

---

**TSG-SA WG1 #17**  
**Durango, USA, 12-16<sup>th</sup> August 2002**

**S1-021846**  
**Agenda Item: SES**

---

**Title:** LS on Speech Enabled Services

**Response to:**

**Release:** Release 6

**Work Item:** SES

**Source:** SA1

**To:** SA2

**Cc:** SA3, T2

**Contact Person:**

**Name:** Jerome Vogedes

**Tel. Number:** +1-847-523-1796

**E-mail Address:** [Jerome.Vogedes@motorola.com](mailto:Jerome.Vogedes@motorola.com)

**Attachments:** TR 22.977 v2.0.0 "Feasibility study for speech enabled services", TS 22.243 v.2.0.0 "Speech Recognition Framework for Automated Voice Services"

---

### **1. Overall Description:**

The TSG SA1 has been discussing the feasibility and associated requirements of various speech recognition technologies used to enable speech-only and multimodal services as part of its Technical Report and Specification for speech-enabled services [1, 2]. This feature is expected to be completed during the Release 6 time frame. Since the terminal cannot support purely terminal based speech recognition systems (due to limited CPU performance, power consumption, etc) for all applications, these services can utilise the use of remote speech recognition servers located in the network.

The speech can be transmitted from the terminal to the network using conventional speech codecs (such as AMR) or Distributed speech recognition codecs (such as ETSI AURORA DSR). TSG SA2 has the necessary expertise in the network architecture, which is required to determine the speech recognition framework and protocol stack for enabling SES (voice-only and multimodal) in the 3GPP system.

### **2. Actions:**

**To SA2 group.**

**ACTION:**

- To analyse the network architecture implications of:
  - the speech recognition framework used to enable speech services
  - multimodal and multi-device services in the 3GPP system
- and produce an associated stage 2 description for SES.
- Additionally, provide feedback on the stage 1 description on an as needed basis.

### **3. Date of Next TSG-SA1 Meetings:**

SA1 SWG	14-18 Oct 02	Beijing	China
SA1#18	11-15 Nov 02	TBD	Korea
SA1 SWG	20-24 Jan 03	TBD	TBD
SA1#19	24-28 Feb 03	Sophia Antipolis	France

#### **4. References:**

**[1]** TR 22.977 v 2.0.0 “Feasibility study for speech enabled services”

**[2]** TS 22.243 v.2.0.0 “Speech Recognition Framework for Automated Voice Services”

# 3GPP TR 22.977 V2.0.0 (08-2002)

---

*Technical Report*

**3rd Generation Partnership Project;  
Technical Specification Group Services and Systems  
Aspects;  
Feasibility study for speech enabled services;  
(Release 6)**

---



The present document has been developed within the 3<sup>rd</sup> Generation Partnership Project (3GPP™) and may be further elaborated for the purposes of 3GPP.

The present document has not been subject to any approval process by the 3GPP Organisational Partners and shall not be implemented.

This report is provided for future development work within 3GPP only. The Organisational Partners accept no liability for any use of this Specification.

Specifications and reports for implementation of the 3GPP™ system should be obtained via the 3GPP Organisational Partners' Publications Offices.

Keywords

---

UMTS, network, architecture

**3GPP**

Postal address

---

F-06921 Sophia Antipolis Cedex - FRANCE

Office address

---

650 Route des Lucioles - Sophia Antipolis  
Valbonne - FRANCE  
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

---

<http://www.3gpp.org>

---

**Copyright Notification**

---

No part may be reproduced except as authorized by written permission.  
The copyright and the foregoing restriction extend to reproduction in all media.

© 2001, 3GPP Organizational Partners (ARIB, CWTS, ETSI, T1, TTA, TTC).  
All rights reserved.

---

# Contents

Foreword.....	4
Introduction .....	4
Speech Enabled Services.....	4
Multimodal and Multi-device Services .....	4
Overview	5
1 Scope .....	5
2 References .....	5
2.1 Informative references .....	5
2.1 Normative references .....	6
3 Definitions and abbreviations.....	6
3.1 Definitions.....	6
3.2 Abbreviations.....	7
4 Speech-Enabled Services .....	8
4.1 Application Scenarios .....	8
5 Multimodal Services.....	8
5.1 Application Scenarios .....	9
6 Speech Recognition Technology.....	10
6.1 DSR standards.....	13
7 Multimodal and Multi-device Technology.....	14
7.1 Execution Model.....	14
7.2 Deployment configurations .....	15
7.3 Authoring .....	18
8 Requirements to introduce Speech-enabled services.....	19
8.1 Initiation.....	19
8.1.1 Service initiation.....	19
8.1.2 Multimodal or multi-device access configuration .....	19
8.2 Information during the interaction session.....	19
8.3 Control .....	20
8.4 User Perspective (User Interface) .....	20
8.5 Service Provisioning .....	20
8.6 Security .....	20
8.7 Privacy .....	21
8.8 Charging.....	21
9 Impact on the 3GPP system.....	22
9.1 Speech Recognition within 3GPP system .....	22
9.2 Multimodal and Multi-device Services within 3GPP system.....	23

---

## Foreword

This Technical Specification (TS) has been produced by the 3<sup>rd</sup> 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.

---

## Introduction

### 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.

---

# 1 Scope

This technical report provides an overview of various speech-enabled services and the different ways of enabling speech recognition for these services. This TR includes information applicable to network operators, service providers, as well as terminal and network manufacturers. It also discusses enabling multimodal and multi-device services aspects.

---

# 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.
- A non-specific reference to an ETS shall also be taken to refer to later versions published as an EN with the same number.

## 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 standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", ETSI ES 201 108

- [3] Draft ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 202 050
- [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](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] 3G TS 21.905: “Vocabulary for 3GPP Specifications”
- [18] 3GPP TS 22.243: “Speech Recognition Framework for Automated Voice Services”.
- [19] 3GPP TS 21.133: “3G Security; Security Threats and Requirements”.
- [20] 3GPP TS 22.228: “Service Requirements for IP multimedia Sub System”.

---

# 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.

## 3.2 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.

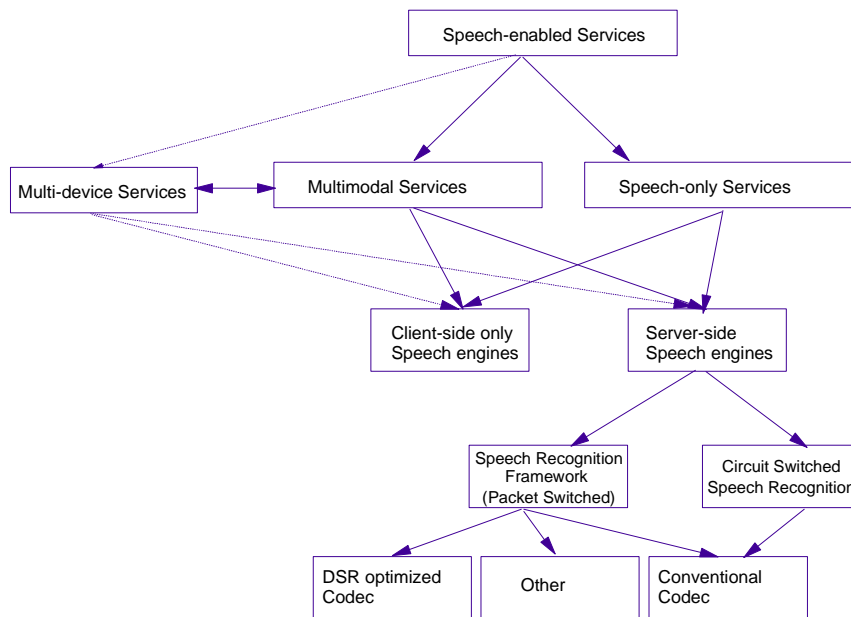


Figure 1 Chart illustrating the different kinds of speech-enabled services and the different methods available to implement speech recognition technology needed to enable these services

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].

