ETSI TR 122 977 V16.0.0 (2020-08)

Digital cellular telecommunications system (Phase 2+) (GSM);
Universal Mobile Telecommunications System (UMTS);
LTE;
Feasibility study for speech-enabled services
(3GPP TR 22.977 version 16.0.0 Release 16)

Reference
RTR/TSGS-0122977vg00

Keywords
GSM,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Legal Notice

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

1 Scope

Speech Enabled Services

The advancement in the Automatic Speech Recognition (ASR) technology, coupled with the rapid growth in the wireless telephony market has created a compelling need for speech-enabled services. Voice-activated dialling has become a de facto standard in many of the mobile phones in the market today. The speech recognition technology has also been applied more recently to voice messaging and personal access services. A Voice Extensible Markup Language (Voice XML) has been designed to bring the full power of web development and content delivery to voice response applications [11]. Voice portals that provide voice access to conventional graphically oriented services over the Internet are now becoming popular. Forecasts show that speech-driven services will play an important role on the 3G market. Users of mobile terminals want the ability to access information while on the move and the small portable mobile devices that will be used to access this information need improved user interfaces using speech input.

Multimodal and Multi-device Services

Speech-enabled services may utilize speech alone for input and output interaction, or may also utilise multiple input and output modalities leading to the multimodal services.

Online access to information is fast becoming a must-have. Along with this trend, come new usage models for information access, particularly in mobile environments. Information appliances in cars such as navigation systems are standard in high-end cars already and this will penetrate lower-end vehicles soon. Data access using mobile phones, though limited and currently estimated to take three years to be widespread, has significant momentum that makes it certain to become widespread. In this new computing paradigm a person will expect to have access to information and interactions in a seamless manner in many environments, be it in the office, at home, in the car, often on several different devices. These new access methods have compelling advantages, such as mobile accessibility, low cost, ease of use, and mass market penetration. They also have their limitations - in particular, it is hard to enter and access data using small devices, speech recognition can introduce mistakes that can sometimes be repeating and therefore blocking the transaction; one interaction mode does not suit all circumstances, and so on.

For example, a recent study of task-performance using wireless phones, such as reading world headlines and checking local weather concluded that currently, these services are often poorly designed, have insufficient task analysis, and abuse existing non-mobile design guidelines. The full report from the field study can be downloaded at [6]. The basic conclusion of this study is that wireless access usability fails miserably; accomplishing even the simplest of tasks takes much too long to provide any user satisfaction. It is thus essential for the widespread acceptance of this computing paradigm to provide an efficient and usable interface on the different device platforms that people are expected to use to access and interact with information.

We can expect and already observe a trend towards a new frontier of interactive services: multimodal and multi-device services.

These services exploit the fact that different interaction modes are good at different things - for example, talking is easier than typing, but reading is faster than listening. Multi-modal interfaces combine the use of multiple interaction modes, such as voice, keypad and display to improve the user interface to services.

Different standard bodies are addressing aspects of this space, driven by several industry proposals: W3C (e.g. MMI activity)[11], OMA/WAP Forum, ETSI [1], IETF[14],…). In particular, the W3C MMI [13] aims at defining a programming model for multimodal and multi-device applications.

Additional details and motivations are discussed in [2, 7, 8].

Overview

A brief overview of the speech-enabled services is presented in Chapter 4. The different ways of enabling speech recognition for the speech enabled services are described in chapter 5. Section 6 discusses multimodal services and options to enable multimodal and multi-device services. The scope of the report, references, definitions and abbreviations are detailed in the first few chapters.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

2.1 Informative references

[1] D. Pearce, "Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition", *Proc. of AVIOS'00*, 2000.

[2] ETSI ES 201 108: "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms; DRS front end".

[3] ETSI ES 202 050: " Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms".

[4] Y. Muthuswamy, P. Walther, "Applications and Requirements", ETSI Aurora DSR Applications & Protocols sub-group, Nov. 10, 2000.

[5] TSG T2010652: "Support of Multi-modal and Multi-device browsers applications by 3GPP".

[6] Nielsen Norman Group: http://www.nngroup.com/reports/wap.

[7] TSG S1-020769-Presentation: Companion presentation to TSG S1-020769.

[8] TSG S1-021274 - "Service Aspects: Multimodal and Multi-device Services".

[9] W3C: http://www.w3.org/TR/nl-spec/.

[10] http://www.w3c.org/2002/ws/.

[11] http://www.w3c.org/Voice/.

[12] Speech Engine Remote Control Protocols by treating Speech Engines and Audio Sub-systems as Web Services - draft-maes-sercp-web-services-00.txt: http://flyingfox.snowshore.com/mrcp_archive/msg00105.html

[13] W3C Multimodal Interaction Working Group: http://www.w3c.org/2002/mmi/.

[14] IETF SPEECHSC – Speech Service Control working group.

[15] D Macho et al. "Evaluation of a Noise-Robust DSR Front-End on Aurora Databases", International Conference on Spoken Language Processing; ICSLP 2002, Denver, CO, Sept 2002

[16] D Pearce, "Developing the ETSI Aurora Advanced Distributed Speech Recognition Front-end & What Next?", IEEE Automatic Speech Recognition and Understanding Workshop; ASRU 2001, Madonna di Campiglio, Dec 2001

2.1 Normative references

[17] 3GPP TS 21.905: "Vocabulary for 3GPP Specifications"

[18] 3GPP TS 22.243: "Speech recognition framework for automated voice services; Stage 1".

[19] 3GPP TS 21.133: "3G security; Security threats and requirements".

[20] 3GPP TS 22.228: "Service requirements for the Internet Protocol (IP) multimedia core network subsystem; Stage 1".

3 Definitions and abbreviations

3.1 Definitions

Automated Voice Services: Voice applications that provide a voice interface driven by a voice dialog manager to drive the conversation with the user in order to complete a transaction and possibly execute requested actions. It relies on speech recognition engines to map user voice input into textual or semantic inputs to the dialog manager and mechanisms to generate voice or recorded audio prompts (text-to-speech synthesis, audio playback,). It is possible that it relies on additional speech processing (e.g. speaker verification). Typically telephony-based automated voice services also provide call processing and DTMF recognition capabilities. Examples of traditional automated voice services are traditional IVR (Interactive Voice Response Systems) and VoiceXML Browsers.

Conventional Codec: The module in UE that encodes the speech input waveform , similar to the encoder in a vocoder e.g. EFR, AMR.

Channel: denotes a particular user agent (browser), device, or a particular modality.

Downlink exchanges: Exchanges from servers and networks to the terminal.

DSR Optimised Codec: The module in UE which takes speech input, extracts acoustic features and encodes them with a scheme optimised for speech recognition. This module is similar to the the conventional codec (e.g. AMR). On the server-side, the uplink encoded stream can be directly consumed by speech engines without having to be converted to a waveform.

Haptic interface: An interface that allows a user to interact by receiving feed back achieved by applying a degree of opposing force to the user along the x, y, and z axes (e.g. pressure).

Mono-modal application: application designed for access through only one channel or channel type (e.g. WAP, Web or Voice exclusively).

Multi-channel application: applications designed for ubiquitous access through different channels, one channel at a time. No particular attention is paid to synchronization or coordination across different channels.

Multi-device applications: denote application that supports the capability to interact with a particular application over a number of physical devices with browsers being synchronised with the MT accessing 3G services. These browsers may support the same (e.g. GUI) or different modalities.

Multimodal application: denotes application that supports more than one interaction mode by relying on a combination of multiple input (e.g. key, stylus, voice, …) to access and manipulate information on the move and to enable the most convenient output (display, tactile, audio) at the discretion of the user/ terminal capability.

Multimodal browser: "declarative" user agent that renders different modalities (e.g. GUI browser and Speech Browser) simultaneously available (concurrently or sequentially) and synchronized for the user to interact with the application. These different user agents may be located on the same device or distributed on among UEs or across the 3G network.

