(51) International Patent Classification:
G10L 15/26 (2006.01)   H04L 29/08 (2006.01)

(21) International Application Number:
PCT/EP20 12/070124

(22) International Filing Date:
11 October 2012 (11.10.2012)

(25) Filing Language:   English

(26) Publication Language:   English

(30) Priority Data:
P201 131 647 14 October 2011 (14. 10.2011) ES


[Continued on nextpage]

(54) Title: A METHOD TO MANAGE SPEECH RECOGNITION OF AUDIO CALLS

(57) Abstract: A method to manage speech recognition of audio calls. In the method of the invention said audio calls are performed in a Media Resource Control Protocol, or MRCP, based system and said speech recognition is carried out by an ASR engine controlled by a MRCP server by looking for a match between an audio stream generated by a user and a compiled grammar. The method is characterised in that it comprises performing said speech recognition using a continuously wordspotting mode, that differs from Normal and HotWord Mode known from MRCPv2. This is done by sending, said MRCP server, events regularly to said user when matches are produced, each of said events indicating a partial result of said speech recognition and ignoring unsuccessful matches, stopping said speech recognition when receiving a stop request from said user or when said audio stream finishes. Also mentions loading and unloading of grammars while performing recognition.

Figure 4
Published:
— with international search report (Art. 21(3))
— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))
A method to manage speech recognition of audio calls

Field of the art
The present invention generally relates to a method to manage speech recognition of audio calls, said audio calls performed in a Media Resource Control Protocol, or MRCP, based system, said speech recognition carried out by an ASR engine controlled by a MRCP server by looking for a match between an audio stream generated by a user and a compiled grammar, and more concretely to a method that comprises performing said speech recognition continuously by sending, said MRCP server, events regularly to said user when matches are produced, each of said events indicating a partial result of said speech recognition and ignoring unsuccessful matches, stopping said speech recognition when receiving a stop request from said user or when said audio stream finishes.

Prior State of the Art
One of the main uses of Speech Recognition so far has been IVR (Interactive Voice Response) systems. Call Centres use IVR systems to reduce costs, by automating the client request management. In order to do this, the speech needs to be captured and then processed by an ASR (Automatic Speech Recognition) engine. The ASR analyses the speech and produces one match if something has been detected.

The ASR engine requires one grammar to process the speech. This grammar contains a limited set of expected words or sentences. Results are matched only if the person speaking says any of the list items. Results are analysed programmatically and the system will behave in one or another way depending on them.

In order to interact with a speech recognition engine, the IETF defines a protocol called Media Resource Control Protocol (MRCP). MRCP is described in RFC 4463 [1]. This protocol controls media service resources like speech synthesizers, recognizers, etc. As defined in this RFC, MRCP uses RTSP as session control protocol. Currently, there is a draft to specify MRCP using SIP (MRCP version 2 [2]).

Regarding architecture of MRCP, as it will be shown in Figure 1, it consists of a client that requires media streams generated (recognizers) or needs media streams processed (synthesizers) and a server that has the resources or devices to process (recognizers) or generate (synthesizers) the streams.
The client establishes a control session with the server for media processing using a protocol such as RTSP (MRCP1) or SIP (MRCPv2). This will also set up and establish the RTP stream between the client and the server or another RTP endpoint.

The MRCP message set consists of requests from the client to the server, responses from the server to the client and asynchronous events from the server to the client.

Figure 2 will show the exchange of messages, for both SIP and MRCP sessions. Initially, the MRCPv2 client (21) sends the SIP INVITE (21 1) to the MRCPv2 server (22) in order to establish the session, indicating the type of server resource required (a=resource:speechrecog). The MRCPv2 server must respond with the full Channel-Identifier and the TCP port to which MRCPv2 messages should be sent (221).

SET-PARAMS (213) is a method that carries a set of parameters in the header to configure the ASR engine. The response from server to client when the ASR has been configured is 200 COMPLETE (222).

The DEFINE-GRAMMAR (214) method, from the client to the server, provides one or more grammars and requests the server to access, fetch, and compile the grammars as needed. The response from server to client when this has been done is 200 COMPLETE (223).

