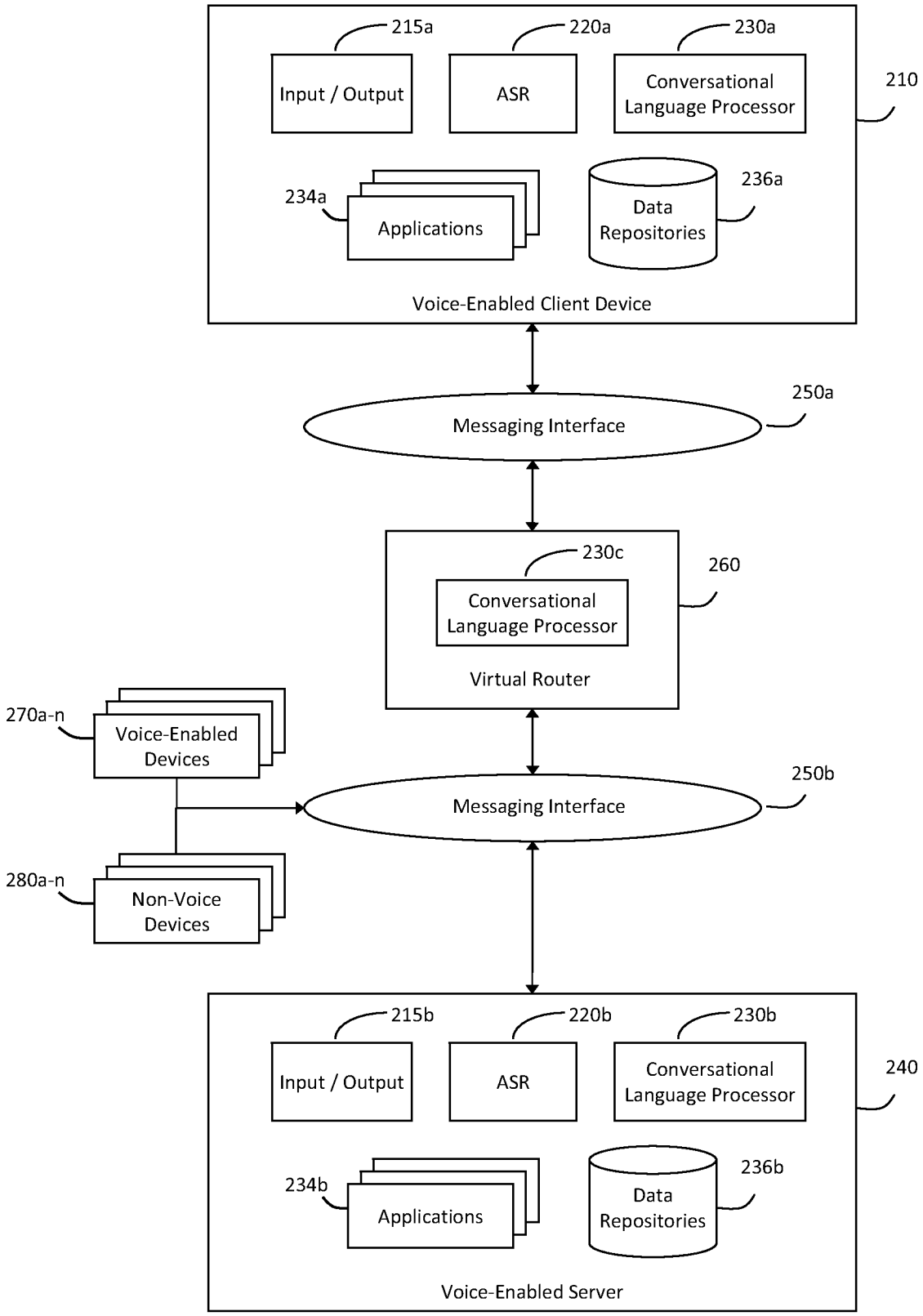
215a
220a
230a
Input / Output
ASR
Conversational Language Processor
210
234a
Applications
Data Repositories
236a
Voice-Enabled Client Device
250a
Messaging Interface
230c
260
Conversational Language Processor
Virtual Router
270a-n
Voice-Enabled Devices
250b
Messaging Interface
280a-n
Non-Voice Devices
215b
220b
230b
Input / Output
ASR
Conversational Language Processor
240
234b
Applications
Data Repositories
236b
Voice-Enabled Server