Speech Recognition Framework: A generic framework to distribute the audio sub-system and the speech services by sending encoded speech between the client and the server. For the uplink, it can rely on conventional (ASR) or on DSR optimised codecs where acoustic features are extracted and encoded on the terminal.

SRF Call: An uninterrupted interaction of a user with an application that relies on SRF-based automated voice services.

SRF Session: Exchange of audio and meta-information, explicitly negotiated and initiated by the SRF session control protocols, between terminal (audio-sub-systems) and SRF-based automated voice services. Sessions last until explicitly terminated by the control protocols.

SRF User Agent: a process within a terminal that enables the user to select a particular SRF-based automated voice service or to enter the address of a SRF-based automated voice service. The user agent converts the user input or

selection into a SIP IMS session initiation with the corresponding SRF-based automated voice service. The user agent can also terminate the session with the service when the user device to disconnect.

Tactile interface: Touch-based interface.

Uplink exchanges: Exchanges from the mobile terminal to the server / network.

User agent: The component that renders the presentation data into physical effects that can be perceived and interacted with by the user. For a given modality this may be a separate browser or platform or one or multiple components internal to a browser or platform.

3.1 Abbreviations

For the purposes of this document the following abbreviations apply:

AMR	Adaptive Multi Rate
DOM	Document Object Model
DSR	Distributed Speech Recognition
DTMF	Dual Tone Multi-Frequency
ELRA	European Language Resource Association
GUI	Graphical User Interface
IP	Internet Protocol
IMS	IP Multimedia Subsystem
MIT	Meta-information Transport
MMSP	Multimodal Synchronization Protocol
MVC	Model View Controller
SDP	Session Description Protocol
SIP	Session Initiation Protocol
SOAP	Simple Object Access Protocol
SPEECHSC	Speech service Control
SRCP	Speech Remote Control Protocol
RTP	Real Time Streaming Protocol
URI	Uniform Resource Identifier

4 Speech-Enabled Services

Most traditional telephony-based speech-enabled applications fall into one of the following groups:

1. Information applications : Here the user queries the service to retrieve some information from a remote database. Examples of this type of service include voice portals which provide weather reports, restaurant information, stock quotes, movie listings etc.

2. Transaction-based applications: Unlike the information applications, here the user calls the service to execute specific transactions with a web server. Examples of this type of service include financial transactions (stock trading), travel reservations, e-commerce etc.

Depending on the modalities used for user interaction with the service, speech-enabled services can be divided into speech services or multimodal/multi-device services. This is illustrated in Figure 1. As the name implies, speech-only services utilise only the speech modality for both user input and output. These services are especially suited to the smaller size wireless devices in the market today. These devices have smaller screens and smaller or difficult-to-enter keyboards and are becoming increasingly difficult for GUI applications.

With respect to Figure 1, it is possible to design speech-enabled services that alternate or combine the use of client-side only engines and servers-side speech engines.

4.1 Application Scenarios

In a typical transaction application scenario, involving speech as input and output modalities, the system may prompt the user to login using some ID or password. Challenge and feedback are provided via voice. The application can then guide

the user through the menus to provide the data required for the transaction. Once the user submits the data, the system completes the transaction and may provide feedback to the user using pre-recorded audio information or synthesized speech depending on the data. While typical, this is only a particular example and numerous other and richer scenarios be considered.

5 Multimodal Services

Speech, however, has its own limitations. It is serial and the user may find it difficult to remember long outputs. Speech output may sometime be unnatural sounding. Even though, the speech recognition technology has matured over recent years, it is not a perfect technology and the recognition systems are still prone to errors, particularly in adverse operating conditions. This is especially problematic if the mistakes are repeated and block completion of an application (transaction, query, …). It is thus obvious that each of the modalities, speech or visual, have their pros and cons. The use of multiple modalities can yield a synergistic blend in which the strengths of each modality are used to overcome the weaknesses of a mono-modal interaction, to result in more efficient user interfaces as compared to the mono-modal ones. Some of the advantages of multimodal interaction include:

- Easy entry and access of data in wireless devices by combining multiple input & output modes (concurrently or sequentially)
- Choosing the interaction mode that suits the task and the circumstances
  - Input: key pad or keyboard (including DTMF), touch, stylus, voice, joystick, haptic / tactile, video...
  - Output: display, haptic/tactile, audio, multimedia...
- Enabling use of several devices in combination by exploiting the resources of multiple devices.

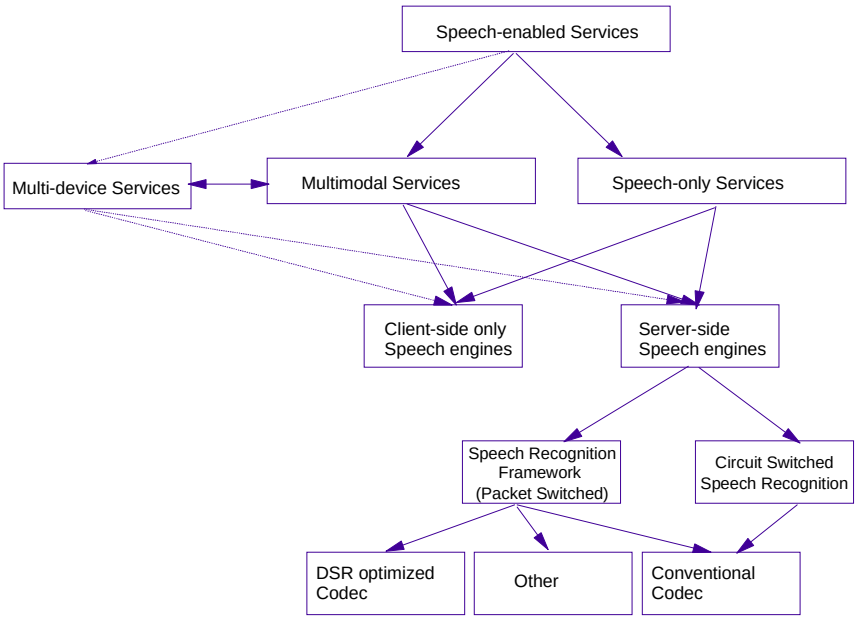

Figure 1 Chart illustrating the different kinds of speech-enabled services and the different methods available to implement speech recognition technology needed to enable these services

The features available for multimodal and multi-device services will evolve; as will the notion of multimodality. GUI is typically multimodal (display, keypad or keyboard and pointer); but it is now widely considered as a particular channel characterized by well established GUI programming patterns.

The combination and synchronization of voice and GUI (including multimedia output) as well as the synchronization of different devices (multi-device) are expected to be deployed first. Handwriting / pen input and multimedia output will probably follow soon after. We may have to wait longer for large volume deployments of more exotic capabilities like video and haptic/tactile inputs or output.

5.1 Application Scenarios

An example of a multimodal airline reservation service, involving both speech and GUI, is shown in Figure 2 and discussed in [5].

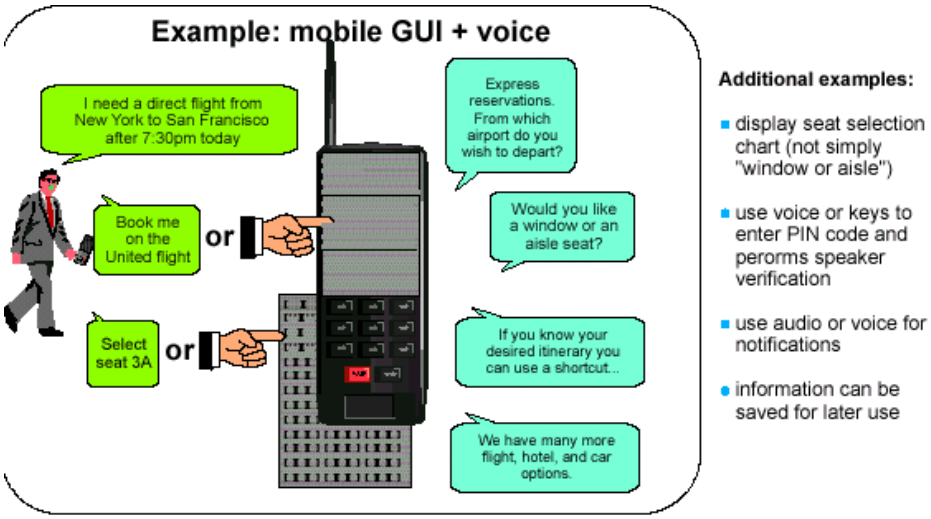

Figure 2 - Example of a multi-modal scenario

Additional discussions of requirements for SES applications can be found in [4].