The RECOGNIZE method (215) requests the recognizer resource to start recognizing. The RECOGNIZE request uses the message body to specify the grammars applicable to the request. MRCPv2 must send the response 200 IN-PROGRESS (224) to make notice the MRCPv2 client that recognition has just started.

START-OF-INPUT (225) is an event from the server to the client indicating that the recognition resource has detected speech.

RECOGNITION-COMPLETE (226) is an Event from the recognizer resource to the client indicating that the recognition completed. The recognition result is sent in the body of the MRCPv2 message.

The "STOP" method (216) from the client to the server tells the resource to stop recognition if a request is active. Server confirms the end of recognition with 200 COMPLETE (227).

Finally, as MRCPv2 channel is not used any more, SIP session can be also finished (217, 228, 218).

There are 2 supported operation modes for recognition:
- Normal Mode Recognition: tries to match all of the speech against the grammar and returns a no-match status if the input fails to match or the method times out.

- Hotword Mode Recognition: the recognizer looks for a match against specific speech grammar and ignores speech that does not match. The recognition completes only for a successful match of grammar or if the client cancels the request or if there is a non-input or recognition timeout.

There are quite a few parameters that can be sent to the MRCP Server within RECOGNIZE (215) or SET-PARAMS (213) requests by including their values in the header fields. Those values are used to configure the ASR engine and prepare it for recognition. For instance, configuration values such as confidence threshold, sensitivity level, etc are typically used. Those parameters related to timeouts that might be meaningful are the following:

- No-Input-Timeout: indicates when recognition is started and there is no speech detected for a certain period of time. In that case, the recognizer sends a RECOGNITION-COMPLETE event to the client and terminates the recognition operation.

- Recognition-Timeout: indicates when recognition is started and there is no match for a certain period of time. In that case, the recognizer sends a RECOGNITION-COMPLETE event to the client and terminates the recognition operation.

- Speech-Complete-Timeout: indicates the length of silence required following user speech before the speech recognizer finalizes a result (either accepting it or generating a nomatch event).

- Speech-Incomplete-Timeout: indicates the length of silence following user speech after which a recognizer finalizes a result. Once the timeout is triggered, the partial result is rejected.

- Hotword-Max-Duration: indicates the maximum length of an utterance that will be considered for Hotword recognition.

- Hotword-Min-Duration: indicates the minimum length of an utterance that will be considered for Hotword recognition.

As described before, media resources which provide speech recognition functionality, such as Speech Processing servers in IVRs, are controlled through a standard protocol called MRCP. This protocol supports 2 different recognition modes, in which an audio stream is matched against a predefined grammar to produce a
recognition result. These modes are targeted for applications with short, bounded audio
streams, such as the conversation that a person maintains with a machine. None of the
existing operation modes support arbitrarily long audio streams, such as those
happening in a person-to-person conversation, which require a continuous loop of
recognition requests which is not optimally controlled.

In fact, certain applications demand that, during a person-to-person
conversation, the results of the recognition appear in real time or at least as the
conversation is still ongoing. For this purpose, the recognition modes of the current
protocol force an intermittent recognition process with request-response iterations.
There is no mechanism in the protocol to control a media resource so that it can
produce continuous partial recognition results during the course of the stream.

Moreover, the grammars to be applied for recognition are loaded at the
beginning of a recognition process and cannot be changed until this process ends.
However, in natural conversations between two speakers, the variations in the subjects
of the content may require adaptation of the grammars during the call. In such cases,
the possibility of dynamically loading/unloading grammars is very useful, but it is not
covered by the current MRCP protocol.

Going into further detail of the specification provided by IETF, the MRCPv2
server behaves as depicted in the following state machine (Figure 3)

Being in the IDLE state (31), when a RECOGNIZE request (312) arrives from
the client, the state changes to RECOGNIZING STATE (32). When a match result is
produced by the recognizer, a RECOGNITION-COMPLETE (321) event is sent to the
client and the state turns to RECOGNIZED (33). This means that the recognition
process is over and the client needs to restart the process to continue recognizing,
starting again with a DEFINE-GRAMMAR (311) message and RECOGNIZE request
(312).

For the aforementioned use cases, a more efficient procedure is required to
deal with dynamic grammars and continuous recognition, as the current MRCP
protocol establishes that the client needs to send RECOGNIZE request to re-start the
state machine after each returned result. Besides, the DEFINE-GRAMMAR method
(311) is only permitted in the IDLE state (31), and not while in the RECOGNIZING (32)
state.

