Technical Specification Group Services and System Aspects
Meeting #16, Marco Island, USA, 10-13 June 2002

TSGS#15(02)0259

Source:	**SA1**
Title:	**TR 22.977, Version 1.0.0 for Speech Enabled Services for information**
Document for:	**Information**
Agenda Item:	**7.1.3**

TSG-SA Meeting #16 **TSG SA #16 (01) 0259**
Marco-Island, 10-13th June 2002

TSG-SA WG 1 (Services) meeting #16 **S1-021184**
Victoria, Canada, 13-17th May 2002 **Agenda Item: Plenary**

Presentation of Specification to TSG or WG

Presentation to:	**TSG SA Meeting #16**
Document for presentation:	**TR 22.977, Version 1.0.0**
Presented for:	**Information**

Abstract of document:

This document is the current draft Technical Report on the feasibility of Speech Enabled Services.

Changes since last presentation to TSG-SA Meeting:

This is the first presentation of the 22.977 to SA.

Outstanding Issues:

None

Contentious Issues:

None

3GPP TR 22.977 V1.0.0 (2002)

Technical Report

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

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

Contents

Foreword

This Technical Specification (TS) 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

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. 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.

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. 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. Usage as multimodal services is also discussed. The scope of the report, references, definitions and abbreviations are detailed in the first few chapters.

[Editor's note: expand to include multimodal…]

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 discussed some multimodal 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] 3G TS 21.905: “Vocabulary for 3GPP Specifications”

[3] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108

[4] 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

3 Definitions and abbreviations

3.1 Definitions

[Editor’s note: to be completed]

3.2 Abbreviations

For the purposes of this document the following abbreviations apply:

DSR	Distributed Speech Recognition
GUI	Graphical User Interface

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 services. This is illustrated in Fig. 1. As the name implies, speech 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.

Speech, however, has its own limitations. It is serial and the user may find it difficult to remember long ouputs. 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 unimodal interaction, to result in more efficient user interfaces as compared to the unimodal 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)

- Chosing the interaction mode that suits the task and the circumstances

    - Input: key, touch, stylus, voice...

    - Output: display, tactile, audio...

- Enabling use of several devices in combination by exploiting the resources of multiple devices.

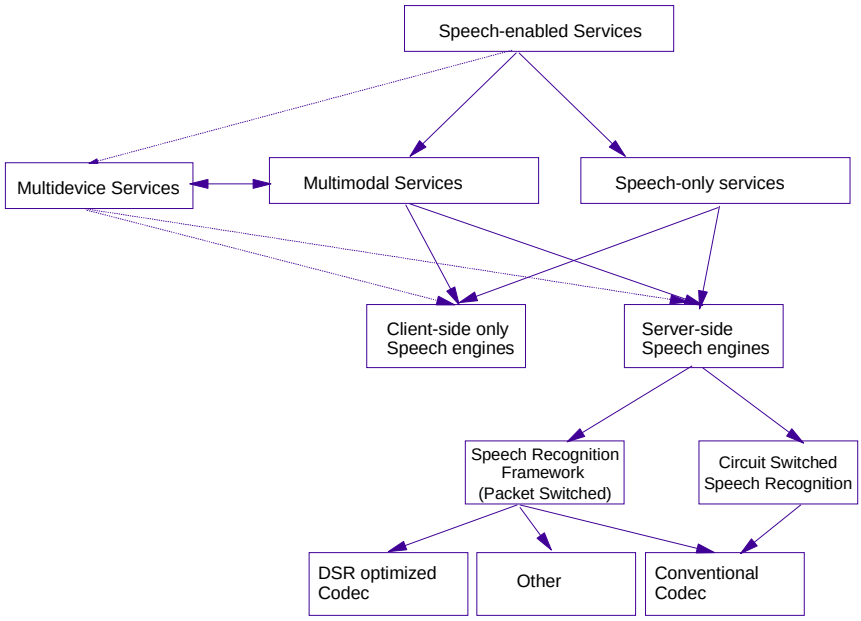

Figure 1 Chart illustrating the different kinds of speech-enabled services and the different methods available to implement speech recognition technology needed to enable these services

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. It then guides 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.

An example of a multimodal airline reservation service, involving both speech and GUI, is shown below.