Figure 2

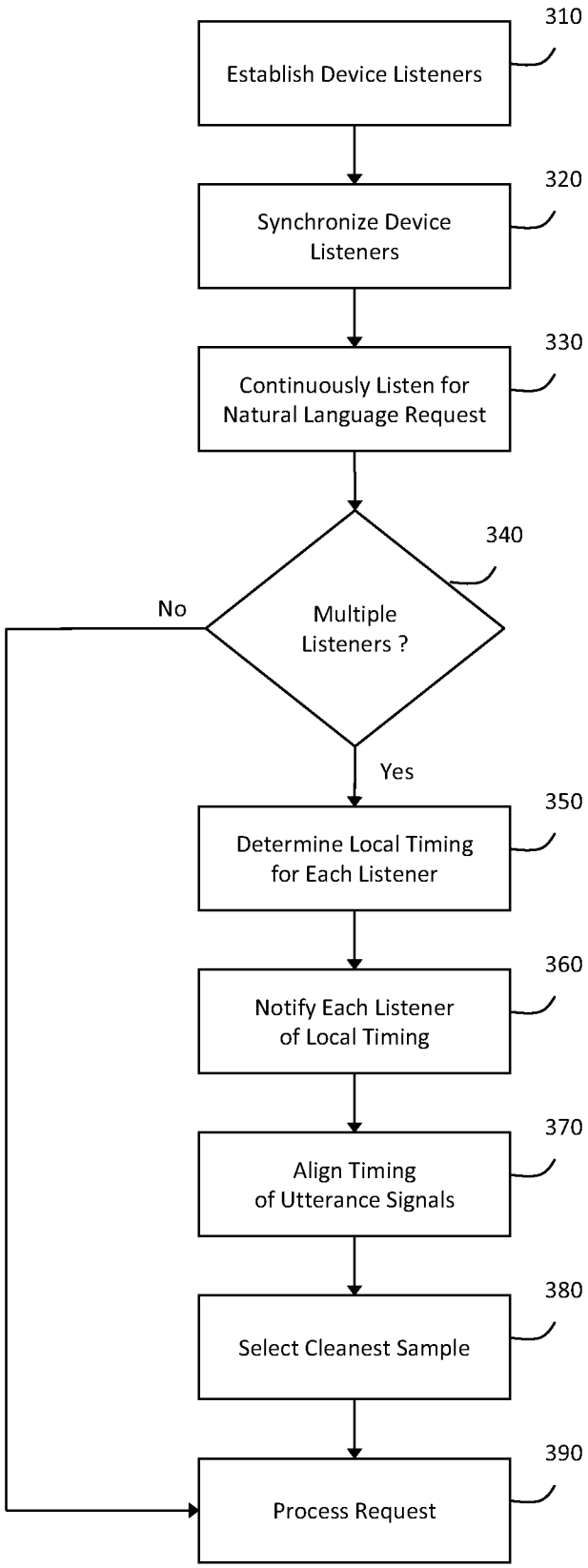
310
Establish Device Listeners
320
Synchronize Device Listeners
330
Continuously Listen for Natural Language Request
340
Multiple Listeners ?
No
Yes
350
Determine Local Timing for Each Listener
360
Notify Each Listener of Local Timing
370
Align Timing of Utterance Signals
380
Select Cleanest Sample
390
Process Request