6 Speech recognition technology

Many of the speech-enabled services require a speech recognition engine to decode (i.e. interpret) speech input from the user (what is being said by the user). Similarly, speech enabled services require speech synthesis engine (TTS engine) to generate output prompts. Other engines may typically be required by speech enabled services, like speaker recognition (enrolment, identification, verification, natural language (NL) parsers, NL dialog managers, prompt generators etc..) ‘

As shown in Figure 1, there are three different ways by which speech recognition can be implemented for a speech-enabled service:

1) - Client-side only speech engines (i.e. terminal based),

2) - Server-side speech engines (i.e. network based). This can itself be subdivided into:

   - Circuit switched-based.

   - Packet switched-based speech recognition framework that supports exchange of encoded speech and meta-information [18]:

     - with conventional codecs (e.g. AMR)

     - with DSR optimized codecs (Distributed Speech Recognition) [1]

These different implementation methodologies are explained in the next paragraphs. To understand the difference between these implementations, it is necessary to understand the basic speech recognition process, which can be divided into two modules:

- Feature extraction (front-end): This involves the conversion of input speech into a set of features that are relevant for recognition of speech.

- Recognition algorithm (pattern matching) (Back-end): this module constitutes the real recognition process that matches the input speech features against one of the stored set of models and provides recognition results based on the active speech grammar and vocabulary. The front-end algorithm typically is a computationally simple algorithm relative to the recognizer and hence often completely masked by the recognizer in terms of the resource requirements.

The speech-driven applications include simple terminal based applications like voice dialling and command and control applications with limited vocabularies that facilitate the speech recogniser to be implemented solely in the terminal. However, more demanding applications like the dictation, time-table enquiry systems, street navigation etc require a complex speech recognition systems that would need lots of memory and computational resources - items that are scarce in today's portable devices. Hence these applications require part or whole of the speech recognition process to be carried out in the network, which can accommodate bigger and more complex computational devices.

Note that other considerations may lead to prefer using server-side processing: when the task is too complex for the local engine, when the task requires a specialized engine, when it would not be possible to download the speech data files (grammars etc…) without introducing significant delays or taking too much bandwidth or when intellectual property (e.g. proprietary grammars), security or privacy considerations (e.g. it would be inappropriate to download a grammar or vocabulary file that contains the names of the customers of a bank or the password grammars) make it not appropriate to download such data files on the client or to perform the processing on the client.

In a network-based system (Figure 3), the conventional circuit switched speech channel is used for the transmission of speech and the complete speech recognition processing – both the feature extraction and recognition - is done at the network side. At the terminal side, speech spoken by the user is encoded using conventional speech coders (e.g. AMR).

The speech recognition framework (SRF) enables to distribute the audio sub-system and the speech services by sending encoded speech and meta-information between the client and the server over packet switched network. It is illustrated in Figure 4. The SRF may use conventional codecs like AMR or Distributed Speech Recognition (DSR) optimized codecs.

The SRF can be deployed over a packet switched (PS) network; like, but not only, on IMS [20]. Over a generic PS network, SRF will require:

- Uplink and downlink transport of audio (e.g. RTP)
- Session establishment, signalling and control
- Codec negotiations
- Quality of service negotiation and provisioning
- Service administration.

Such features and services can be automatically exploited on IMS and adapted to the specificities of SRF. SRF on IMS can rely on and extend the IMS protocol stack.

In the particular case of the DSR approach, the terminal hosts the feature extraction module, while the recognition is done in the server. The speech features are usually compressed to reduce the transmission bit rate, error protection is added and the resulting data stream is transmitted through error protected data channels.

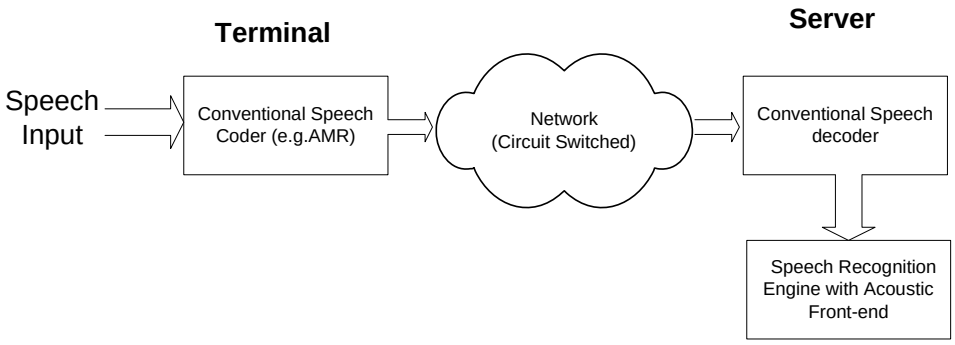

Figure 3 - Illustration of Network Based Circuit Switched Speech Recognition

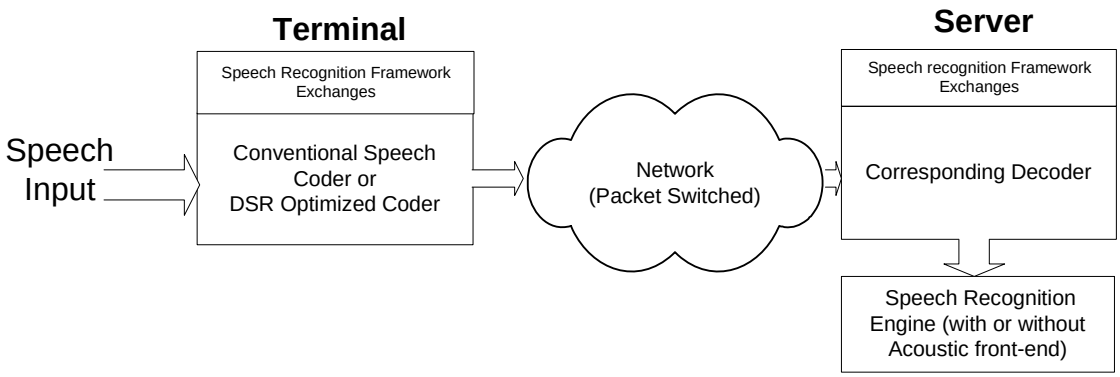

Figure 4 - Ilustration of Speech Recognition Framework

Figure 5 shows the different components of a typical speech enabled service, which may use either conventional or DSR optimized codec and is explained below:

1) Voice Platform: The voice platform hosts the network side components needed for enabling a speech service.
2) Speech Recognizer: Speech recognizer includes the speech recognition back-end in the case of DSR or a complete speech recognizer in the case of a conventional codecs.
3) Text-to-Speech: This module converts the text into speech to be sent as speech output to the user.
4) Audio-Playback and Record: This module facilitates the audio prompts to be played back to the user.
5) Voice Browser: The voice browser interprets the voice dialog between the user and the speech service (e.g. VoiceXML Browser).
6) Call Control: The call control module handles the call control functions needed for the speech service.
7) Content Server: The content server hosts the web content, which is accessed by the voice platform through HTTP.

Note that additional modules may be needed to enable multimodal services.