On the other hand, a person-to-person communication involves 2 channels, one
per each party, whereas traditional IVR systems involve only one. The MRCP protocol
manages each channel as an independent session and leaves the client side to keep
track of the results for each channel. The ability to load dynamic grammars is required to adapt the recognition process depending on the results obtained by each party.

5 **Description of the Invention**

It is necessary to offer an alternative to the state of the art which covers the gaps found therein, particularly related to the lack of proposals which really allow performing speech recognition of arbitrarily long audio streams, in real time or near real time, without having to wait until the end of the audio stream to return results, and allowing using different grammars in the MRCP server in order to adapt the speech recognition while the stream is ongoing.

To that end, the present invention provides a method to manage speech recognition of audio calls, said audio calls performed in a Media Resource Control Protocol, or MRCP, based system and said speech recognition carried out by an ASR engine controlled by a MRCP server by looking for a match between an audio stream generated by a user and a compiled grammar.

On contrary to the known proposals, the method of the invention, in a characteristic manner it further comprises performing said speech recognition continuously by sending, said MRCP server, events regularly to said user when matches are produced, each of said events indicating a partial result of said speech recognition and ignoring unsuccessful matches, stopping said speech recognition when receiving a stop request from said user or when said audio stream finishes.

Other embodiments of the method of the invention are described according to appended claims 2 to 12, and in a subsequent section related to the detailed description of several embodiments.

**Brief Description of the Drawings**

The previous and other advantages and features will be more fully understood from the following detailed description of embodiments, with reference to the attached drawings (some of which have already been described in the Prior State of the Art section), which must be considered in an illustrative and non-limiting manner, in which:

Figure 1 shows current MRCP version 2 architecture.

Figure 2 shows the signaling flow between the MRCP client and the MRCP server according to MRCP protocol.
Figure 3 shows current MRCP version 2 state machine of the MRCP server. Figure 4 shows the modified state machine of the MRCP server according to an embodiment of the present invention.

Figure 5 shows the modified signaling flow between the MRCP client and the MRCP server according to an embodiment of the present invention. Figure 6 shows the signaling flow between the MRCP client and the MRCP server when the load grammar timeout is exceeded, according to an embodiment of the present invention. Figure 7 shows the signaling flow between the MRCP client and the MRCP server when a grammar is unloaded, according to an embodiment of the present invention.

**Detailed Description of Several Embodiments**

This patent presents a new procedure for continuous speech recognition of audio calls in a MRCP-based system as shown in Figure 5. Specifically, a new operation mode is proposed, different from to the two modes defined by IETF (Normal Mode Recognition, HotWord Mode Recognition). Within the scope of this document, we will name this new mode as Continuous WordSpotting Mode Recognition; this is not a definitive naming concerning patent purposes however.

In this operation mode, the recognizer looks for a match according to the compiled grammars of the MRCP server and ignores everything that does not match. In addition, recognition continues even if after a successful match is found, and it only ends when there is a STOP request coming from the client, or the input audio stream finishes. Events are sent from server to client whenever there is a match so that speech processing can be in near real time. In this document, those events have been named PARTIAL RESULT.

Under normal circumstances, a PARTIAL RESULT can be sent from server to client at any time since the beginning of the recognition process. It is the ASR engine which decides that there is one match and it must be sent to the client. However, for cases where full control is required from the client, a mechanism for the client to impose the minimum frequency with which the server will check for unreturned partial results is proposed. This mechanism defines an optional parameter, added to the collection defined in the MRCP protocol. This parameter can be set by the SET-PARMS or the RECOGNIZE requests of the standard. In this document, this
parameter is called tracebackrate; it represents the maximum time interval at which the server should check for unreturned partial results.

In addition, the operation with unbounded audio streams opens up the possibilities to improve the quality of the results by using different grammars as the stream progresses. To support this, an extension to the MRCP protocol is proposed that allows dynamic loading and unloading of grammars. In order to load an additional grammar during the recognizing state, the standard DEFINE-GRAMMAR request will be used. Additionally, a new type of request, named UNLOAD GRAMMAR in the document, is proposed. This request orders the server to unload the specified grammar.