Figure 3

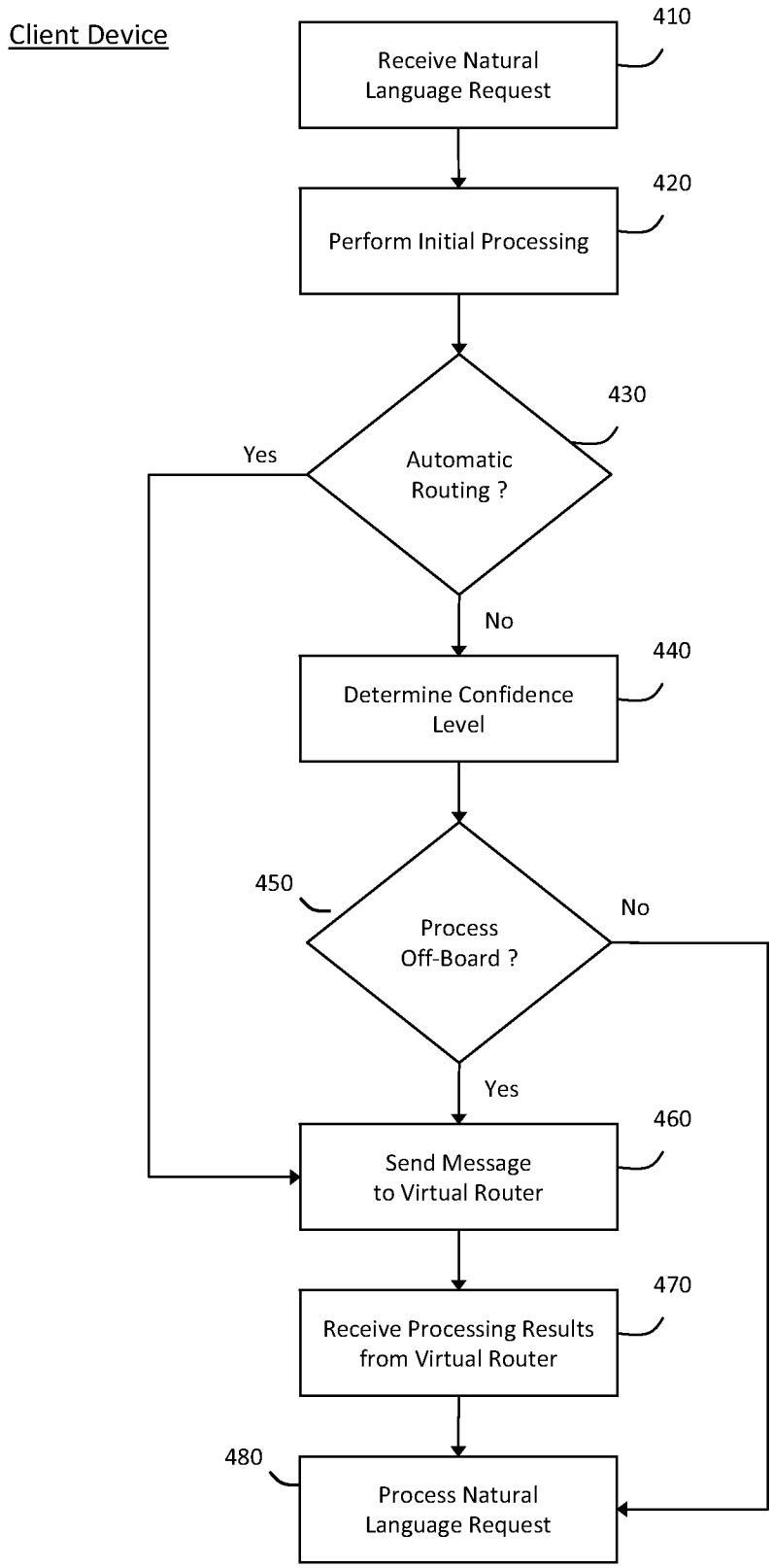
Client Device
Receive Natural Language Request
410
Perform Initial Processing
420
Automatic Routing ?
430
Yes
No
Determine Confidence Level
440
Process Off-Board ?
450
No
Yes
Send Message to Virtual Router
460
Receive Processing Results from Virtual Router
470
Process Natural Language Request
480

