3GPP TS 22.243 V11.0.0 (2012-09)

Technical Specification

3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Speech recognition framework for automated voice services; Stage 1 (Release 11)

Keywords
LTE, UMTS, speech, stage 1

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

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.

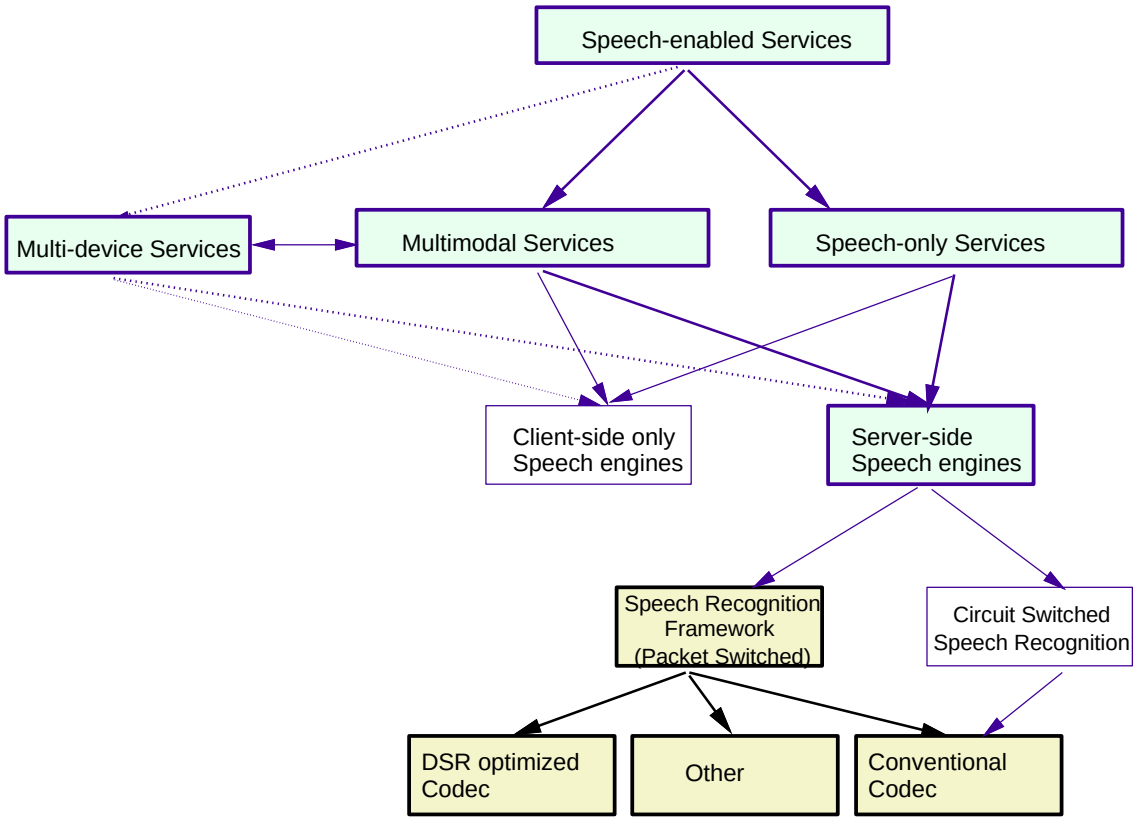

Figure 1 - Positions the scope of the speech recognition framework as part of general speech enabled services.

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.

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 codec suitable for speech recognition 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).
- The capability to reconstruct encoded speech. The reconstruction requirement does not apply to the UE.

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 channel available in GERAN and UTRAN networks for the transport of the codec payload and with QoS (Quality of Service) for conversational class, streaming and interactive QoS 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 phones 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	
			22.243					Output of ad-hoc drafting session		0.0.2	
			22.243					Presented to TSG SA#14 for information	0.0.2	1.0.0	
			22.243					Reviewed in SWG DSR 14 Jan 2002	1.0.0	1.1.0	
			22.243					Reviewed again in SWG DSR 14 Jan 2002	1.1.0	1.2.0	
			22.243					Contribution to 1.2.0 to address open issues	1.2.0	1.3.0	
			22.243					Update to 1.3.0 to address comments at DSR SWG (Sophia Antipolis)	1.3.0	1.4.0	
			22.243					Update to 1.4.0 to reflect the new SRF terminology.	1.4.0	1.5.0	
			22.243					Update to 1.5.0 to reflect changes agreed to in SWG mtg in Rome.	1.5.0	1.6.0	
			22.243					Update to 1.6.0 to reflect and update according to the agreements in SES SWG in Rome	1.6.0	1.6.1	
			22.243					Updated to reflect changes agreed to in SES SWG in Durango	1.6.1	1.7.0	
			22.243					Updated for presentation to SA #17	1.7.0	2.0.0	
SP-17	SP-020572		22.243			Rel-6		Presented to SA for approval	2.0.0	6.0.0	
SP-18	SP-020664	S1-021928	22.243	001		Rel-6	F	CR to TS 22.243 Removal of references	6.0.0	6.1.0	SRSES
SP-19	SP-030031	S1-030151	22.243	003	-	Rel-6	F	Correction of contradictory information (former: 'Removal of references')	6.1.0	6.2.0	SRSES
SP-20	SP-030260	S1-030431	22.243	004	-	Rel-6	F	UE and network capabilities	6.2.0	6.3.0	SRSES
SP-20	SP-030260	S1-030432	22.243	005	-	Rel-6	C	Addition of Streaming and interactive QoS	6.2.0	6.3.0	SRSES
SP-21	SP-030468	S1-030974	22.243	007	-	Rel-6	B	Reconstructed speech as an output mechanism	6.3.0	6.4.0	SRSES
SP-36			22.243			Rel-7		Updated from Rel-6 to Rel-7	6.4.0	7.0.0	
SP-42	-	-				Rel-8		Updated from Rel-7 to Rel-8	7.0.0	8.0.0	
SP-46	-	-	-	-	-	-	-	Updated to Rel-9 by MCC	8.0.0	9.0.0	
2011-03	-	-	-	-	-	-	-	Update to Rel-10 version (MCC)	9.0.0	10.0.0	
2012-09	-	-	-	-	-	-	-	Updated to Rel-11 by MCC	10.0.0	11.0.0	