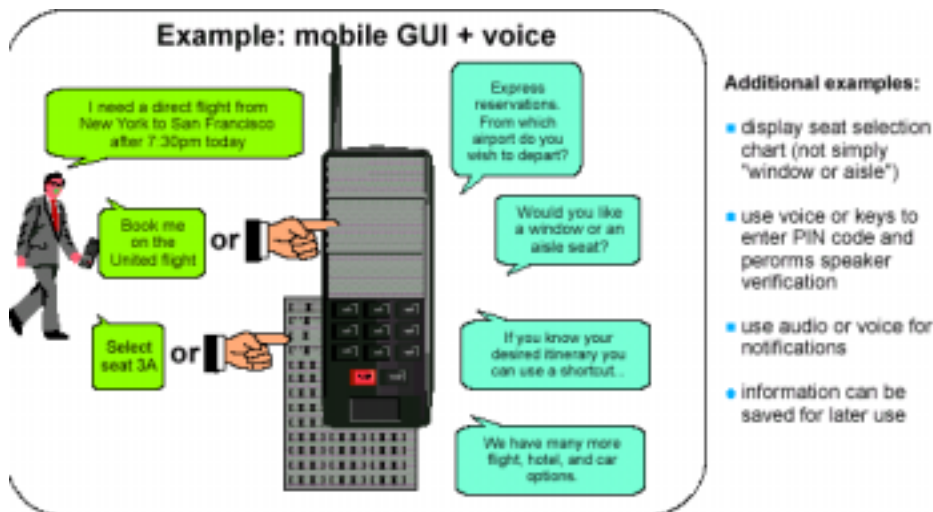


Figure 2 Example of a multi-modal scenario

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:
  - o Circuit switched-based.
  - o Packet switched-based speech recognition framework that supports exchange of encoded speech and meta-information [18]:
    - o with conventional codecs (e.g. AMR)
    - o 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:

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

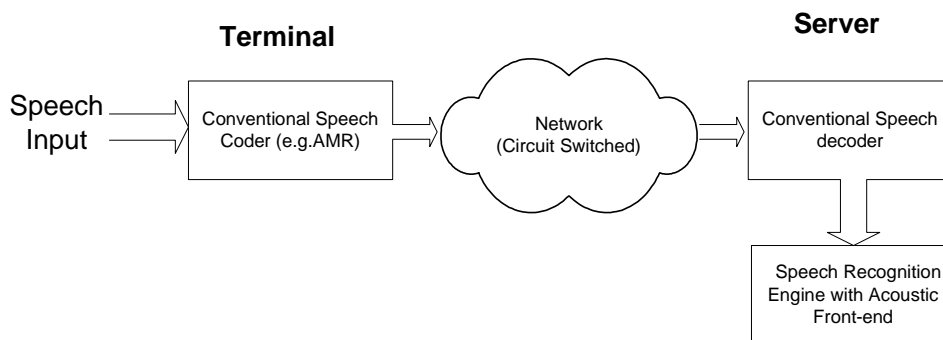


Figure 3 Illustration of Network Based Circuit Switched Speech Recognition

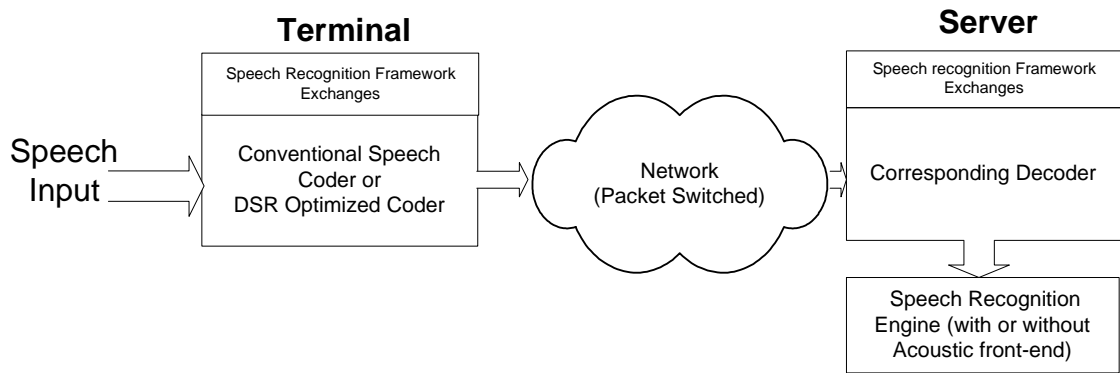


Figure 4 Illustration of Speech Recognition Framework

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.

Please note that additional modules may be needed to enable multimodal services.

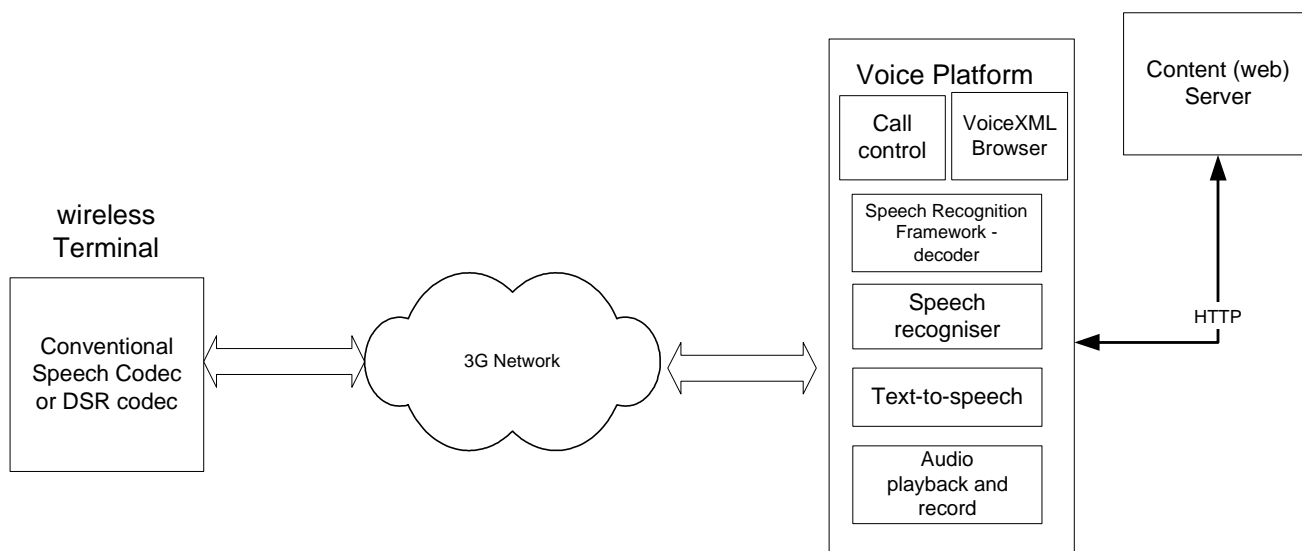


Figure 5 Illustration of the architecture for a typical speech service in a 3G network using network side speech recognition resources based on the speech recognition framework (SRF)

## 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.
- 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 mitigation to recognition under channel. Examples of the complexities of standards are table 1.

