Technical Specification Group Services and System Aspects
Meeting #20, Hämeenlinna, Finland, 09-12 June 2003

TSGS#20(03)0261

Source:	**SA1**
Title:	**CR to 22.243 on reconstructed speech (Rel-6)**
Document for:	**Approval**
Agenda Item:	**7.1.3**

CR-Form-v7

CHANGE REQUEST

⌘ **22.243** **CR** **006** ⌘**rev** **-** ⌘ Current version: **6.2.0** ⌘

*For **HELP** on using this form, see bottom of this page or look at the pop-up text over the ⌘ symbols.*

Proposed change affects: UICC apps⌘ ☐ ME ☐ Radio Access Network ☐ Core Network ☒

Title:	⌘	Reconstructed speech as an output mechanism
Source:	⌘	Nokia, Ericsson
Work item code:	⌘	SRSES
Date:	⌘	03/04/2003
Category:	⌘	**F**
Release:	⌘	Rel-6

Use one of the following categories:
F *(correction)*
A *(corresponds to a correction in an earlier release)*
B *(addition of feature),*
C *(functional modification of feature)*
D *(editorial modification)*
Detailed explanations of the above categories can be found in 3GPP TR 21.900.

Use one of the following releases:

2	*(GSM Phase 2)*
R96	*(Release 1996)*
R97	*(Release 1997)*
R98	*(Release 1998)*
R99	*(Release 1999)*
Rel-4	*(Release 4)*
Rel-5	*(Release 5)*
Rel-6	*(Release 6)*

Reason for change: ⌘ TS 22.243 has already identified that speech/audio output should be provided back to the user for speech enabled services. This change clarifies the details how audio output can be provided by adding reconstructed speech as an output mechanism.

The use-cases for reconstructed speech include database validation and training, confirmation and verification of the spoken input utterance, application dialogue design, system debug data logging in transaction processing, speaker identity verification, and solving problem situations by human operator.

Use case 1: *Verifying the user identity*. In applications, where financial transactions are involved, it is necessary to record the user dialogue for the future verification. It is possible that the user may deny that he has not done the transactions that he is charged for. While there are several ways to verify the caller identity (phone number, pin code etc.), the final issue is if it was the user himself who carried out the transaction. To make this check, it is necessary that there is recording of the transaction from where the user identity can be checked. Without this recorded proof, the user can claim that the phone call and the related transaction was not carried out by him.

Use case 2: *Human operator assistance*. With the help of automatic speech recognition, it is possible to automate call center functions. However, since speech recognition is prone to errors, a complete automation cannot be done, human operators are still required to solve the problem situations. To help the user as efficiently as possible, the operator is often required to listen to the earlier dialogue between the user and system before the intervention. As the operators need to continuously listen to these dialogues, it is important that

	dialogue recordings are available to minimize the load related to the operator work. If the operators need to make an effort for understanding the dialogue, these additional efforts reduce the working efficiency. Use case 3: *Accessibility guidelines*. In http://www.access-board.gov/telecomm/html/telfinal.htm accessibility guidelines have been specified for the use of telecommunication devices (e.g. for the visually impaired users of mobile phones). Voice output (see the use case #4) is mentioned in this document as one of the recommended output modality to be used in these devices. Use case 4: *Retraining of ASR systems with real-life utterances*: Retraining of ASR systems with real-life utterances from a normal application is a well-established and well-known technique to increase the ASR performance. Speaker independent systems are trained from a database or a set of databases containing utterances from a variety of speakers. By choosing utterances with certain characteristics (with / without background noise, type of noise, male / female, type of microphone, etc.), the conditions in the target application concerning acoustic input etc. shall be reflected as much as possible. Nevertheless the optimal performance is not achieved. To achieve the best possible performance, the user utterances are recorded, which were spoken while using the application ('real life'). These utterances are used to retrain the ASR vocabulary. This process (recording of utterances and retraining) is iterated until saturation in the recognition performance is achieved. Recorded utterances during usage of a service are a valuable asset for service providers. In order to allow training and retraining of ASR systems from different vendors and for different versions of an ASR system, the user utterances have to be stored. Use case 5**:** *Name dialling with acoustic feedback***:** Name dialling is one of the most successful applications in mobile communications. This application may be provided by implementations on handsets as well as on servers in the telecommunications network. Nowadays implementations are still dominated by speaker dependent technology. This means the user trains a name and as a result the trained reference for the recogniser and a coded representation for acoustic feedback is stored. There is a clear trend towards text enrolment, which means the user just has to type in the name as text into the ASR system and the reference is automatically generated as a speaker independent reference. Instead of typing, the names can be provided from a centralized database such as e.g. outlook. In this application there is a need to generate an acoustic feedback for the name. One possibility is to use a TTS (text-to-speech) system, which adds further costs and complexity to the name dialling system. An easy possibility to avoid usage of a TTS system, is to use the user's own voice for the acoustic feedback. This means the first time a name is dialled by a user, the spoken utterance is stored as acoustic confirmation for later usage.
Summary of change: ⌘	Addition of reconstructed speech as an output mechanism for speech enabled services.
Consequences if not approved: ⌘	The content of the speech data from the user of the service cannot be validated by human audition by the service provider or by the user of the service. The spoken speech cannot be verified. This results in problems, e.g., in case of dispute, the spoken speech cannot be kept to verify what was spoken or to verify the identity of the user. The use and creation of personal recognition vocabularies in speech enabled services becomes difficult as the users do not receive any output for the recognition results for these items. If the users cannot create personal vocabularies, the scope of speech enabled applications and services is

	considerably limited. Visually impaired users have also have substantial accessibility problems if speech output cannot be offered them in services. Without the possibility to store the real-world data in real services, it becomes very costly to maintain and improve the running speech enabled services. A separate data collection step is always needed whenever there are changes, e.g. in feature extraction part of the recogniser. Moreover, a long pilot period for real-world data collection is also needed to set up new services in order to optimise the system performance.

Clauses affected:	⌘	Chapters 4.2, 4.4 and 5.			
		Y	**N**		
Other specs affected:	⌘		**X**	Other core specifications	⌘
			X	Test specifications	
			X	O&M Specifications	
Other comments:	⌘				

How to create CRs using this form:

Comprehensive information and tips about how to create CRs can be found at http://www.3gpp.org/specs/CR.htm. Below is a brief summary:

1) Fill out the above form. The symbols above marked ⌘ contain pop-up help information about the field that they are closest to.

2) Obtain the latest version for the release of the specification to which the change is proposed. Use the MS Word "revision marks" feature (also known as "track changes") when making the changes. All 3GPP specifications can be downloaded from the 3GPP server under ftp://ftp.3gpp.org/specs/ For the latest version, look for the directory name with the latest date e.g. 2001-03 contains the specifications resulting from the March 2001 TSG meetings.

3) With "track changes" disabled, paste the entire CR form (use CTRL-A to select it) into the specification just in front of the clause containing the first piece of changed text. Delete those parts of the specification which are not relevant to the change request.

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 speech reconstructed from coder output) 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 or reconstructed 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 to be able to output back to the user.

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).