Figure 4

Virtual Router

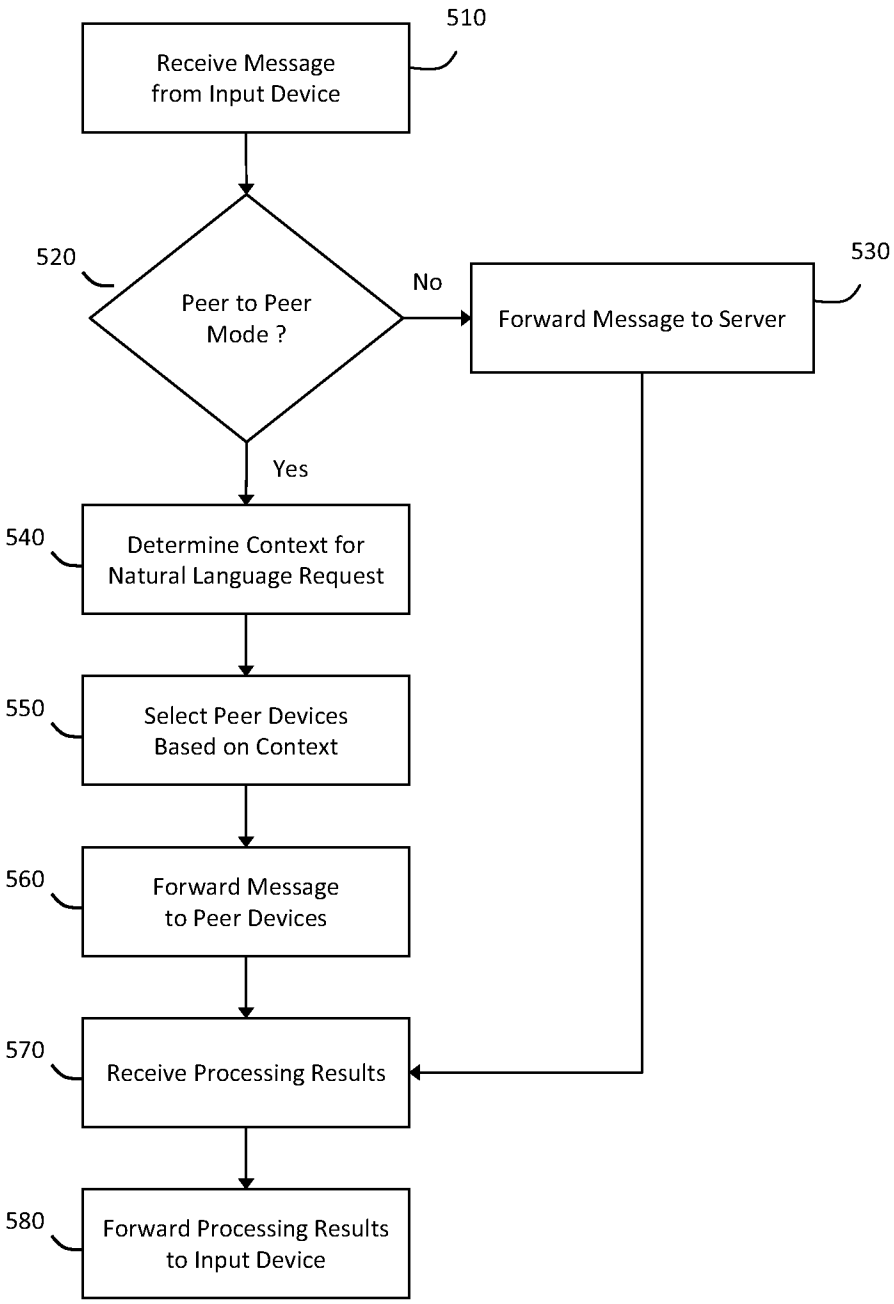

Figure 5

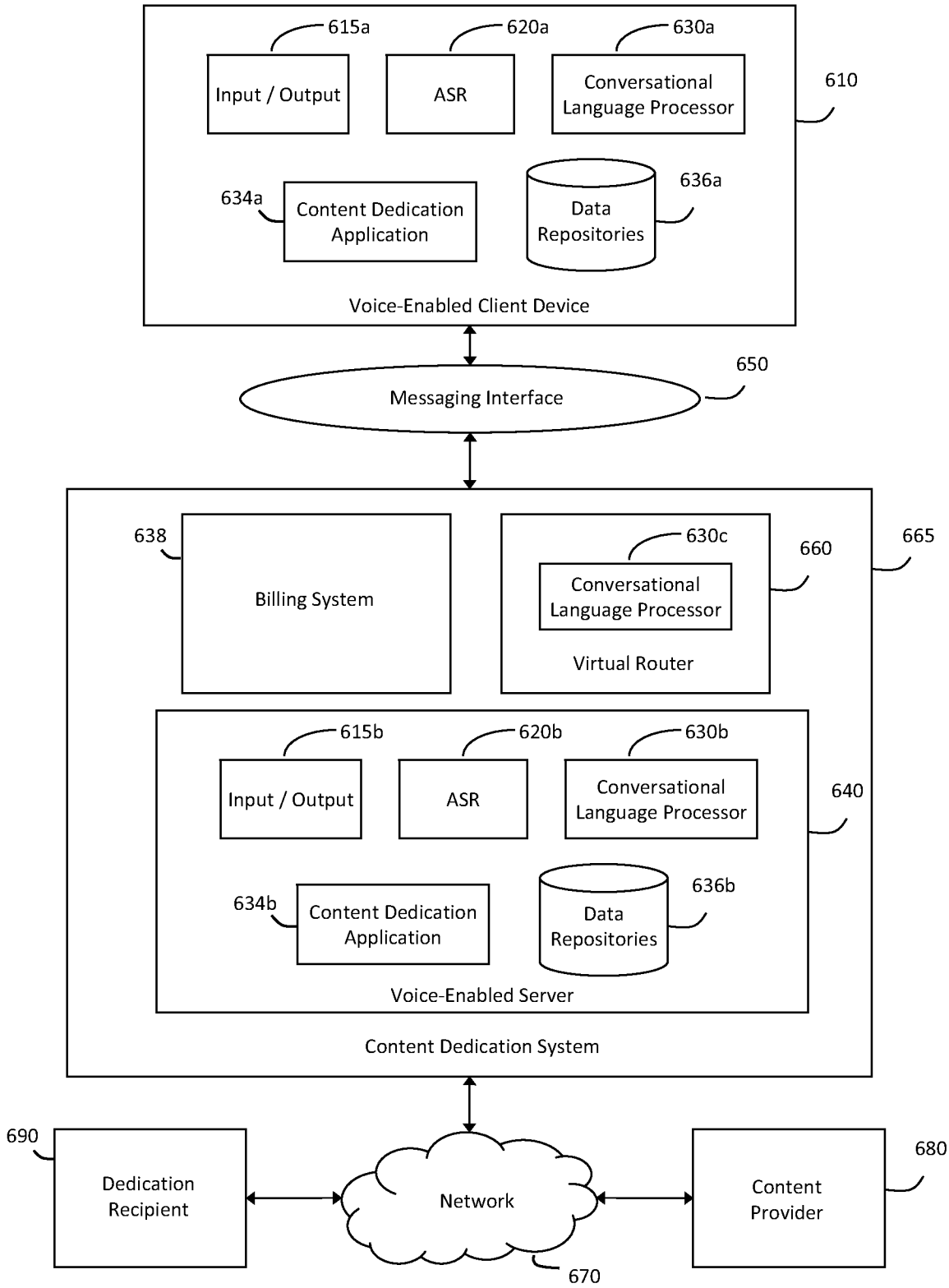
615a
620a
630a
Input / Output
ASR
Conversational Language Processor
610
634a
Content Dedication Application
636a
Data Repositories
Voice-Enabled Client Device
650
Messaging Interface
638
Billing System
630c
660
665
Conversational Language Processor
Virtual Router
615b
620b
630b
Input / Output
ASR
Conversational Language Processor
640
634b
Content Dedication Application
636b
Data Repositories
Voice-Enabled Server
Content Dedication System
690
Dedication Recipient
Network
680
Content Provider
670