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

server side error maintain performance errors. computational the two estimated in

Table 1: DSR Terminal side Complexity

## 7. Multimodal and Multi-device Technology

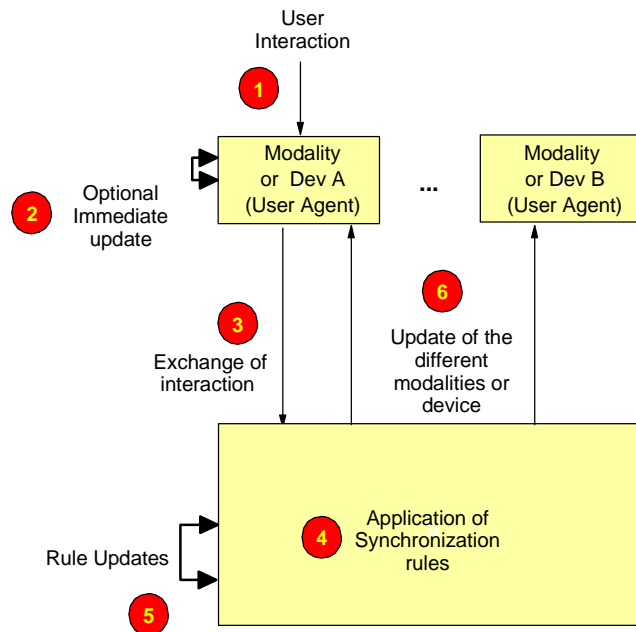


Figure 6 –Execution model of multimodal or multi-device applications; independent of programming model or configuration.

### 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).

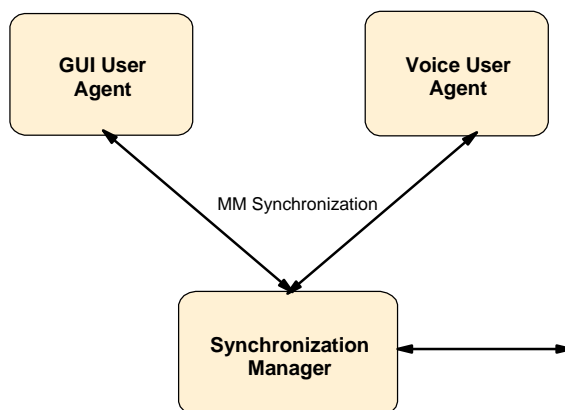


Figure 7 – Example of a basic architecture that supports multimodal and multi-device interactions illustrated for voice and GUI interaction.

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.

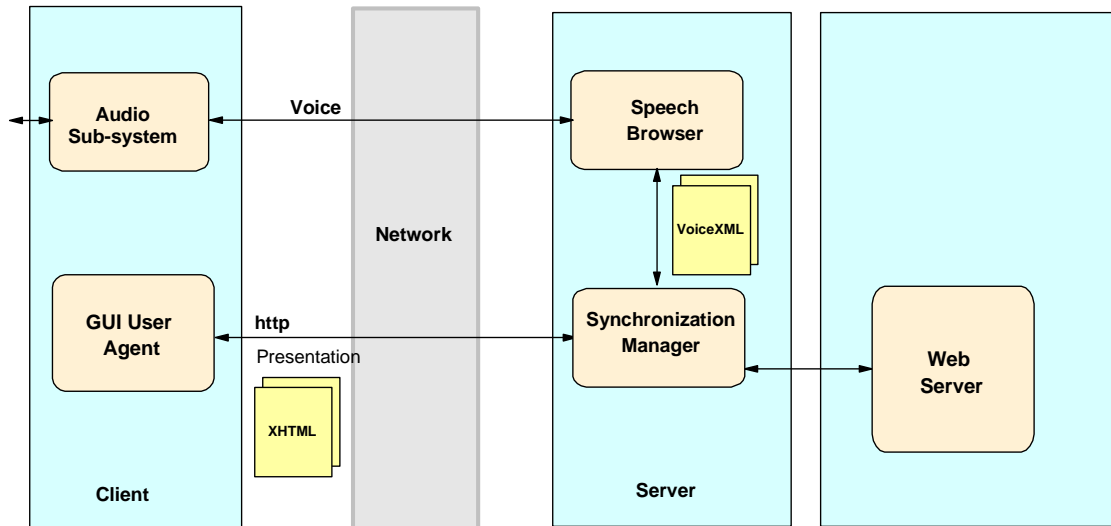


Figure 8 – Example of sequential configuration (no voice and data support simultaneously) for voice and GUI interaction. This configuration does not require IMS (SRF): it can be deployed on 2G or 2.5G networks. Only one modality is available at a given moment. The user may switch at any time or when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

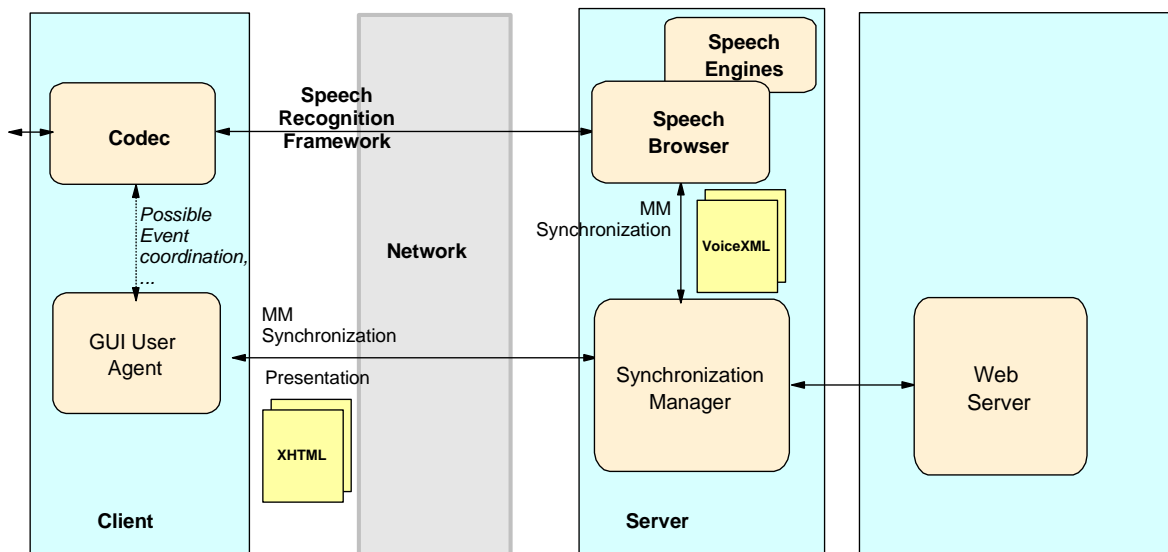


Figure 9 – Example of Thin Client Configuration (voice and data support) - Server-side speech engines local to speech browser for voice and GUI interaction.