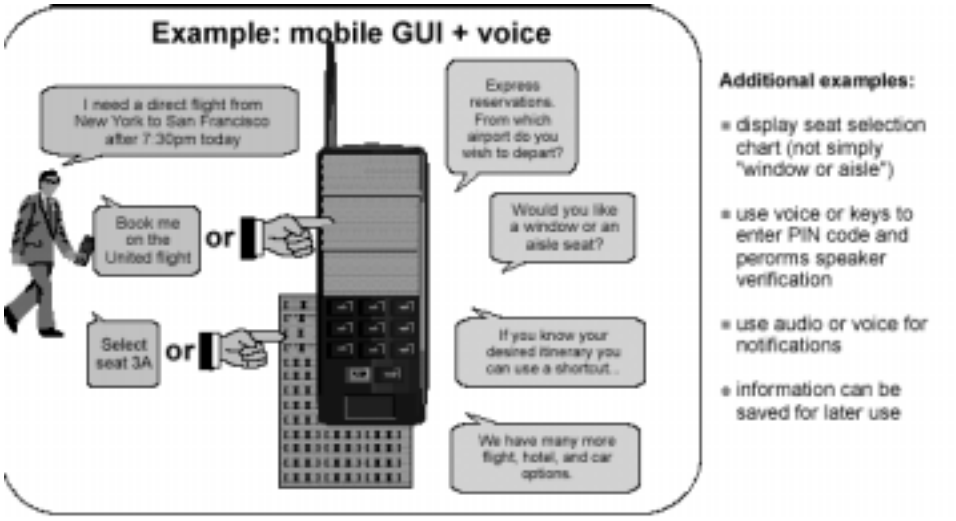

Figure 2 Example of a multi-modal scenario

5 Multimodal Services

[Editor's note: text to be provided]

6 Speech Recognition Technology

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

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

1. Client-side only speech engines (i.e. terminal based),
2. Server-side speech engines (i.e. network based). This can itself be subdivided into:
   - Circuit switched-based.
   - Packet switched-based speech recognition framework that supports exchange of encoded speech and meta-information:
     - with conventional codecs (e.g. AMR)
     - with DSR optimized codecs (Distributed Speech Recognition)

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

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

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

recognition system that would need lots of memory and computational resources - items that are scarce in todays 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 tranmission 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).

In distributed speech recognition (DSR), the processing is distributed between the terminal and the network [1]. This is encompassed in Figure 4. In the DSR approach, the terminal hosts the feature extraction module, while the recognition is done in the server. The speech features are usually compressed to reduce the transmission bit rate, error protection is added and the resulting data stream is transmitted through error protected data channels.

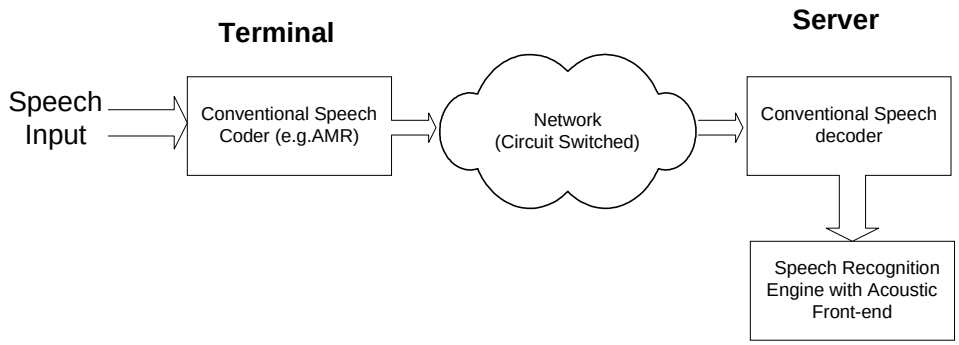

Figure 3 Illustration of Network Based Circuit Switched Speech Recognition

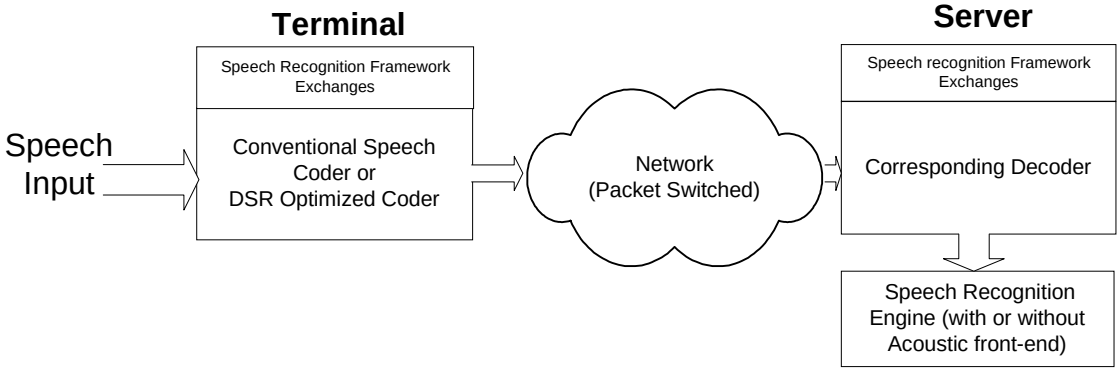

Figure 4 Ilustration of Speech Recognition Framework

Figure 5 shows the different components of a typical speech service, which may use either DSR or network based speech recognition and is explained below:

1. Voice Platform: The voice platform hosts the network side components needed for enabling a speech service.
2. Speech Recogniser: Speech recogniser includes the speech recognition back-end in the case of DSR or a complete speech recogniser in the case of a network based speech recognition system.
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.
6. Call Control: The call control module handles the telephony 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.