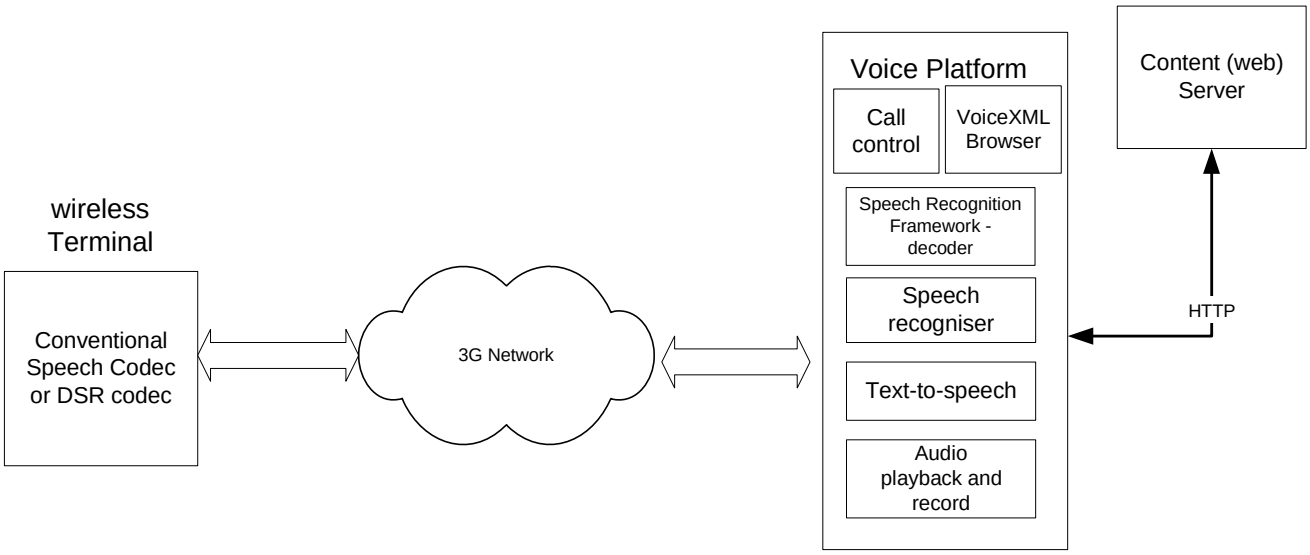

Figure 5 - Illustration of the architecture for a typical speech service in a 3G network using network side speech recognition resources based on the speech recognition framework (SRF)

6.1 DSR standards

ETSI STQ-Aurora working group have developed two DSR Front-ends:

1) ES 201 108 [2] was published in Feb 2000. It is based on the popular Mel-Cepstrum feature extraction that is extensively used in speech recognition systems.

Front-end	Computation (wMOPS)	RAM (kwords)	ROM (kwords)
ES 201 108	3.1	0.6	2.1
ES 202 050	11.7	3.83	3.75

2) ES 202 050 [3], the Advanced DSR front-end, was selected in Feb 2002 and will be published in September 2002. It provides improved robustness in background noise giving significant reduction in speech recognition word error rate compared to the Mel-Cepstrum in noise [15].

The requirements for the advanced front-end have been defined to meet the needs of user in a mobile environment. These requirements are described in [16] and a competitive selection was conducted to find the best front-end technology. A set of evaluation databases have been established and used for the characterisation of the recognition performance of these front-ends. These databases are now publicly available via ELRA and cover both small vocabulary and large vocabulary tasks in a range of noises typical of those found in mobile environment. As part of this, the front-ends have been tested on 5 languages (Finnish, Spanish, German, Danish, and Italian) from databases collected in a car environment [16].

In addition to the front-end feature extraction, these standards define a compression algorithm to achieve a data rate of 4.8kbps and a server side error mitigation to maintain recognition performance under channel errors.

Examples of the computational complexities of the two standards are estimated in table 1.

Table 1: DSR Terminal side Complexity

[Editor's note: publicly available references to the information presented would be valuable]

7 Multimodal and Multi-device Technology

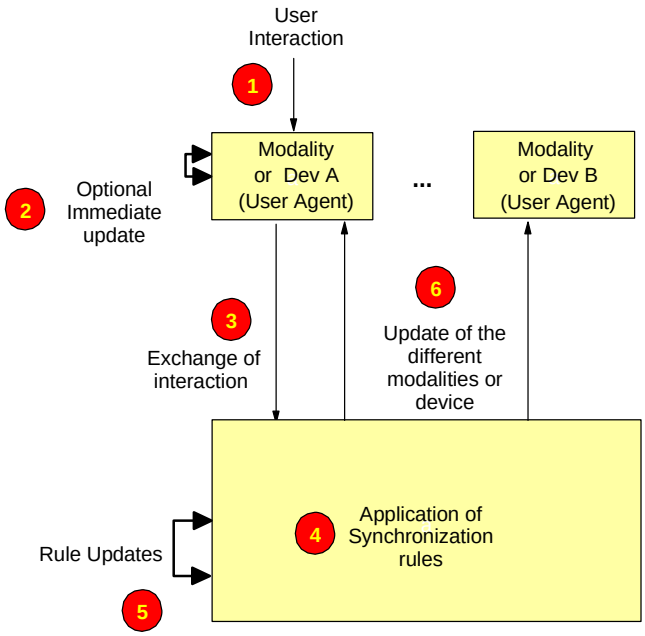

Figure 6 - Execution model of multimodal or multi-device applications; independent of programming model or configuration.

7.1 Execution Model

The proposed execution model of multimodal and multi-device applications is discussed in [8]:

- A user interaction in one of the available modalities (user agent) results into handling a representation of the interaction event to determine of the impact of the interaction via synchronization rules. This in turns results into updates of the synchronization rules and all the registered modalities (user agents) available to the user.

This is summarized in Figure 6, where each user agent may represent a different modality (e.g. VoiceXML browser and XHTML-MP browser or GUI and Voice Java applications) or different devices (e.g. smart phone and PDA or kiosk).

The user action may or may not result into an immediate update of the affected modality state prior to the synchronization.

Note also that Figure 6 does not address the steps internal to the user agents. For example, a voice or handwriting user agent will interface with local or distributed speech engines to process input and generate outputs.

This execution model supports the different authoring approaches that have been proposed for authoring multimodal and multi-device services [8].

The following is an example of a basic architecture that implements the execution model (Figure 7).

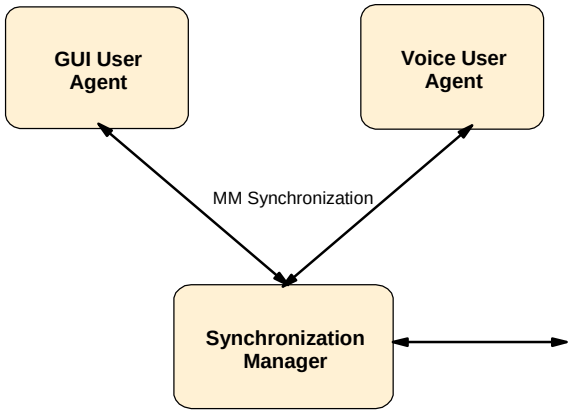

Figure 7 – Example of a basic architecture that supports multimodal and multi-device interactions illustrated for voice and GUI interaction.

In distributed cases, voice can be exchanged based on SRF on IMS [18].

The synchronization manager is responsible for the exchange of synchronization messages (interaction events and presentation updates) using the multimodal synchronization interfaces and protocols. The synchronization logic can also be responsible for functions like:

- Ordering and possible composition of inputs from several modalities
- Maintenance of the state and session context (i.e. history of interaction and state changes) of the application.

Dialog management i.e. to process user inputs to interpret the intent of the user and generate an output, presentation update or action resulting from a complete input. To do so, a dialog manager relies on strategies / algorithms to determine focus and intent based on the session context and possibly external knowledge sources as well as based on disambiguation, correction and confirmation sub-dialogs.

7.2 Deployment configurations

Examples of multimodal or multi-device configurations are provided in Figure 8 to Figure 13. It can be deployed under numerous configurations. Except for the sequential configuration, they can support any synchronization granularity authorized by application, network or user.