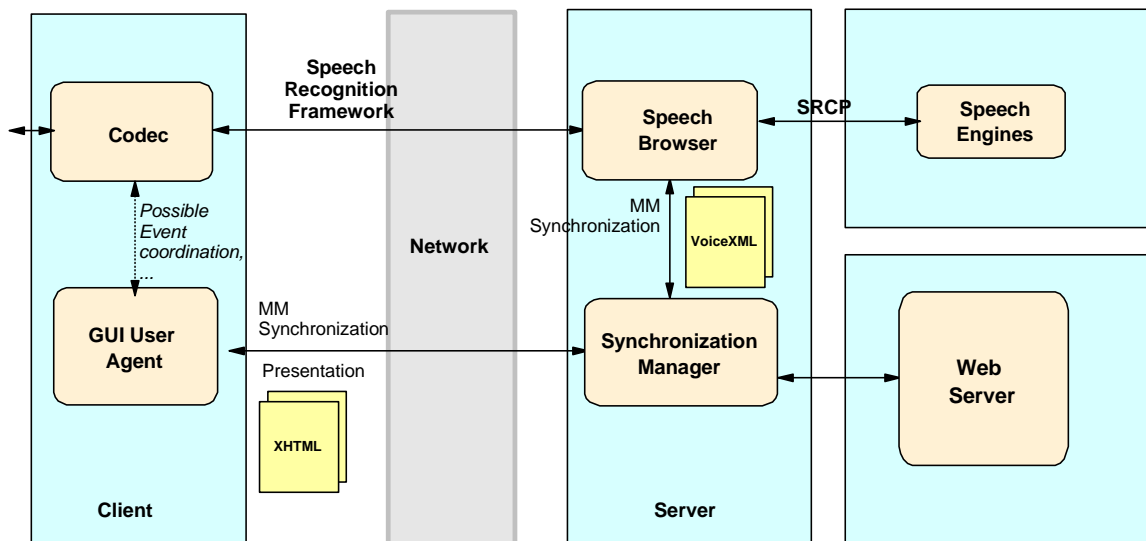


Figure 10 – Example of Thin Client Configuration (voice and data support) - Server-side speech engines remote with respect to speech browser for voice and GUI interaction.

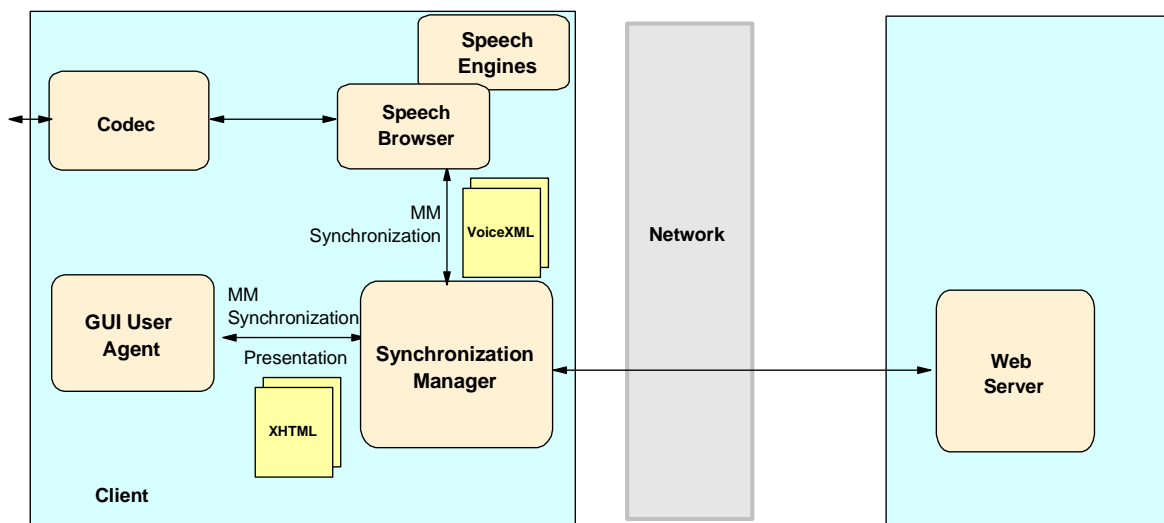


Figure 11 – Example of Fat client configuration with local speech engines for speech and GUI interaction. This can be the internal architecture of a browser implementation.

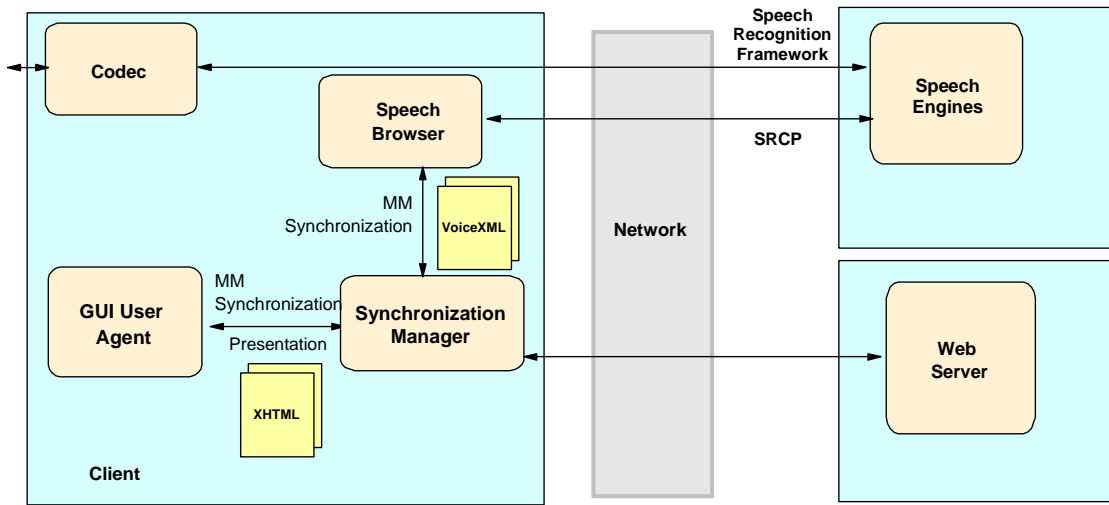


Figure 12 - Example of Fat client configuration with server-side speech engines for speech and GUI interaction. The speech engines are remote controlled by SRCP.

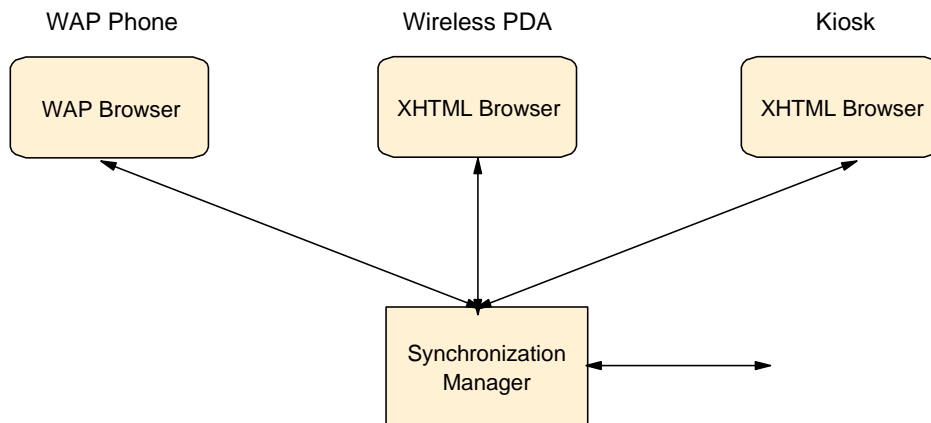


Figure 13 – Example of Multi-device configuration.