In order to avoid potentially too long delays in grammar loading, which could result in matches being missed, the inclusion of an additional timeout to the MRCP specifications is proposed. This timeout is named load-grammar-timeout. It is to be set at the beginning of the recognition process and can be changed anytime during the recognition process by means of a SET-PARAMS request. This parameter is mandatory and it defines the maximum time that the client is willing to wait for a reply to the DEFINE-GRAMMAR request. This is to keep the time that the server is without a properly configured grammar (and thus unable to produce recognition results) under control.

In the light of the specifications of the MRCP standard [1], this invention aims to upgrade the state machine of the server side and provide new messages for the proposed Continuous WordSpotting Mode.

STATE MACHINE

The state machine, as shown in Figure 4, includes several changes with respect to the state machine of MRCPv2. In particular, three new actions can be performed during the RECOGNIZING state:
- Send partial results (424)
- Unload one of the active grammars (425)
- Define a new grammar (426)

SPEECH PROCESSING OF A CALL

When the call is established in the PBX, the first step of the MRCP protocol is to create two processing channels between the PBX and the Media Server, one per each leg of the call. Both sessions are set up with the SIP protocol (it could be set up with
the RTSP protocol instead, both choices have been tackled in the standard) and associated with a MRCP channel identifier.

Once the two MRCP channels are successfully opened, the procedure for the new continuous mode was illustrated in Figure 5. First, the Load-Grammar-Timeout parameter is set by means of the following header field of the DEFINE-GRAMMAR (5102) message:

    Load-Grammar-Timeout: <load-grammar-timeout>

This timeout can be set also with the SET-PARAMS (5101) request.

Adjusting this parameter is a requirement for working with dynamic grammars which adapt recognition to the context of the conversations. In fact, grammars should be as short as established by the time the server is able to process them according to the Load-Grammar-Timeout.

The continuous-wordsspotting mode is signaled by the SET-PARAMS (5101) or the RECOGNIZE (5103) requests, which contain the following headers:

    Recognition-Mode: continuous-wordsspotting
    Tracebackrate: <time slice for partial results>

The tracebackrate parameter is optional, and its purpose is to configure the maximum time interval at which the server should check for unreturned partial results.

Once the RECOGNIZE request has been received and processed by the server, partial results are sent, if available, to the client on a certain period of milliseconds given by the tracebackrate parameter as a maximum (5204, 5205, 5208, 5209). If the ASR engine has not produced any result within this maximum value, it is forced to check if there are matches.

Results are processed in real time by the MRCP client, which may decide to add new grammar(s) to be used in order to try to obtain more accurate results. For this purpose, a new DEFINE-GRAMMAR (5104) request is sent.

Only if the client explicitly requests the end of recognition with an STOP request (5105), will the server state change from RECOGNIZING to IDLE. Otherwise, the process continues unlimitedly controlled by the two defined timers: Load-Grammar-Timeout, and Tracebackrate (if set).

Figure 6 showed what happens in case of exceeding the Load-Grammar-Timeout when a grammar is being loaded. In that case, the server sends a 503 COMPLETE (6206) response to the client with a cause field of Load-Grammar-Timeout and the server continues sending partial results (6207, 6208) with the previous grammar(s). The MRCP client could request another grammar(s) to be loaded instead.
The aforementioned procedure has no limitations concerning the amount of grammars to be used at the same time by the ASR engine. Splitting the rules into several small grammars favours the modularity and dynamicity of grammar loading and unloading, making these processes quicker. Therefore, the use case in which one grammar should be unloaded because it is not going to be used any more for recognition processing is described next.

Figure 7 showed the sequence of requests and responses in the case that a grammar is unloaded. The request UNLOAD-GRAM MAR (7104) is sent with the purpose of removing certain rules in the ASR engine. If the unloading is successful, a 200 COMPLETE (7206) response is issued.