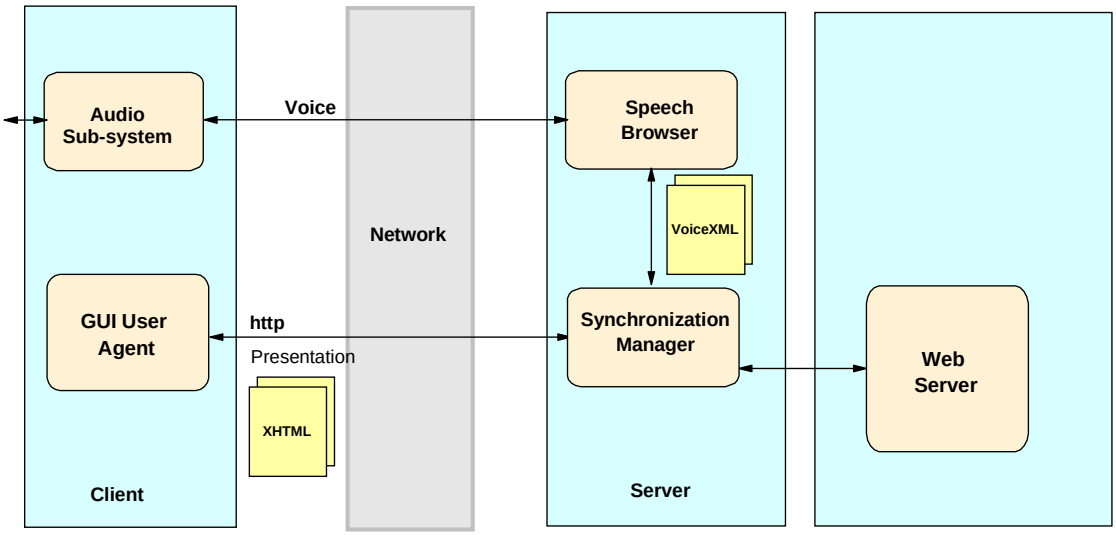

Figure 8 – Example of sequential configuration (no voice and data support simultaneously) for voice and GUI interaction. This configuration does not require IMS (SRF): it can be deployed on 2G or 2.5G networks. Only one modality is available at a given moment. The user may switch at any time or when allowed or imposed by the application.

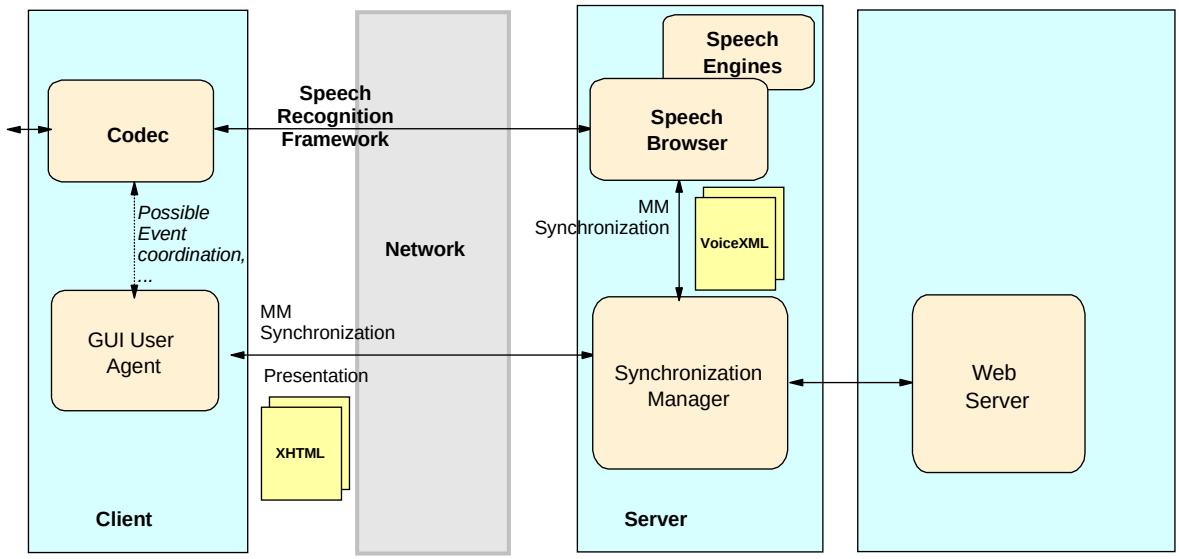

Figure 9 – Example of Thin Client Configuration (voice and data support) - Server-side speech engines local to speech browser for voice and GUI interaction.

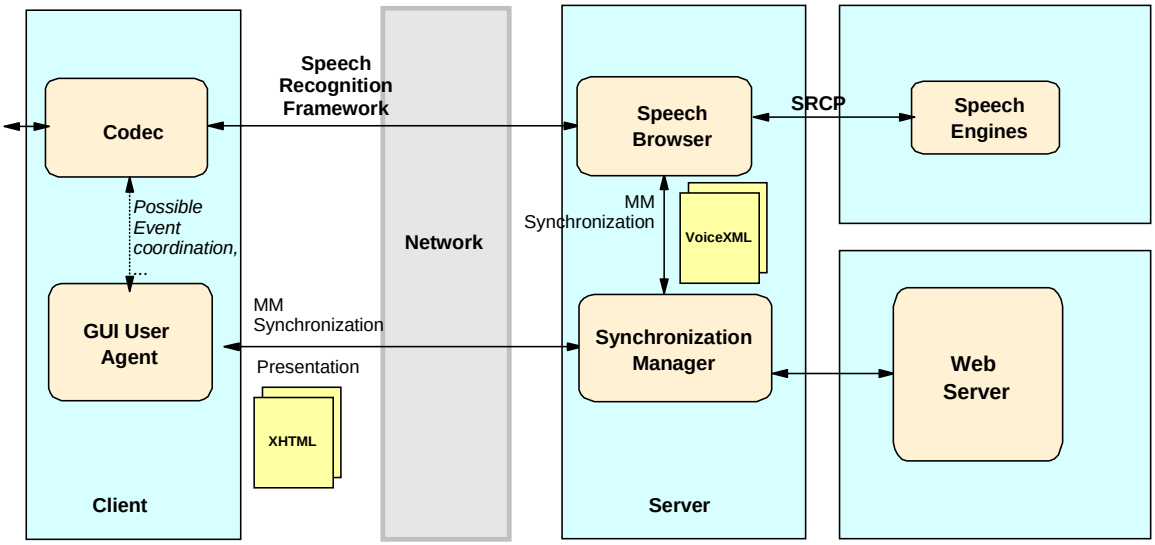

Figure 10 – Example of Thin Client Configuration (voice and data support) - Server-side speech engines remote with respect to speech browser for voice and GUI interaction.

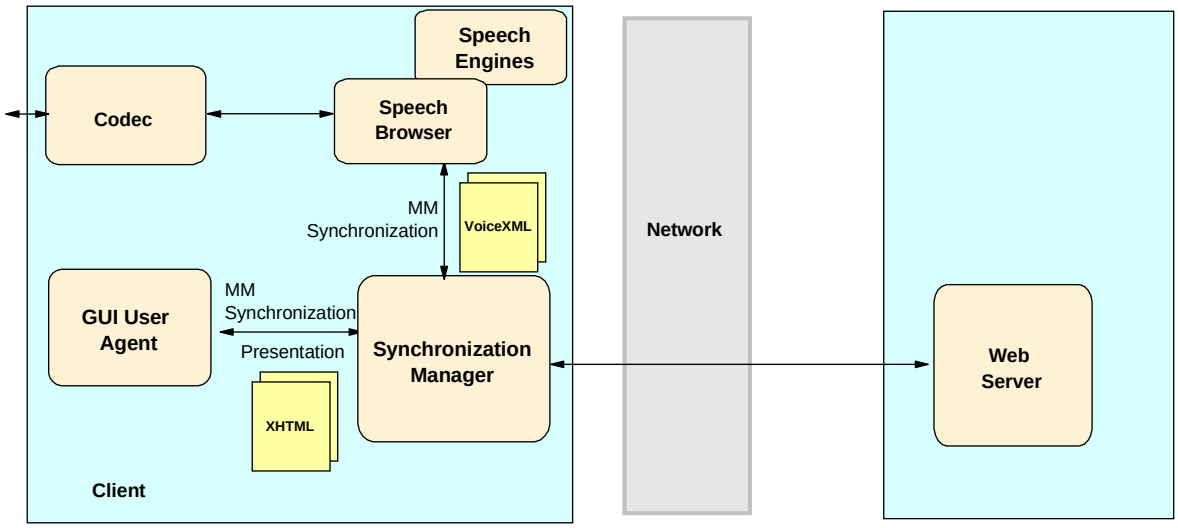

Figure 11 – Example of Fat client configuration with local speech engines for speech and GUI interaction. This can be the internal architecture of a browser implementation.