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 multidevice 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, proprietary 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) By duration of session (including "one-off" charge/flat rate)
- b) By data volume transferred (number of packets) or other similar criteria.
- c) By subscription fees for the service (unlimited usage or unlimited usage up to a point and then per-use fees)
- 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.



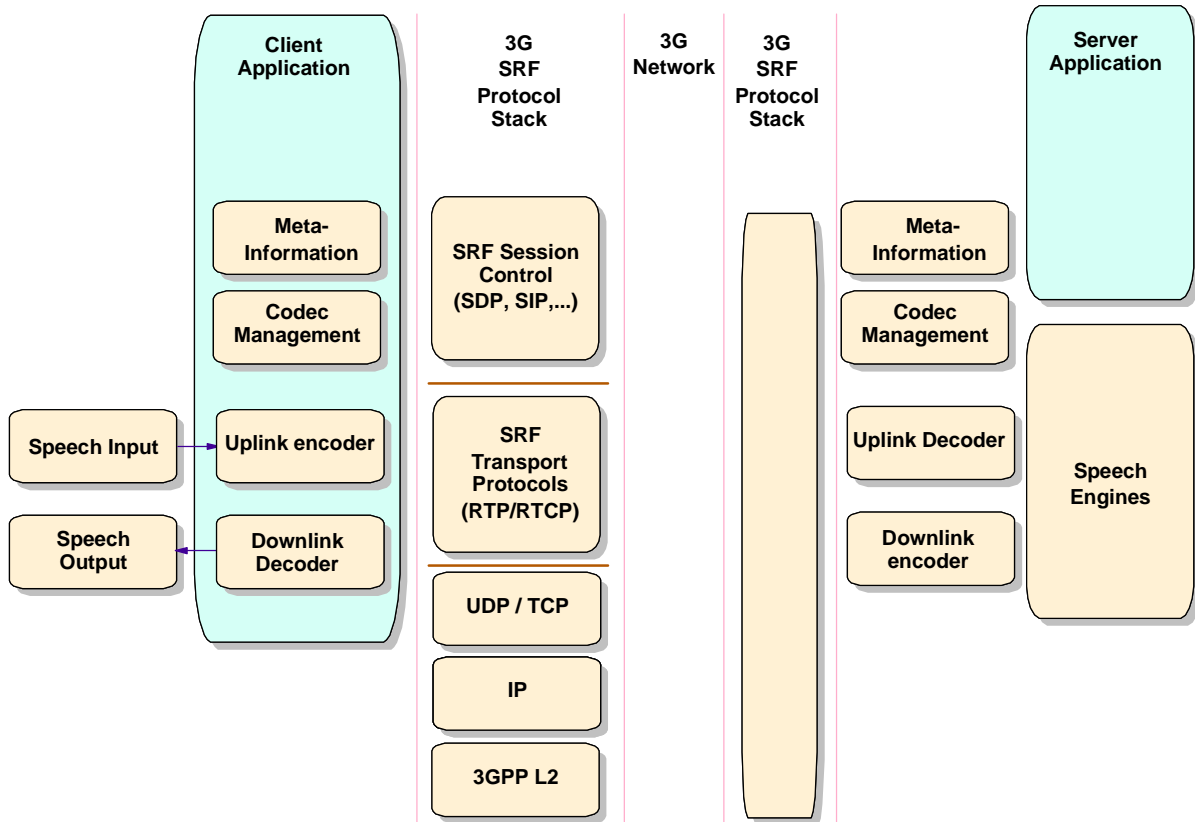


Figure 14 – Possible architecture for integrating SRF framework within 3GPP protocol stack and architecture.

## 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.

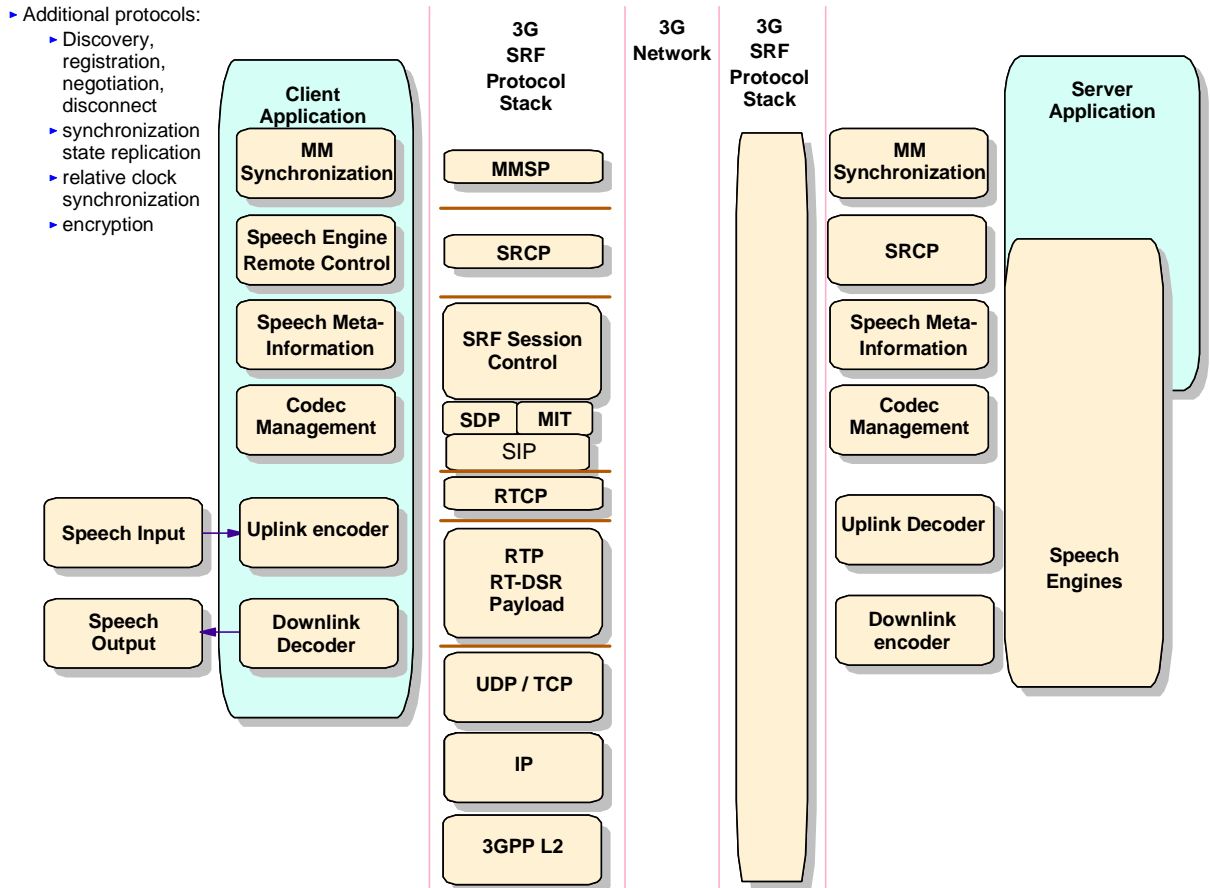


Figure 15 – Possible architecture for integrating multimodal and multi-device services within 3GPP.

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 (informative): Change history

Date	Version	Information about changes
SA1#14	0.0.0	First Draft
SA1#15	0.1.0	Updated version 0.1.0
SA1 SES SWG Sophia Antipolis	0.2.0	Updated version 0.2.0
SA1 SES Victoria	0.3.0	Updated version of 0.2.0
SA1, SES SWG Rome	1.1.0	Contribution to add multimodal and multi-device services in SES + some additional updates to fit the agreed upon details on SRF.
SA1#17, SES	1.1.0	Chair's update to reflect Rome SES SWG agreements: S1-021517.
SA1#17 SES	1.2.0	More detailed updates to complete and support the agreements reached at Rome SES SWG.