Use Case of the Invention

The present invention aims to establish a procedure for wordspotting in person-to-person audio calls for a MRCP system. When natural language is managed, some features should be highlighted:

- There are two audio channels (caller and callee) and they need to be processed separately for better interpretation of the results, allowing later post-processing.
- Even if processed separately, these two channels must be associated to the same call (also for post-processing).
- Speech streams to be processed are arbitrarily long,
- Speech potentially includes "long" silence intervals, because of the two-way interaction, but recognition must not be stopped.
- Results should be produced and sent to client in (near) real time
- Grammars to be applied for speech processing should be loaded and unloaded dynamically to allow adaptation based on the context of the conversation,
- Time consumed in grammar loading should be short enough to avoid delays in processing.

Taking into account all the above requirements, this patent proposes a procedure for audio call recognition in a MRCP system as follows.

When a call is set up by VoIP PBX, a module is in charge of setting up the MRCP protocol with a Media Server. This module acts as a MRCP client which establishes two MRCP channels with the server that are processed separately (caller and callee speech).
First of all, a SIP session must be established between client and server as shown in Figure 2 (21 1, 221, 212). The SIP INVITE transaction and the underlying SDP offer/answer contain m-lines describing the resource control channel to be allocated. There must be one SDP m-line for each MRCPv2 resource to be used in the session.

The port number field of the m-line must contain the TCP listen port on the server in the SDP answer.

Next, an example of a SIP/SDP negotiation is provided. Both channels will share the same TCP connection for MRCP (m=application 32416 TCP/MRCPv2 1), but different UDP connection for RTP flows (m=audio 48260 RTP/AVP 0 96; m=audio 48261 RTP/AVP 0 96). Therefore, every MRCP message for any of the channels will be exchanged by the 32416 TCP port of server, whereas the RTP messages of one of the channels will go to 48260 UDP port and the other to 48261 UDP port.

The 200 OK reply contains the MRCP channel identifiers for each channel (a=channel:32AECB234338@speechrecog, a=channel:32AECB234339@speechrecog). All the subsequent MRCP messages will indicate this identifier so that the server can know which channel the MRCP message belongs to.

C->S: INVITE sip:mresources@server.example.com SIP/2.0

 Via:SIP/2.0/TCP client.atlanta.example.com:5060;
branch=z9hG4bK74bf3
 Max-Forwards:6
 To:MediaServer <sip:mresources@example.com>;tag=62784
 From:sarvi <sip:sarvi@example.com>;tag=1 928301774
 Call-ID:a84b4c76e66710
 CSeq:314162 INVITE
 Contact:<sip:sarvi@client.example.com>
 Content-Type:application/sdp
 Content-Length:...

v=0
o=sarvi 2890844526 2890844527 IN IP4 192.0.2.4
s=-
c=IN IP4 192.0.2.12
m=application 9 TCP/MRCPv2 1

a=setup:active
a=connection:new
a=resource:speechrecog
a=cmid:1
m=audio 49170 RTP/AVP 0 96
5
a=rtpmap:0 pcmu/8000
a=rtpmap:96 telephone-event/8000
a=fmtp:96 0-15
a=sendonly
a=mid:1
10
m=application 9 TCP/MRCPv2 1
a=setup:active
a=connection:existing
a=resource:speechrecog
a=cmid:2
15
m=audio 49171 RTP/AVP 0 96
a=rtpmap:0 pcmu/8000
a=rtpmap:96 telephone-event/8000
a=fmtp:96 0-15
a=sendonly
a=mid:2

S->C: SIP/2.0 200 OK
Via:SIP/2.0/TCP client.atlanta.example.com:5060;
branch=z9hG4bK74bf3;received=1 92.0.32.1 0
25
To:MediaServer <sip:mresources@example.com>;tag=62784
From:sarvi <sip:sarvi@example.com>;tag=1 928301 774
Call-ID:a84b4c76e66710
CSeq:314162 INVITE
Contact:<sip:mresources@server.example.com>
30
Content-Type:application/sdp
Content-Length:...
v=0
o= 2890842808 2890842809 IN IP4 192.0.2.4
s=-
35
c=IN IP4 192.0.2.1 1
m=application 32416 TCP/MRCPv2 1
a=setup:passive
a=connection: new
a=channel:32AECB234338@speechrecog
a=cmdid:1
m=audio 48260 RTP/AVP 0 96
a=rtpmap:0 pcmu/8000
a=rtpmap:96 telephone-event/8000
a=fmt 96 0-15
a=sendonly
a=mid:1
m=application 32416 TCP/MRCPv2 1
a=setup:passive
a=connection:existing
a=channel:32AECB234339@speechrecog
a=cmdid:2
m=audio 48261 RTP/AVP 0 96
a=rtpmap:0 pcmu/8000
a=rtpmap:96 telephone-event/8000
a=fmt 96 0-15
a=sendonly
a=mid:2