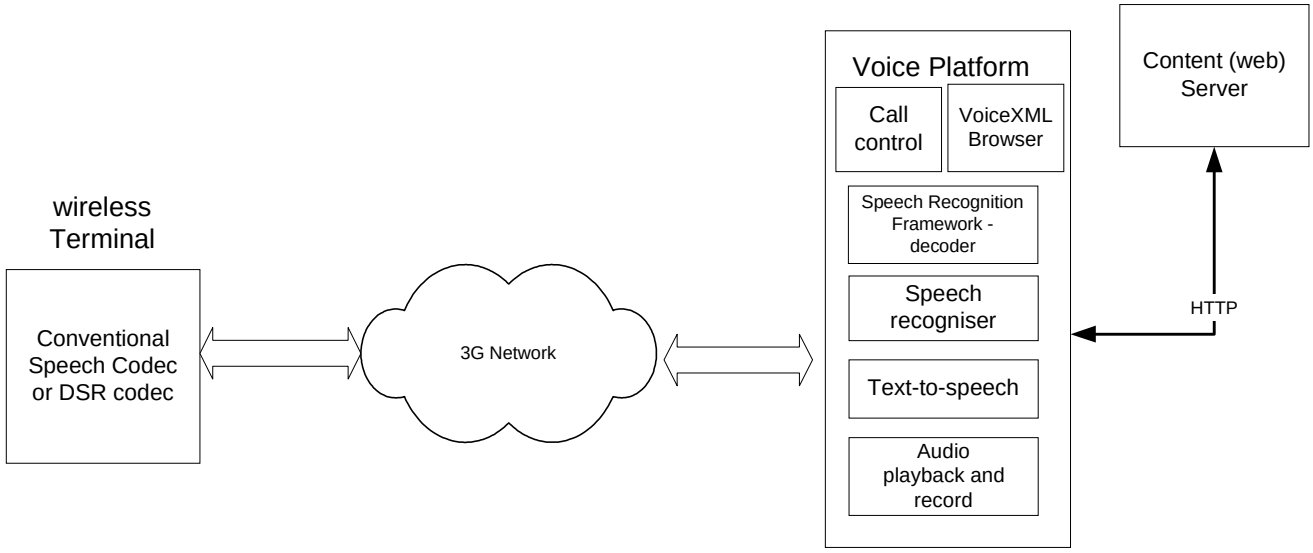

Figure 5 Illustration of the architecture for a typical speech service in a 3G network using network side speech recognition resources based on the speech recognition framework

6.1 DSR standards

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

1) ES 201 108 [3] 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 [4], the Advanced DSR front-end, was selected in Feb 2002 and will be published in July 2002. It provides improved robustness in background noise giving significant reduction in speech recognition word error rate compared to the Mel-Cepstrum in noise.

A set of evaluation databases have been established and used for the characterisation of the recognition performance of these front-ends. These databases cover both small vocabulary and large vocabulary tasks in a range of noises typical of those found in mobile environment. In addition the front-ends have been tested on 5 languages (Finnish, Spanish, German, Danish, Italian) from databases collected in a car environment.

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

An example of the computational complexities of the two standards are estimated in 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

Table 1: DSR Terminal side Complexity

7. Requirements to introduce Speech-enabled services

This section provides the high level requirements to enable speech and multimodal 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 offered over the IMS.

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

7.2 Information during the speech 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. Accordingly some terminals shall support sending additional data to the service (e.g. keypad information and other terminal events) and receiving data feedback 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)
- Presentation update instructions (as events or presentation mutations instructions).

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

7.3 Control

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

7.4 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.

8 Speech Recognition within 3GPP system

There are no speech-enabled services that are 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.

However, network based and DSR-based speech recognition technology utilise the network based resources. Network-based speech recognition performs speech encoding by conventional speech coders, which are already being included as part of the 3GPP specifications. 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. It is envisaged that a separate technical specification outlining the additional changes required to support speech-enabled services is needed to utilise DSR based speech recognition technology.

[Editors note: Scenarios how to apply Speech Recognition within 3GPP system need to be developed.]

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