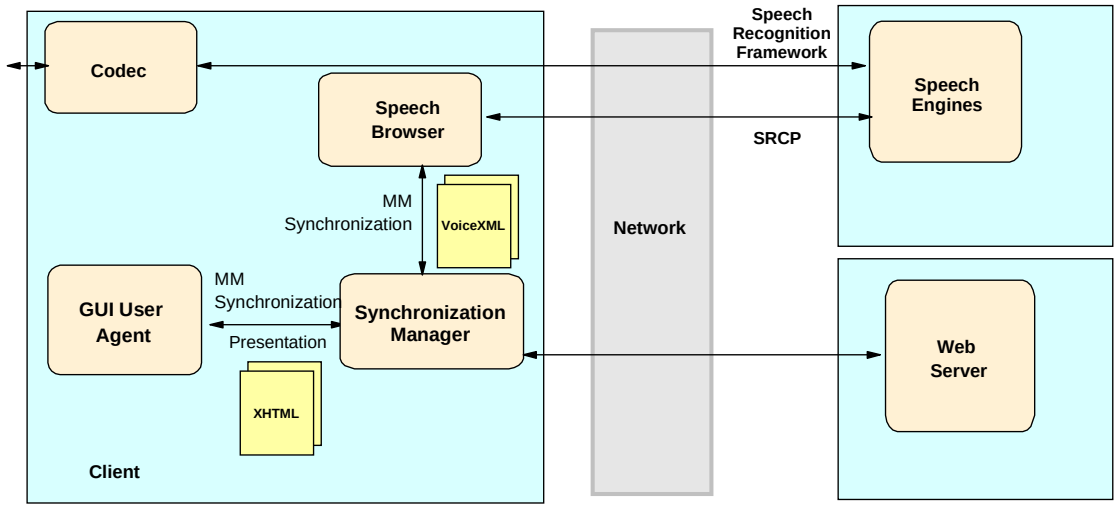

Figure 12 - Example of Fat client configuration with server-side speech engines for speech and GUI interaction. The speech engines are remote controlled by SRCP.

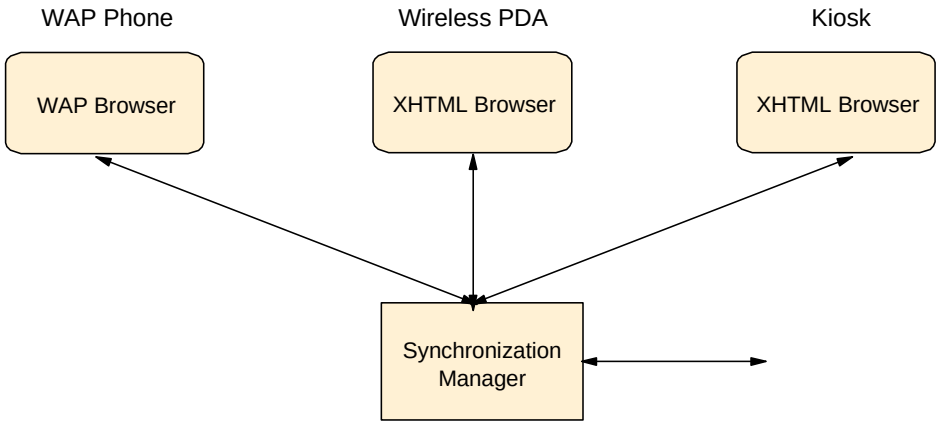

Figure 13 – Example of Multi-device configuration.

Configurations as illustrated in Figure 10 and Figure 12 require speech engine remote control APIs or protocols. This could be based on the SRCP specifications currently developed by the IETF SPEECHSC activity [14]. As no other activity addresses these issues, the support of such distributed configurations introduces a dependency on IETF.

7.3 Authoring

The W3C MMI working group is currently developing specification for declarative (i.e. XML-based) authoring of multimodal and multidevice applications [13].

8. Requirements to introduce Speech-enabled services

This section provides the high level requirements to enable speech and multimodal / multi-device services. Users of the Speech service shall be able to initiate voice communication, access information or conduct transactions by voice commands. Multimodal interaction will utilise other modalities depending on the UE and network capabilities.

The speech-enabled service will be offered by the network operators and will bring value to the network operator by the ability to charge for these services. Speech recognition Framework-enabled services shall be compatible with the IMS. However, it is possible that aspects of the speech service will be available without requiring IMS.

8.1 Initiation

8.1.1 Service initiation

It shall be possible for a user to initiate a connection to the speech-enabled service, for example, by entering the identity of the service. The identity used will depend on the scheme of the service provider but could include a phone number, an IP address or even a URI.

Multimodal and multi-device services will allow URI-based initiation entered in the terminal user agent by the user (e.g. entered address or selected icon or bookmark).

8.1.2 Multimodal or multi-device access configuration

In addition, multimodal and multi-device services, when in distributed configurations, may require a registration of the different user agents / devices that will participate to the interaction with the service. This may include a "definition" of the role of the different user agents and devices. This step may take place prior to accessing the service and be maintained across services or it may take place after initiating a service or while interacting with the service. This step may involve clock synchronization for correct relative time stamping of the interaction events and presentation manipulations.

8.2 Information during the interaction session

This may be motivated by the expected or observed acoustic environment, the service package purchased by the user, the user profile (e.g. hands-free as default) or the service need. In the case of a speech service, the user speaks to the service and receives output back from the service provider as audio (recorded ‘natural’ speech) or Text-to-Speech Synthesis. The output from the server can be provided in the downlink as a streaming service or by using conversational speech codec. Additional modalities may be involved in a multimodal service depending on the capabilities of the client device.

Additionally, it shall be possible to exchange control and application specific information during the call between the client and the service (e.g. speech meta-information). Accordingly some terminals shall support sending additional data to the service (e.g. keypad information and other terminal events) and receiving data feedback from the server that shall be displayed on the terminal screen.

With multimodal services where synchronization is distributed, it will be necessary to exchange synchronization information:

- Events that result from the user interaction (possibly time stamped) along with possible meta-information about the events (e.g. see NLSML [9]).
- Presentation manipulation instructions (as events or presentation mutations instructions). Details are provided in [8].

The QoS for these exchanges shall be the highest available (conversational) to minimize any synchronization delay felt by the user.

In addition, when the user change the configuration of his or her multimodal or multi-device user agents (e.g. by adding or removing a device), the necessary exchanges to support registration, de-registration, change of role etc.. must take place.

Eventually, replication of the state of the synchronization manager may require additional exchanges (multi-device configurations or hybrid cases, e.g. case where the configuration can switch between thin and fat client configuration (e.g. between Figure 9 and Figure 11).

Eventually, configurations as illustrated in Figure 10 and Figure 12 require the exchanges of the messages required to remote configure and control the server-side speech engines (e.g. SPEECHSC [14]).

8.3 Control

It shall be possible for network operators to control access to services based on subscription profile of the callers.