C->S: ACK sip:mresources@server.example.com SIP/2.0
Via:SIP/2.0/TCP client.atlanta.example.com:5060;
branch=z9hG4bK74bf4
Max-Forwards:6
To:MediaServer <sip:mresources@example.com>;tag=62784
From:Sarvi <sip:sarvi@example.com>;tag=1 928301 774
Call-ID:a84b4c76e66710
CSeq:314162 ACK
Content-Length:0

Once the SIP session is established, the MRCP protocol itself begins. For the
sake of simplicity, let us consider only one of the two call channels,
It could be demonstrated that the message flow for the other channel is analogous.

The first of the MRCP messages correspond to the definition of the grammar. Hereafter, we show an example of the content of the request DEFINE-GRAMMAR using an xml format for a wordspotting grammar.

C->S: MRCP/2.0 ... DEFINE-GRAMMAR 543257
   Channel-Identifier: 32AECB234338@speechrecog
   Load-Grammar-Timeout: 2
   Content-Type:application/srgs+xml
   Content-ID:<request1@form-level.store>
   Content-Length:...
   <?xml version="1.0"?>
   <gramar xmlns=http://www.w3.org/2001/06/grammar
   xml:lang="en-EN" version="1.0">...
   <rule id="restaurant">
   Restaurante
   <one-of xml:lang="en-EN">
   <item>Africa Kine </item>
   <item>Fishers of Men</item>
   <item>Make my Cake</item>
   </one-of>
   </rule>
   </grammar>

S->C: MRCP/2.0 ... 543257 200 COMPLETE
   Channel-Identifier: 32AECB234338@speechrecog
   Completion-Cause:000 success

The RECOGNIZE method from the client to the server signals the recognizer to start speech processing, indicating Recognition mode (Recognition-Mode: continuous wordspotting) in its header. In this case, the tracebackrate has been set at 500 ms.
Confidence-Threshold: 0.9
Recognition-Mode: continuous-wordspotting
Tracebackrate: 500
S->C: MRCP/2.0 ... 543258 200 IN-PROGRESS

Channel-Identifier: 32AECB234338@speechrecog

From this moment on, every time the recognizer detects a keyword of the defined grammar, the server sends a PARTIAL-RESULT event associated to the request RECOGNIZE it belongs to. A status IN-PROGRESS reveals that recognizer is still active.

S->C: MRCP/2.0 ... PARTIAL-RESULT 543258 IN-PROGRESS
Channel-Identifier: 32AECB234338@speechrecog
Completion-Cause:000 success
Content-Type:application/nlsml+xml
Content-Length:...
<?xml version="1.0"?>
<result xmlns="http://www.ietf.org/xml/ns/mrcpv2"
xmlns:ex="http://www.example.com/example"
grammar="session:request1 @form-level.store">
<interpretation>
<instance name="Restaurant">
<ex:Restaurant>
<ex:Name>Africa Kine</ex:Name>
</ex:Restaurant>
</instance>
<input> Africa Kine Restaurant</input>
</interpretation>
</result>

S->C: MRCP/2.0 ... PARTIAL-RESULT 543258 IN-PROGRESS
Channel-Identifier: 32AECB234338@speechrecog
Completion-Cause:000 success
Content-Type:application/nlsml+xml
Content-Length:...
<?xml version="1.0"?>
As a result of any context change, the client decides to modify the grammar by using respective UNLOAD-GRAMMAR and/or DEFINE-GRAMMAR messages as indicated below.