Figure 6

Client Device

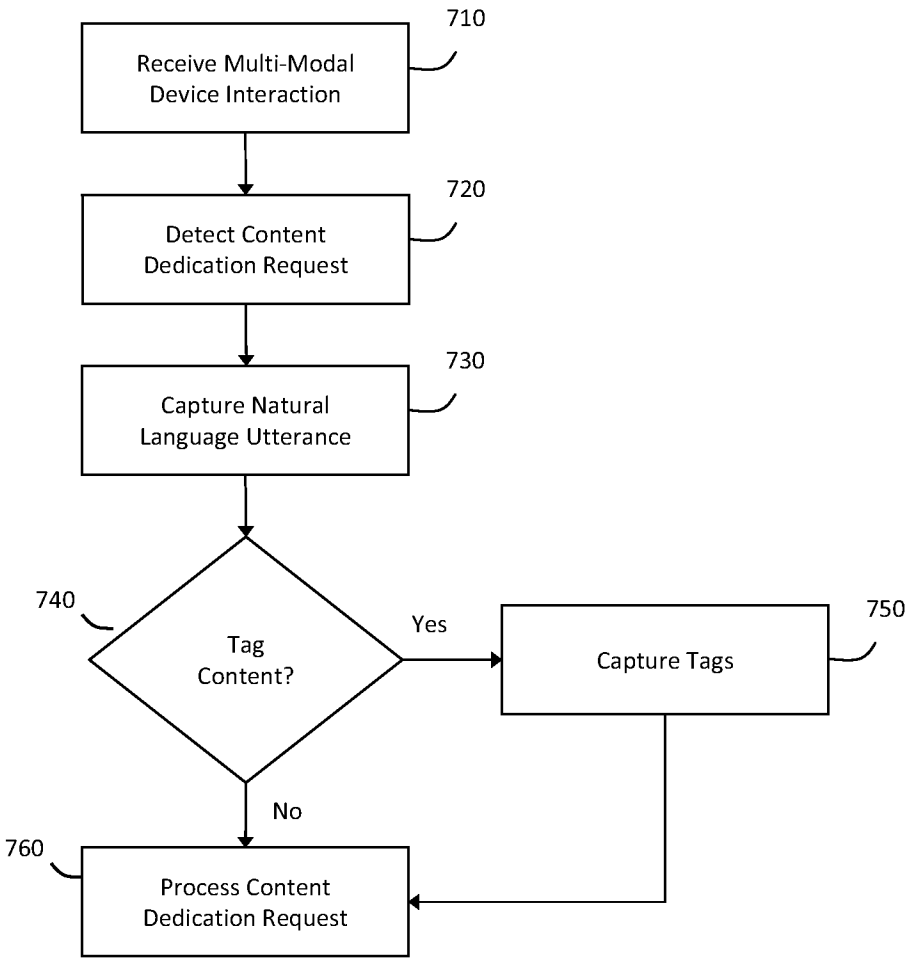

Figure 7

Content Dedication System

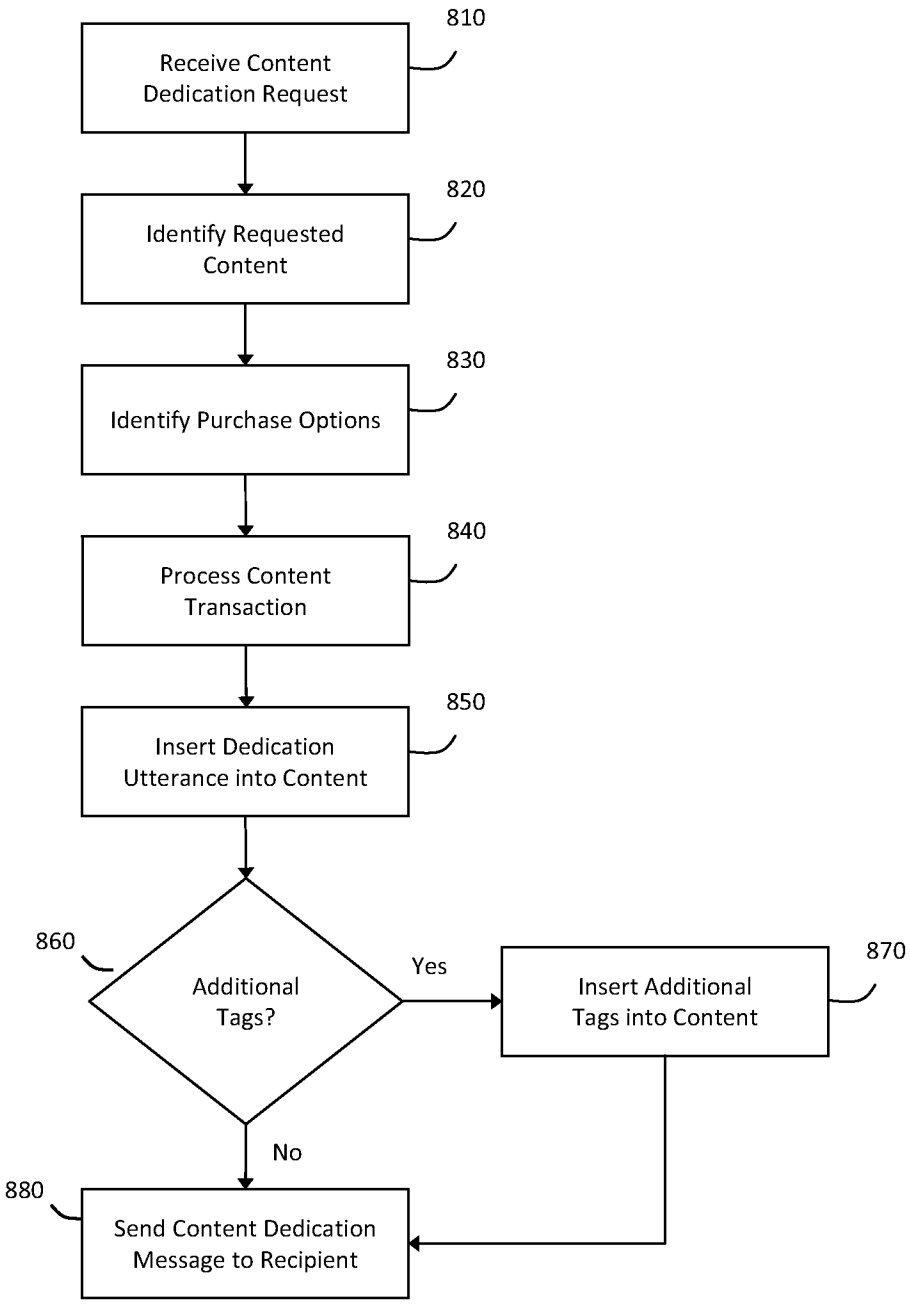

Figure 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/56109

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2010.01)
USPC - 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/257; 704/231; 704/E15.001 (see terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST(PGPB,USPT,USOC,EPAB,JPAB), GOOGLE SCHOLAR
terms: purchase, paid subscription, content dedication, natural language, transaction, utterance, song, movie, video, buy to own, pay, annotate, tag, recipient, multimodel, interact, metadata.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/0193420 A1 (Kennewick et al.) 30 September 2004 (30.09.2004), fig. 1, 3, 5-6, para [0013]-[0020], [0082], [0095]-[0182].	1-20
A	US 2009/0216540 A1 (Tessel et al.) 27 August 2009 (27.08.2009), entire document.	1-20
A	US 2008/0140385 A1 (Mahajan et al.) 12 June 2008 (12.06.2008), entire document.	1-20

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 December 2010 (18.12.2010)	07 JAN 2011
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)