SRF-based automated voice services may be provided by the network operator (home or visited) or by third parties.

The administration of the speech or multimodal/multi-device services (authorization, deauthorization, registration, deregistration, activation, deactivation) shall be under the control of the network operator. But when decided to do so by the network operator, it should be possible to the third part providers to administer the speech or multimodal/multi-device services themselves through the gateway that they would connect to IMS. In such case, the third part provider performs all the administrative steps and no registration would be required with the network operator.

It shall be possible to use speech or multimodal/multi-device sessions in order to provide access to the corresponding services. For example applications might use a speech or multimodal/multi-device session to access and navigate within and between the various automated services by spoken dialogs or multimodal/multi-device interactions.

8.4 User perspective (user interface)

The user's interface to this service shall be via the UE. User can interact by spoken and keypad inputs. The UE can have a visual display capability. When supported by the terminal, the server-based application can display visual information (e.g., stock quote figures, flight gates and times) in addition to audio playback (via recorded speech or text-to-speech synthesis) of the information. Depending on the terminal capabilities, other modalities can also be supported.

Feedback is provided via spoken speech (recorded or prompted), displays (GUI) when available on the UE or other output modality supported by the UE.

8.5 Service provisioning

Speech or multimodal/multi-device service shall be able to be provisioned by either the network operator (roaming or home) or by a 3rd party service provider.

It shall be possible for network operators and 3rd party service providers to offer speech or multimodal/multi-device services by providing identity of service, such as a phone number, an IP address or a URI that the user can say, enter or select on the terminal.

8.6 Security

The "Security Threats and Requirements" specified in 21.133 [17] shall not be compromised.

It shall be possible to deny unauthorized access to the 3GPP speech or multimodal/multi-device services. An authorization may be based on the following,

- identity of the accessing user agent, server or device
- the destination user, device or user agent

Third parties shall have authorization from the User and PLMN Operators in order to access the speech or multimodal/multi-device service.

In the case of distributed multi-modal or multidevoice browser, the exchange of interaction events and client manipulation raise the following security issues that must be addressed:

- Interaction events and presentation manipulations can be intercepted by unauthorized third parties. This would enable reconstruction of the complete interaction with the application; especially in between submits to the backend. Any note, temporarily selections etc would be accessible!
- Unauthorized third parties may be able to issue presentation manipulations that would affect the user agent.

Mechanisms shall be considered to securely exchange multi-modal synchronization. This may require:

- Encryption of the exchanged information.
- Presentation manipulation shall be accepted only from trusted / authored parties. Mechanisms to achieve this shall be supported. Again encryption may be an adequate mechanism to address these issues.

In the case of client-based speech engines, additional security (and privacy) issues arise when the application is downloaded from a third party service provider:

- The speech data files (acoustic model grammars, language models, vocabularies, NL parser data files, etc…) sent to the client may contain proprietary or sensitive information (e.g. passwords, list of customers and associated input information, propprietary grammar, …).
  - The data files may be intercepted by un-authorized third parties or tampered with in the UE.
  - This may relate to the Digital Right Management work items.
- Results of some client-side engine sent across the network can be tampered with or intercepted on the UE or when transmitted.

These issues shall be addressed by appropriate mechanisms or by requiring server-side engines when needed.

8.7 Privacy

Speech or multimodal/multi-device privacy requirements shall be at least as good as for IMS voice or data sessions [18]:

- It shall be possible to encrypt speech and speech meta-information exchanges
- It shall be possible to prevent exchange of the user's true identity, location and other terminal or user related information when required.

Speech or multimodal/multi-device services may imply that the service provider collects information about the user or usage. This information should be treated according to the policies in place for data and voice (e.g. human to operator or human to automated service) services. The speech or multimodal/multi-device services shall not add additional privacy risks.

In the case of multimodal or multi-device services, interaction events enable reconstruction of the complete interaction with the application, including in between submission to the backend and therefore possibly beyond the knowledge or control of the user. This information or aspect of it may be considered as private by the user. Therefore, it is important that the multimodal synchronization be associated to schemas that let the user specify the use that can be done of the information, beside synchronization. Multi-modal services shall produce similar schemas to describe their handling and use of the information. Trust and resolution mechanism should be provided to enable the user to accept the particular service and configuration on the basis of the usage that will be made of such information or the management options provided to the user.

Privacy of user shall not be threatened when exchanging speech data files across the wireless network or by storing them on a UE.

8.8 Charging

The user can be charged for sessions with the speech or multimodal/multi-device service in a variety of ways. The following shall be possible:

a) a) By duration of session (including "one-off" charge/flat rate)

b) b) By data volume transferred (number of packets) or other similar criteria.

c) c) By subscription fees for the service (unlimited usage or unlimited usage up to a point and then per-use fees)

d) d) Free (e.g. with the service being subsidised by advertising revenue from advertisement spots).

Speech or multimodal/multi-device services shall be available to pre-paid and post-paid subscribers.

9 Impact on the 3GPP system

9.1 Speech Recognition within 3GPP system

There are no speech-enabled services that are exclusively specific to the speech recognition technologies described earlier. However, there are some requirements for enabling these speech services in a wireless system, based on the available technologies.

Speech services utilising terminal based speech recognition like voice dialling do not require any special resources in the wireless network other than possibly accessing and downloading the applications (e.g. VoiceXML pages) and associated engine data files (e.g. grammar, language model, acoustic models etc…) and the resulting delays, privacy and security issues afore mentioned. These considerations may impact the terminal as well as services that provided client-side speech engine-based speech services. A technical specification may be needed to address these issues. However, it is expected that this will be covered in a technical specification that would address multimodal and multi-device services

Server-side speech engine relies on network based resources.

Non SRF-based server-side automated voice services performs speech encoding by conventional speech coders, which are already included as part of the 3GPP specifications. Hence there are no additional requirements, specific to this technology, which requires standardisation for enabling such services in the 3GPP system. It is to be noted that speech services relying on network based speech recognition exist today and can be accessed using current wireless terminals. At this time, these typically provide informational services like weather, sports, news information etc. These services currently utilise a circuit switched voice connection to send encoded speech.

Part of the speech recognition processing in DSR based speech recognition, unlike the network based recognition, takes place in the terminal. There are some recognition technology specific requirements that are to be satisfied for DSR, which include the introduction of a uplink optimised DSR codec in the UE.

SRF-based automated voice services rely on PS and shall preferably rely on IMS with specification of the control and exchange of speech meta-information. This also impacts the terminal to enable these exchanges. When relying on DSR-optimized codecs, these shall be supported by the audio sub-system on the UE.

A separate technical specification outlining the additional changes is required to support speech-enabled services based on SRF [18]. A possible architecture view of the impact is schematically presented in Figure 14.