C->S: MRCP/2.0 ... UNLOAD-GRAMMAR 543259
Channel-Identifier: 32AECB234338@speechrecog

S->C: MRCP/2.0 ... 543259 200 IN-PROGRESS
Channel-Identifier: 32AECB234338@speechrecog

C->S: MRCP/2.0 ... DEFINE-GRAMMAR 543260
Channel-Identifier: 32AECB234338@speechrecog

Content-Type: application/srgs+xml
Content-ID:<request1@form-level.store>
Content-Length:...
<?xml version="1.0"?>
<grammar xmlns=http://www.w3.org/2001/06/grammar
xml:lang="en-EN" version="1.0">
<rule id="street">
Calle
<one-of xml:lang="en-EN">
  <item>Downing</item>
  <item>Charlotte</item>
</one-of>
Again, partial results are generated when the recognizer finds any match in the speech channel:

S->C: MRCP/2.0 ... PARTIAL-RESULT 543260 IN-PROGRESS
Channel-Identifier: 32AECB234338@speechrecog
Completion-Cause:000 success

Finally, the STOP method tells the server resource to stop recognition for the current session. The response header section contains an active-request-id-list header field containing the request-id of the RECOGNIZE request that was terminated.
Channel-Identifier: 32AECB234338@speechrecog
S->C: MRCP/2.0 ... 543261 200 COMPLETE
Channel-Identifier: 32AECB234338@speechrecog
Active-Request-Id-List:543258

SIP BYE will de-allocate all the control channels and resources allocated under the session.

C->S: BYE sip:mresources@server.example.com SIP/2.0
Via:SIP/2.0/TCP client.atlanta.example.com:5060;
branch=z9hG4bK74bg7
Max-Forwards:6
From:Sarvi <sip:sarvi@example.com>;tag=1 928301 774
To:MediaServer <sip:mresources@example.com>;tag=62784
Call-ID:a84b4c76e66710
CSeq:323126 BYE
Content-Length:0

Advantages of the invention

The procedure explained in this invention fills the gap of current standard media resource control mechanisms (namely, MRCP by IETF) to manage arbitrarily long audio streams in an efficient manner. Specifically, it provides a new recognition mode for long-lasting speech which does not wait until the end of the audio stream to produce matches, but sends real-time results whenever they are produced.

With this solution, efficiency is notably increased in the control protocol as RECOGNIZE requests are not required to be sent again and again to recognize specific grammars, but a continuous mode is proposed which leads to partial results.

This method is particularly suitable for speech recognition of person to person conversations, where the nature of the audio is theoretically unbounded (unconstrained by any a priori known grammar or format).

Furthermore, the idea of this patent envisages the dynamic adaptation of speech recognition to the context of the conversation in a way that is transparent to the users of the service. The control protocol is modified accordingly, in order to accept dynamic changes of grammars during the course of the conversation.
Also, in order to offer interoperability with other operators and compatibility with Speech Processing Engine providers in the future, standardization of the extension to the protocol would be very interesting.

A person skilled in the art could introduce changes and modifications in the embodiments described without departing from the scope of the invention as it is defined in the attached claims.
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>IETF</td>
<td>Internet Engineering Task Force</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IVR</td>
<td>Interactive Voice Response</td>
</tr>
<tr>
<td>MRCP</td>
<td>Media Resource Control Protocol</td>
</tr>
<tr>
<td>MRCPv2</td>
<td>Media Resource Control Protocol v2</td>
</tr>
<tr>
<td>PBX</td>
<td>Private Branch Exchange</td>
</tr>
<tr>
<td>RFC</td>
<td>Request For Comments</td>
</tr>
<tr>
<td>RTP</td>
<td>Real Time Protocol</td>
</tr>
<tr>
<td>RTSP</td>
<td>Real Time Streaming Protocol</td>
</tr>
<tr>
<td>SDP</td>
<td>Session Description Protocol</td>
</tr>
<tr>
<td>SIP</td>
<td>Session Initiation Protocol</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over IP</td>
</tr>
</tbody>
</table>
REFERENCES


Claims

1.- A method to manage speech recognition of audio calls, said audio calls performed in a Media Resource Control Protocol, or MRCP, based system, said speech recognition carried out by an ASR engine controlled by an MRCP server by looking for a match between an audio stream generated by a user and a compiled grammar, characterised in that it comprises performing said speech recognition continuously by sending, said MRCP server, events regularly to said user when matches are produced, each of said events indicating a partial result of said speech recognition and ignoring unsuccessful matches, stopping said speech recognition when receiving a stop request from said user or when said audio stream finishes.