# Draft 3G TS 22.243 2.0.0 (2002-8)

---

*Technical Specification*

**3rd Generation Partnership Project;  
Technical Specification Group Services and System Aspects  
Service aspects; Stage 1  
Speech Recognition Framework for Automated Voice Services  
(3G TS 22.243 Release 6)**

---



The present document has been developed within the 3<sup>rd</sup> Generation Partnership Project (3GPP™) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPP Organisational Partners and shall not be implemented. This Specification is provided for future development work within 3GPP only. The Organisational Partners accept no liability for any use of this Specification. Specifications and reports for implementation of the 3GPP™ system should be obtained via the 3GPP Organisational Partners' Publications Offices.

---

Keywords

---

UMTS, architecture, speech recognition, automated voice services

**3GPP**

Postal address

---

3GPP support office address

---

650 Route des Lucioles - Sophia Antipolis  
Valbonne - FRANCE  
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

---

<http://www.3gpp.org>

**Copyright Notification**

---

No part may be reproduced except as authorized by written permission  
The copyright and the foregoing restriction extend to reproduction in all media.

© 2002, 3GPP Organizational Partners (ARIB, CWTS, ETSI, T1, TTA, TTC).  
All rights reserved.

---

# Contents

Contents .....	3
Foreword.....	4
Introduction .....	5
1. Scope .....	5
2. References .....	6
2.1 Normative References .....	7
2.2 Informative References.....	7
3. Definitions and abbreviations .....	7
3.1 Definitions .....	7
3.2 Abbreviations .....	8
4. Requirements.....	9
4.1 Initiation .....	9
4.2 Information during the speech recognition session.....	9
4.3 Control.....	10
4.4 User Perspective (User Interface).....	10
5 UE and network capabilities .....	10
6. Administration.....	10
6.1 Authorization .....	11
6.2 Deauthorization .....	11
6.3 Registration.....	11
6.4 Deregistration .....	11
6.5 Activation .....	11
6.6 Deactivation.....	11
7. Service Provisioning .....	11
8. Security.....	12
9. Privacy.....	12
10. Charging .....	12
11. Roaming .....	13
12. Interaction with other services .....	13
Annex A (informative): Speech recognition Framework-based automated voice service examples.....	14
Annex B (informative): Change History .....	15

---

# Foreword

This Technical Specification has been produced by the 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 this TR, 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 specification.

---

## Introduction

Forecasts show that speech-driven services will play an important role on the 3G market. People 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. At present, however, the complexity of medium and large vocabulary speech recognition systems is beyond the memory and computational resources of such devices. Also associated delay to download speech data files (e.g. grammars, acoustic models, language models, vocabularies etc. ...) may be prohibitive. Eventually, it may not always be acceptable for the speech service providers to allow download of these speech data files if they contained confidential information (password (security issue), customer names and address (privacy issue)) or intellectual properties; for example a well crafted speech grammar is often considered by speech service providers as a trade secret.

Server-side processing of the combined speech and DTMF input and speech output, can overcome these constraints by taking full advantage of memory and processing power as well as specialized speech engines and data files. However, the distortions introduced by the encoding used to send the audio between the client and the server as well as additional network errors can degrade the performance of the speech engines; therefore also limiting the achievable speech functionalities. A server-side speech service is generally equivalent to a phone call to an automatic service. As for any other telephony service, DTMF is a feature that should always be considered as needed.

This document describes a generic speech recognition framework to distribute the audio sub-system and the speech services by sending encoded speech and meta-information between the client and the server. Instead of using a voice channel as in today's server-based speech services, an error-protected data channel will be used to transport encoded speech from the client audio sub-system (terminal client) to remote speech engines (on server) for processing (e.g. speech recognition, speaker recognition,). The speech recognition framework will also enable downlink data streaming of voice and recorded audio prompt generated by server to the terminal client audio subsystem. The speech recognition framework may use conventional codecs like AMR or Distributed Speech Recognition (DSR) optimized codecs.

The speech recognition framework will provide users with a high performance distributed speech interface to server-based automatic speech services with communication, information access or transactional purposes.

The types of supported user interfaces include those that are voice only, for example, automatic speech access to information, such as a voice portal described in this section. These typically support combined speech or DTMF input.

In the future, a new range of multi-modal applications is also envisaged incorporating different modes of input (e.g. speech, keyboard, pen) and speech and visual output.

---

## 1. Scope

This Technical Specification defines the stage one description of the Speech Recognition Framework for Automated Voice Services. Stage One is the set of requirements for data seen primarily from the user's and service providers' points of view.

This Technical Specification includes information applicable to network operators, service providers, terminal and network manufacturers.

This Technical Specification contains the core requirements for the Speech Recognition Framework for automated voice services.

The scope of this Stage 1 is to identify the requirements for 3G networks to support the deployments of a speech recognition framework - based automated voice services and therefore to introduce a 3GPP speech recognition framework as part of speech-enabled services. The Speech Recognition Framework for automated voice services is an optional feature in a 3GPP system.



Figure 1 positions the Speech recognition Framework (SRF) with respect to other speech-enabled services as discussed in [6]. As illustrated, SRF is designed to support server-side speech recognition over packet switched network (e.g. IMS). As such SRF also enable configurations of multimodal and multi-device services that include distribute the speech engines.

Note that it is possible to design speech-enabled services that alternate or combine the use of client-side only engines and SRF.