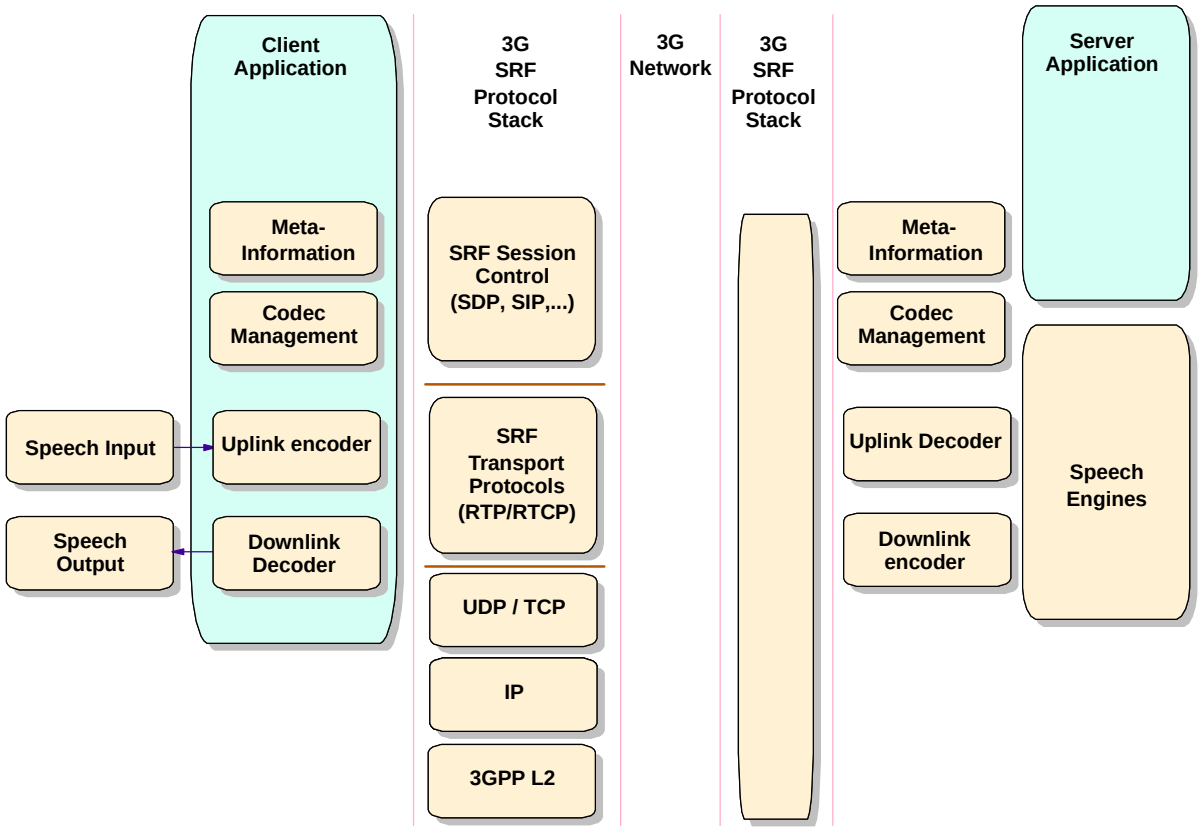

Figure 14 – Possible architecture for integrating SRF framework within 3GPP protocol stack and architecture.

9.2 Multimodal and Multi-device Services within 3GPP system

A separate technical specification outlining the additional changes is required to support multimodal and multi-device services. It is expected to also cover the requirements of applications with client-side speech engines.

A possible architecture that provides basic support of multimodal and multi-device services within 3GPP system is presented in Figure 15.

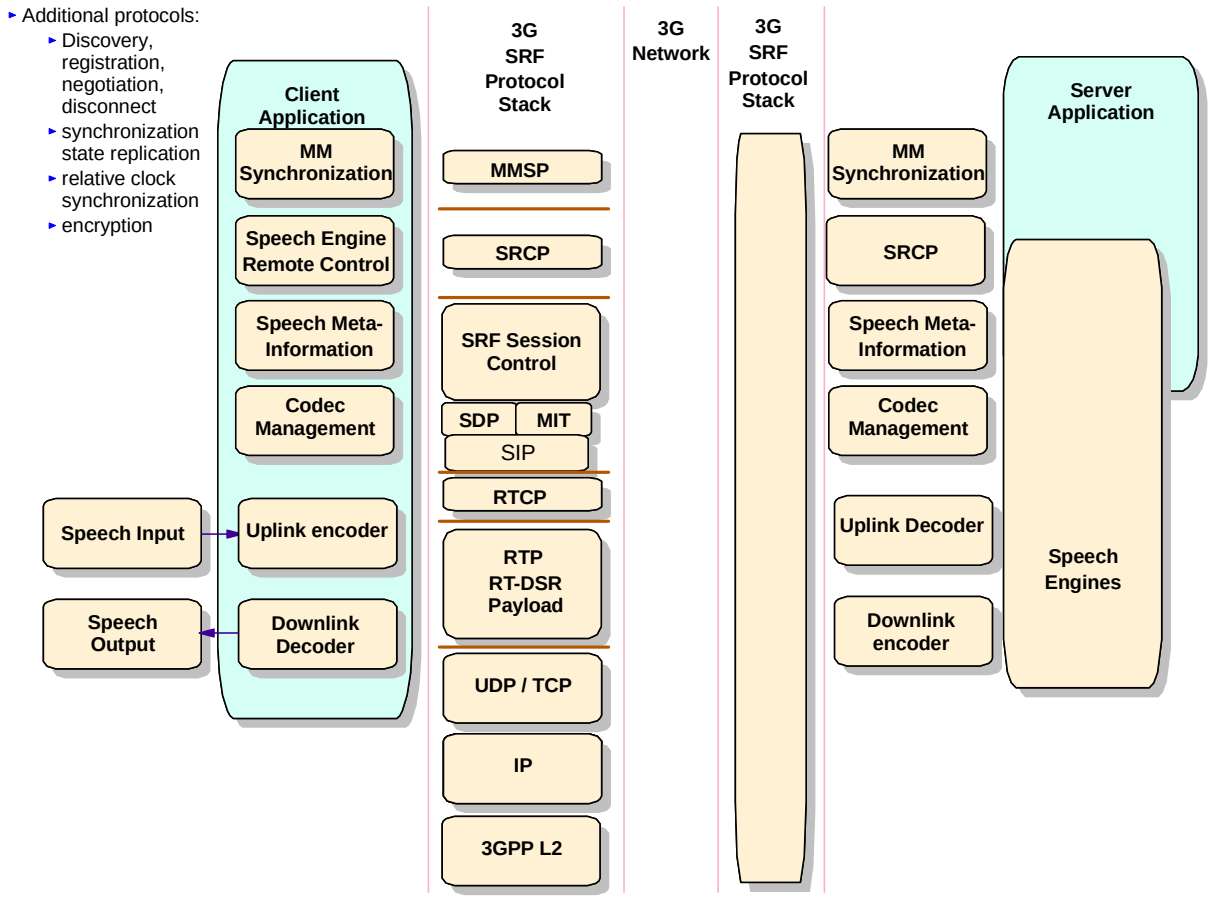

Figure 15 – Possible architecture for integrating multimodal and multi-device services within 3GPP.

MIT designates Meta-information Transport which can be exchanged on SOAP or RTP (e.g. as part of the payload or with dynamic payload switches).

MMSP designates Multi-modal Synchronization protocols

SRCP (speech engine remote control protocol) helps to remotely control the speech engines, if they are located away from a Voice browser. This protocol is addressed currently by IETF SPEECHSC working group [14].

Annex A:
Change history

Change history											
TSG SA#	**SA Doc.**	**SA1 Doc**	**Spec**	**CR**	**Rev**	**Rel**	**Cat**	**Subject/Comment**	**Old**	**New**	**Work Item**
SP-17	SP-020570		22.977			Rel-6		Presented to SA #17 for approval	2.0.0	6.0.0	
SP-36			22.977			Rel-7		Updated to Rel-6 to Rel-7	6.0.0	7.0.0	
SP-42	-	-				Rel-8		Updated from Rel-7 to Rel-8	7.0.0	8.0.0	
SP-46	-	-	-	-	-	-	-	Updated to Rel-9 by MCC	8.0.0	9.0.0	
2011-03	-	-	-	-	-	-	-	Update to Rel-10 version (MCC)	9.0.0	10.0.0	
2012-09	-	-	-	-	-	-	-	Updated to Rel-11 by MCC	10.0.0	11.0.0	
2014-10								Updated to Rel-12 by MCC	11.0.0	12.0.0	
2015-12	-	-	-	-	-	-	-	Updated to Rel-13 by MCC	12.0.0	13.0.0	
2017-03	-	-	-	-	-	-	-	Updated to Rel-14 by MCC	13.0.0	14.0.0	
2018-06	-	-	-	-	-	-	-	Updated to Rel-15 by MCC	14.0.0	15.0.0	
SA#88e	-	-	-	-	-	-	-	Updated to Rel-16 by MCC	15.0.0	16.0.0	

History

Document history		
V16.0.0	August 2020	Publication