2.- A method as per claim 1, comprising performing said speech recognition according to an operation mode different from Normal Mode Recognition and HotWord Mode Recognition defined by Internet Engineering Task Force.

3.- A method as per claim 2, comprising indicating said operation mode to said MRCP server by means of existing SET-PARAMS or RECOGNIZE request of MRCP protocol.

4.- A method as per any of previous claims, comprising deciding, an Automatic Speech Recognition module of said MRCP based system, when a match has occurred and sending, said MRCP server, an event every time that a match has occurred.

5.- A method as per any of previous claims, comprising including a parameter in existing SET-PARAMS and/or RECOGNIZE request of MRCP protocol, said parameter indicating maximum time interval at which said MRCP server should check for unreturned partial results.

6.- A method as per any of previous claims, comprising using different compiled grammars while performing said speech recognition by loading, said user, a given grammar by means of existing DEFINE-GRAMMAR request of MRCP protocol and compiling, said MRCP server, said given grammar.

7.- A method as per claim 6, comprising unloading a concrete grammar from said MRCP server when receiving, said MRCP server, a UNLOAD-GRAMMAR request, said UNLOAD-GRAMMAR request being defined for MRCP protocol.

8.- A method as per claim 6 or 7, comprising including a load grammar timeout parameter in existing SET-PARAMS or DEFINE-GRAMMAR request of MRCP protocol, said load grammar timeout parameter indicating maximum time to wait for a reply of a DEFINE-GRAMMAR request.
9.- A method as per claim 8, comprising sending a COMPLETE response of MRCP protocol from said MRCP server to said user if exceeding said load grammar timeout parameter and continuing sending said partial results according to a previous grammar.

10.- A method as per any of previous claims, comprising changing state of said MRCP server, according to MRCP protocol, from recognizing state to idle state only when receiving a STOP request from said user to said MRCP server.

11.- A method as per any of previous claims, comprising establishing a call in a Private Branch Exchange, or PBX, and creating two processing channels between said PBX and said MRCP server, one per each party of said call and each of said two processing channels used to perform said speech processing over audio streams generated by party calls.

12.- A method as per claim 11, wherein said audio flow is arbitrarily long and contains silence intervals.
Figure 1
Figure 3
Figure 4
Figure 6
Figure 7

71: MRCP Client

- 7101: SET-PARAMS
- 7201: 200 COMPLETE
- 7102: DEFINE-GRAMMAR
  Load-Grammar-Timeout: 2s
- 7202: 200 COMPLETE
- 7103: RECOGNIZE
  Recognition-Mode: continuous wordspotting
- 7203: 200 IN-PROGRESS
- 7204: PARTIAL-RESULT IN-PROGRESS
- 7205: PARTIAL-RESULT IN-PROGRESS
  ... 
- 7104: UNLOAD-GRAMMAR
- 7206: 200 COMPLETE
- 7207: PARTIAL-RESULT IN-PROGRESS
- 7208: PARTIAL-RESULT IN-PROGRESS
  ... 
- 7105: STOP
- 7209: 200 COMPLETE

72: MRCP Server
A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/26 H04L29/08
ADD.
According to International Patent Classification (IPC) onto both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
G10L H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal , WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

* 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
  - "D" 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

* If 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

* "Z" document member of the same patent family

Date of the actual completion of the international search 7 February 2013

Date of mailing of the international search report 15/02/2013

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016
Nash, Michael

Authorized officer
**DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>
| A        | US 2009/037176 Al (ARROWOOD JON A [US])  
5 February 2009 (2009-02-05)  
abstract; figure 1  
paragraph [0005] - paragraph [0009]  
paragraphs [0018], [0024] | 1-12                  |
| A        | Anonymous: "Product Support Notice PSN002343u",  
28 July 2009 (2009-07-28), pages 1-6,  
XP055052694,  
Retrieved from the Internet:  
URL:http://downloads.avaya.com/css/P8/documents/100060040  
[retrieved on 2013-02-07] | 6-8                   |

*Note: For the full text, please refer to the actual document or online link provided.*
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 2009037176 Al</td>
<td>05-02-2009</td>
<td>US 2009037176 Al</td>
<td>05-02-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2009038882 Al</td>
<td>26-03-2009</td>
</tr>
</tbody>
</table>