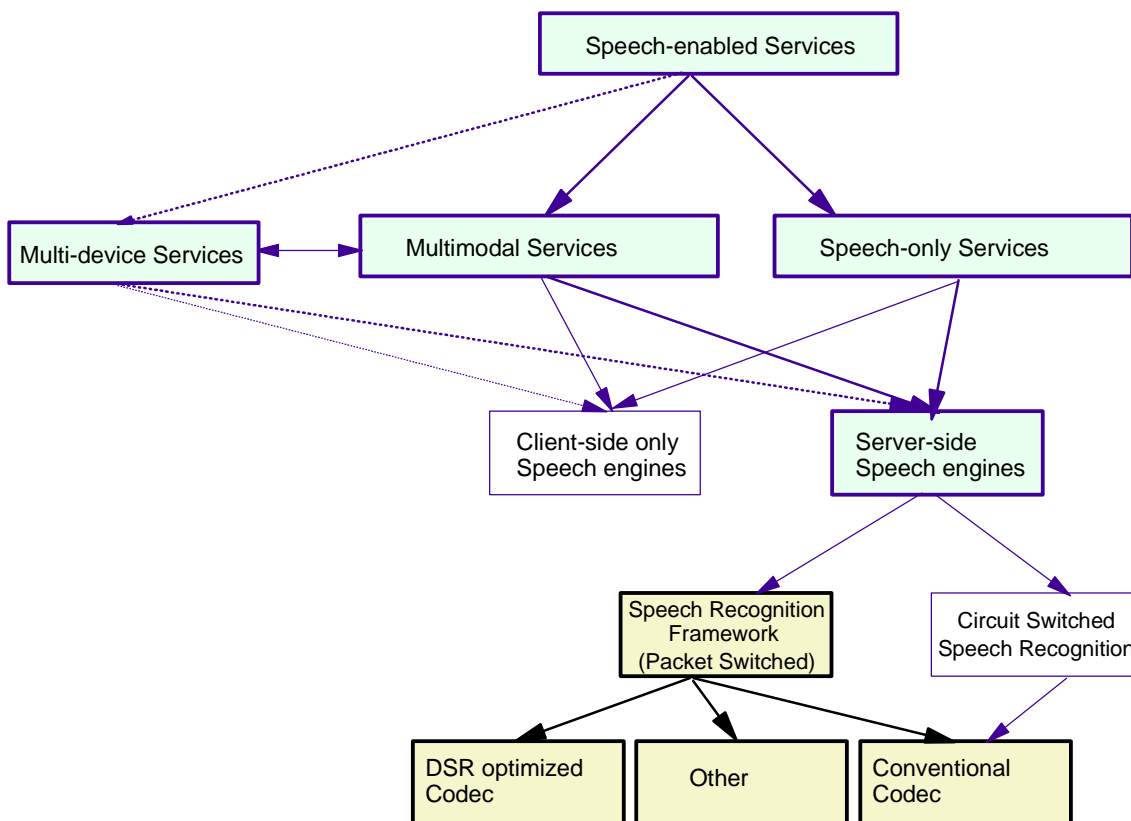


Figure 1 - Positions the scope of the speech recognition framework as part of general speech enabled services.

## 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.
- A non-specific reference to an ETS shall also be taken to refer to later versions published as an EN with the same number.

## 2.1 Normative References

- [1] 3GPP TS 21.133: "3G Security; Security Threats and Requirements"
- [2] 3GPP TS 21.905: "Vocabulary for 3GPP Specifications"
- [3] 3GPP TR 22.941: "IP Based Multimedia Services Framework"
- [4] 3GPP TS 22.105: "Services and Service Capabilities"
- [5] 3GPP TS 22.228: "Service Requirements for IP multimedia Sub System"
- [6] 3GPP TR 22.977: "Feasibility for speech enabled services"

## 2.2 Informative References

- [7] ETSI ES 201 108 v1.1.2 Distributed Speech Recognition: Front-end Feature Extraction Algorithm; Compression Algorithm", April 2000.
- [8] AU/410/02 "Speech recognition performance comparison between AMR speech coding and the Advanced DSR Front-End (ETSI ES 202 050)", Motorola, April 2002
- [9] AU/411/02 "Speech recognition performance comparison between AMR speech coding and the DSR Front-End (ETSI ES 201 108)", France Telecom, April 2002
- [10] ETSI ES 202 050 v0.0.0 "Distributed Speech Recognition: Advanced Front-end Feature Extraction Algorithm; Compression Algorithm", standard selected; document in preparation.

---

## 3. Definitions and abbreviations

Definitions and abbreviations used in the present document are listed in TR 21.905 [2]. For the purposes of this document the following definitions and abbreviations apply:

### 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.

**Barge-in event:** Event that takes place when the user starts to speak while audio output is generated.

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

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

**Dialog manager:** A technology to drive a dialog between user and automated voice services. For example a VoiceXML voice browser is essentially a dialog manager programmed by VoiceXML that drives speech recognition and text-to-speech engines.

**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 conventional codec, such as AMR. On the

server-side, the uplink encoded stream can be directly consumed by speech engines without having to be converted to a waveform.

**Meta information:** Data that may be required to facilitate and enhance the server-side processing of the input speech and facilitate the dialog management in an automated voice service. These may include keypad events over-riding spoken input, notification that the UE is in hands-free mode, client-side collected information (speech/no-speech, barge-in), etc....

**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.

**Speech Recognition Framework-based Automated Voice Service:** An automated voice service utilising the speech recognition framework to distribute the speech engines from the audio sub-system. In such a case the user voice input is captured and encoded, with a conventional or a DSR optimised for speech recognition as negotiated at session initiation. The encoded speech is streamed uplink to server-side speech engines that process it. The application dialog manager generates prompts that are streamed downlink to 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.

**Text-to-Speech Synthesis:** A technology to convert text in a given language into human speech in that particular language.

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

## 3.2 Abbreviations

For the purposes of this document the following abbreviations apply:

AMR – Adaptive Multi Rate

DSR – Distributed Speech Recognition

DTMF – Dual Tone Multi-Frequency

IETF – Internet Engineering Task Force

IMS – IP Multimedia Subsystem

IVR – Interactive Voice Response system

PCM – Pulse Coded Modulation

PIM - Personal Information Manager

SIP – Session Initiation Protocol

SRF – Speech Recognition Framework

URI – Uniform Resource Identifier

---

## 4. Requirements

A 3GPP speech recognition framework enables the use of conventional codecs (e.g. AMR) or DSR optimized codecs to distribute in the network the speech engines that process speech input or generate speech output. It includes:

- Default uplink and downlink codec specifications.
- A stack of speech recognition protocols to support:
  - Establishment of uplink and downlink sessions, along with codec negotiation
  - Transport of speech recognition payload (uplink) with conversational QoS
  - Support of transport (also at conversational QoS) of meta-information required for the deployment of speech recognition applications between the terminal and speech engines (meta-information may include terminal events and settings, audio sub-system events, parameters and settings, etc.).

IMS provides a protocol stack (e.g. SIP/SDP, RTP and QoS), that may advantageously be used to implement such capabilities.

It shall be possible to recommend a codec to be supported by default to deploy services that rely on the 3GPP speech recognition framework. To that effect, the specifications will consider either conventional speech codecs (e.g. AMR) or DSR optimized codecs.

ETSI has published DSR optimized codecs specifications (ETSI ES 201 108 & ETSI ES 202 050 [7, 8, 9, 10]) and a payload format for transport of DSR data over RTP (IETF AVT DSR).

The following list gives the high level requirements for the SRF-based automated voice services: .

- Users of the SRF-based automated voice service shall be able to initiate voice communication, access information or conduct transactions by voice commands using speech recognition. Examples of SRF-based automated voice services are provided in Appendix A.

The speech recognition framework for automated voice services will be offered by the network operators and will bring value to the network operator by the ability to charge for the SRF-based automated voice services.

This service may be offered over a packet switched network; however in general this requires specification of a complete protocol stack. When this service is offered over the IMS, the protocols used for the meta information and front-end parameters (from terminal to server) and associated control and application specific information can and shall be based on those in IMS.

### 4.1 Initiation

It shall be possible for a user to initiate a connection to the SRF-based automatic voice services by entering the identity of the service. Most commonly, when used as a voice service, this will be performed by entering a phone number. However, particular terminals may offer a user agent that accepts other addressing schemes to be entered by the user: IP address, URI, e-mail address possibly associated to a protocol identifier. This is particularly important for multi-modal usages.

In all cases, the terminal will convert the address entered by the user to initiate a session via the SIP IMS session initiation protocol and establish the different SRF protocols. During this initiation of the SRF session, it shall be possible to negotiate the uplink and downlink codecs. The terminal shall support a speech recognition optimized codec as a default uplink codec.

### 4.2 Information during the speech recognition session

Codec negotiation during a SRF session should be optionally supported.

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 service need. The user speaks to the service and receives output back from the automated voice 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 control and application specific information shall be exchanged during the session between the client and the service. Accordingly some terminals shall be able to support sending additional data to the service (e.g. keypad information and other terminal and audio events) and receiving data feedback that shall be displayed on the terminal screen.

Dynamic payload switches within a session may be considered to transport meta-information.

