ETSI TS 122 243 V15.0.0 (2019-07)

**Digital cellular telecommunications system (Phase 2+) (GSM);
Universal Mobile Telecommunications System (UMTS);
LTE;
Speech recognition framework for automated voice services;
Stage 1
(3GPP TS 22.243 version 15.0.0 Release 15)**

Reference
RTS/TSGS-0122243vf00

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

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

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 Specification (TS) 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 "**shall**", "**shall not**", "**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 Specification 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.

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

The present document 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.

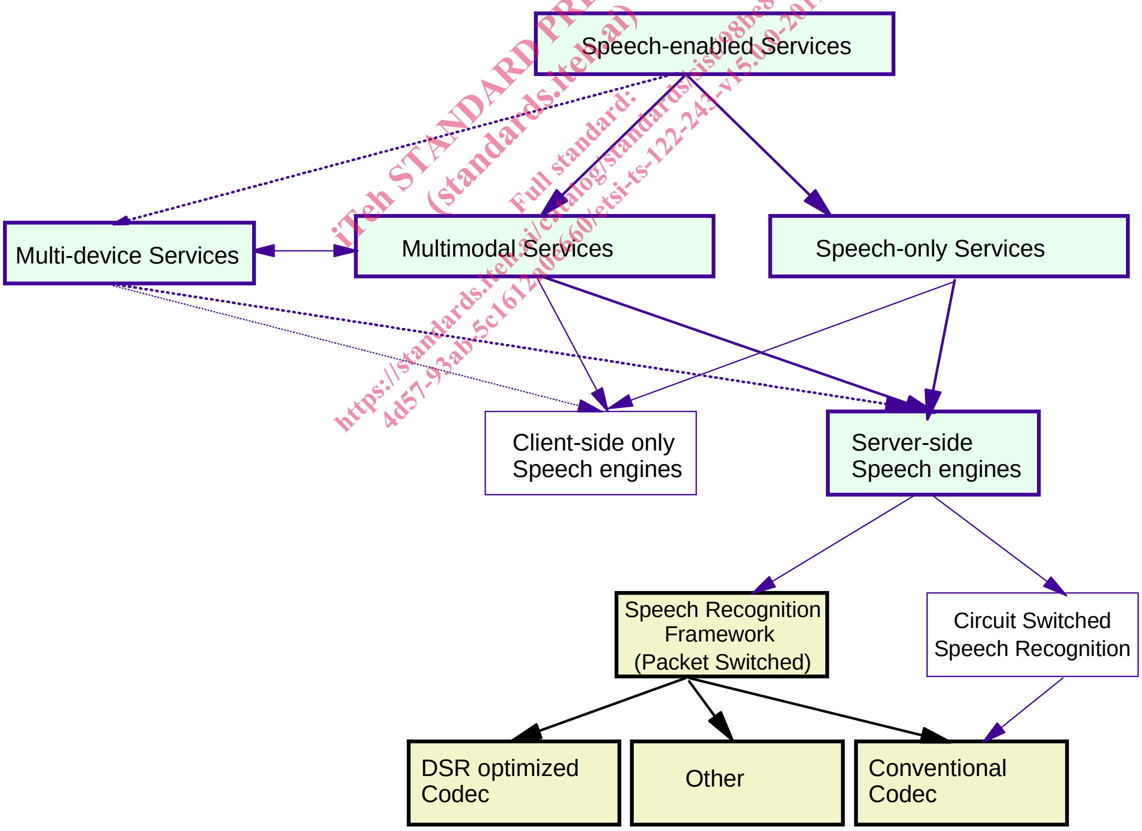

Figure 1 - Positions the scope of the speech recognition framework as part of general speech enabled services.

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. 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 Normative References

[1] 3GPP TS 21.133: "3G security; Security threats and requirements".

[2] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[3] 3GPP TR 22.941: "IP based multimedia framework; Stage 0".

[4] 3GPP TS 22.105: "Services and service capabilities".

[5] 3GPP TS 22.228: "Service requirements for the Internet Protocol (IP) multimedia core network subsystem; Stage 1".

[6] 3GPP TR 22.977: "Feasibility study 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] Void

[9] Void

[10] ETSI ES 202 050 v0.0.0 "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms; DSR advanced front end", 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, 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.