## 4.3 Control

It shall be possible to use SRF sessions in order to provide access to SRF-based automated voice services. For example applications might use a SRF session to access and navigate within and between the various SRF-based automated voice services by spoken commands or pressed keypads.

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

## 4.4 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. These are examples of multimodal interfaces. SRF enables distributed multimodal interfaces as described in [6].

---

# 5 UE and network capabilities

In addition to the capabilities required for IMS Basic Voice session (such as the default voice codec that will be used for the downlink audio prompt stream), the following SRF-based automated voice service-specific capabilities shall be required in the UE and network:

- A default uplink codec (conventional codec or DSR optimized codec).
- A downlink conventional codec and downlink streaming capabilities (simultaneous with uplink)
- The capability to transmit keypad information from the client to the server (e.g., either DTMF or the keypad string)

It shall be possible to enable application specific information exchanges between the client and the server (e.g. client events (e.g. barge-in events), display information, etc...), in the form of speech meta-information. It shall be possible to enable these exchanges with conversational QoS.

SRF shall be supported by an uplink bandwidth of 9.6 kbits/s for the payload and QoS (Quality of Service) for conversational class services as specified in TS 22.105 [4]

It shall be possible for the network to distinguish a SRF session from a basic voice session (e.g. for charging purposes).

---

# 6. Administration

SRF-based automated voice services may be provided by the network operator (home or visited) or by third parties. See appendix A for examples of such services.

The administration of the SRF-based automated voice services will 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 party providers to administer the SRF-

based automated voice services themselves through the gateway that they would connect to IMS. In such case, the third party provider performs all the administrative steps and no registration would be required with the network operator.

## 6.1 Authorization

Authorization for use of SRF-based automated voice services will be under the control of the network operator. It shall require authorization of the connection to IMS.

The network operators can provide automated voice services or they can only provide the network that connects users to SRF-based automated voice services provided by third party application service providers. The network operator shall be able to permit or prevent access to a third party service. This requirement shall be treated as equivalent to allowing or prevent access to some phone numbers (IMS voice sessions) or internet services (domains for example in WAP data access).

The network operator shall be able to administer the authorization of a SRF-based automated voice service on a user basis as well as on a service basis (e.g. to authorize access to all users or prevent access to all users).

It shall be possible for the operator to provide for the user:

- Authorization to access a particular "address" (e.g. 3rd party SRF-based automated voice service)
- Authorization to use a service that the operator authorize access to when the 3rd party operator wishes to rely on the operator to control this access

It shall be possible for the third party provider to authorize usage of its services based on the identity of the user.

## 6.2 Deauthorization

Deauthorization for use of the SRF-based automated voice services shall be under the control of the network operator as for the authorization described in section 6.1.

## 6.3 Registration

Authorized SRF-based automated voice service register their address with the IMS upon authorization of the service authorized SRF-based automated voice service can then be reached by the user (by entering address and initiated a SRF session).

## 6.4 Deregistration

Disconnection from the IMS shall prevent the use of the SRF-based automated voice service. Deregistration may be decided by the third party provider.

## 6.5 Activation

Once authorized and registered a SRF-based automated voice service is deemed activated as for other IMS services.

## 6.6 Deactivation

Deactivation shall be done by deregistering the services (operator or service provider initiative) or by refusing to initiate a SRF session (service provider initiative).

---

# 7. Service Provisioning

The SRF-based automated voice services 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 SRF-based automatic voice services by providing identity of service, such as a phone number, an IP address or a URI that the user can enter or select on the terminal.

---

## 8. Security

The “Security Threats and Requirements” specified in 21.133 [1] shall not be compromised.

It shall be possible to deny unauthorized access to 3GPP SRF-based automated voice 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 3GPP SRF-based automated voice services.

It shall be possible to reconstitute PCM samples from DSR packets so that the user’s spoken command can be transcribed at a later time, if required.

---

## 9. Privacy

For SRF-based automated voice services, privacy requirements shall be at least as good as for IMS voice or data sessions [5]:

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

SRF-based automated voice 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 SRF-based automated voice services shall not add additional privacy risks.

---

## 10. Charging

The user can be charged for sessions with SRF-based automated voice services in a variety of ways. The following shall be possible:

- a) By duration of session (including “one-off” charge/flat rate)
- b) By data volume transferred (number of packets) or other similar criteria.
- c) By subscription fees for the service (unlimited usage or unlimited usage up to a point and then per-use fees)
- d) Free (e.g. with the service being subsidised by advertising revenue from advertisement spots). The advertisement spots may be inserted either at session start-up or close, or designed in such that system delay time is masked (e.g., while the user is waiting for the flight schedules to be returned, or a purchase transaction to be completed). The network operator will receive revenue from users directly as well as from the content and service providers who want their sites to be accessible via the automated voice service, and from advertisers. Advertising spots can be inserted at appropriate points during the session (e.g., at the beginning of the session, while the user is waiting for a system response, or at the end of a session).

SRF-based automated voice services shall be available to pre-paid and post-paid subscribers.

---

## 11. Roaming

The user shall be able to utilize SRF-based automated voice services when roaming in any IMS compatible mobile network.

The capabilities of the SRF-based automated voice service shall be available in the roamed-to network in the same manner as in the home network, within the limitation of the capabilities of the serving network.

---

## 12. Interaction with other services

No interaction with other services identified. When connected to the IMS, other IMS services are available to the user through the terminal.

Other services (non IMS voice etc..) may be available with or without disconnecting from the IMS.



---

## Annex A (informative): Speech recognition Framework-based automated voice service examples

Examples of Automated Voice Services include:

- Communication assistance (Name dialling, Service Portal, Directory assistance)
- Information retrieval (e.g., obtaining stock-quotes, checking local weather reports, flight schedules, movie/concert show times and locations)
- M-Commerce and other transactions (e.g., buying movie/concert tickets, stock trades, banking transactions)
- Personal Information Manager (PIM) functions (e.g., making/checking appointments, managing contacts list, address book, etc.)
- Messaging (IM, unified messaging, etc...)
- Information capture (e.g. dictation of short memos)
- A usage scenario for multimodal applications with a GUI user agent on the terminal synchronized with an SRF automated voice service.

## Annex B (informative): Change History

Change history											
TSG SA#	SA Doc.	SA1 Doc	Spec	CR	Rev	Rel	Cat	Subject/Comment	Old	New	Work Item
			22.243					Initial draft based on content of TR 22.941		0.0.1	
								Output of ad-hoc drafting session		0.0.2	
								Presented to TSG SA#14 for information	0.0.2	1.0.0	
								Reviewed in SWG DSR 14 Jan 2002	1.0.0	1.1.0	
								Reviewed again in SWG DSR 14 Jan 2002	1.1.0	1.2.0	
								Contribution to 1.2.0 to address open issues	1.2.0	1.3.0	
								Update to 1.3.0 to address comments at DSR SWG (Sophia Antipolis)	1.3.0	1.4.0	
								Update to 1.4.0 to reflect the new SRF terminology.	1.4.0	1.5.0	
								Update to 1.5.0 to reflect changes agreed to in SWG mtg in Rome.	1.5.0	1.6.0	
								Update to 1.6.0 to reflect and update according to the agreements in SES SWG in Rome	1.6.0	1.6.1	
								Updated to reflect changes agreed to in SES SWG in Durango	1.6.1	1.7.0	

SA1 Rapporteur: Jerome Vogedes, Motorola, Inc.

Email: Jerome.Vogedes@motorola.com

Telephone: +18475231796