



White Paper

Towards IMS: The Potential of VoiceXML for Multimedia Applications

Authors: Daniele Sereno, Paolo Baggia (Loquendo)

Date: July 25, 2006

Table of Contents:

1. Introduction	3
2. The evolution of IVR: beyond voice content	3
3. The potential of VoiceXML for multimedia	4
4. The network scenarios	5
5. Conclusions	6
6. Acronyms	7

1. Introduction

The availability of sophisticated mobile terminals with video capabilities, the increased amount of bandwidth for transmission available within UMTS networks as well as the appearance on the market of small footprint embedded technologies, are all relevant issues that foresee an increased demand of automatic telephone services not just limited to speech, but based on multimedia content. As a matter of fact, quite a number of companies which base their business models on standard IVR development are rapidly moving to address the new IVVR (Interactive Voice and Video Response) market opportunity.

In spite of a certain expected penetration of fixed network video-telephony, clearly those regions where UMTS is expected to get a higher penetration are also those where this market has the highest potential for development. In this respect, Europe and the Far-East are certainly among the most promising markets.

2. The evolution of IVR: beyond voice content

Interactive Voice Systems were conceived in the last century to retrieve information over the phone and so replacing or minimizing the work of human operators with the aim of reducing the cost associated with human operators. In spite of a continuous enhancement in the performance of state of the art ASR technology, the expected capability of an automatic recognizer to be robust to any condition and to allow a fluent vocal interaction with a machine, has not been completely achieved yet. Moreover the recent reduction of the costs associated with human operators, at least in some geographic areas, still delays the pervasive use of vocal applications in modern IVR. Nevertheless, in the last few years a boost in increasing the penetration of automated services has been made by the work of W3C in promoting VoiceXML and all the relevant associated standards in the speech area, like SRGS/SISR for grammars and SSML for text to speech. Despite this, the rate of increase of penetration of automated voice services is still not in line with expectations.

We are today probably facing a new opportunity for combining voice commands with multimedia contents. The VoiceXML, in spite of its name, can again be a key driver in boosting this emerging market opportunity.

The examples of the benefits of associating vocal commands with multimedia content retrieval encompass all those that are typically made for audio services but amplified by the expressive capabilities associated with a multimedia content. The possibility to easily access maps, stocks indexes, movies etc. with an audio/video terminal presents new requirements for more intelligent interaction with the device instead of using a sequence of numbers (i.e. DTMF). Let's consider, for instance, about searching for a movie without scrolling the complete list of clips available, or making a reservation for a trip or for a seat at a theater show without moving from menu to menu. At the same time, the option of superimposing a synthesized comment on a graphic or on a menu is an example of the potential of combining TTS with video or image content.

3. The potential of VoiceXML for multimedia

VoiceXML has proven it can provide a worldwide unified approach in developing advanced IVR services. This success has been extended also to graphic service creation environments (SCE) that exploit user-friendly graphic interfaces to build VoiceXML applications. Moreover, researchers in advanced VUI development are working on the introduction of an additional service layer to improve the development of advanced speech applications, which separates the application logic from dialogue management and media handling, but still leverages on the underlying VoiceXML to interface with a voice platform.

There are several reasons for adopting VoiceXML for describing automated services with multimedia content, in addition to the normal use in interactive voice services. The following are the most straightforward:

- VoiceXML is independent of the underlying media, therefore it is well suited to describe VUI and DTMF commands for any content transferred on the protocol transport layer.
- In spite of proprietary languages available in today's IVVRs that are based on DTMF commands only, VoiceXML is the choice to deploy mixed services with DTMF as well as voice commands, and for audio as well as video contents, leveraging on the same application logic conceived for typical audio contents.
- The inherent flexibility of VoiceXML allows for sophisticated video presentation description in XML, like those specified for instance in the SMIL framework where a higher degree of synchronization is allowed. The approach can be very similar to the exploitation of SSML elements, inside VoiceXML applications, for sophisticated description of TTS presentation.
- Because VoiceXML can manage both audio and video content, developers can conceive single applications working with conventional terminals (i.e. telephones) as well as terminals with advanced capabilities like video and/or embedded synthesis, thus avoiding the need to develop specialized applications and without the need to use different platforms specialized for audio and video.
- VoiceXML is a markup language that makes possible the development of advanced voice applications with sophisticated interactions with Web Applications (e.g. the use of the AJAX programming style by means of the use of <data> elements, dynamic concatenation of prompts, tapered prompts etc.).
- The evolution of VoiceXML to better manage video content is one of the priorities of the W3C Voice Browsing group which is currently specifying the VoiceXML 3.0 release.
- With the increasing potential of the IVVR market, the VoiceXML developer community, and in particular those companies developing vocal applications, are clearly motivated to also adopt VoiceXML to deploy multimedia applications, instead of using different languages specific to the video contents.
- Finally, the IMS (IP Multimedia Subsystem) scenario for fixed and mobile network convergence on IP, is intrinsically well suited to allow call control to be described in CCXML and dialogs to be represented in VoiceXML both for audio contents as well as multimedia.

Indeed, it is no surprise that a number of Voice Platform vendors are already adopting VoiceXML with some proprietary extensions to deal with the audio and video (multimedia) scenario.

W3C is working hard to not miss this important opportunity. VoiceXML 3.0, as it is already planned, must incorporate video as a possible media to be treated in the same way as audio at the application level. But time is a critical issue. The market is already asking for solutions and a delay in the availability of a standard for service creation could mitigate the potential of this new market segment. Even an incomplete VoiceXML 3.0, but available soon, could serve the development of the market better than the addition of new features that are made available too late.

4. The network scenarios

Already today a number of multimedia services are offered over fixed and mobile networks leveraging on the 3G-324M recommendation implementations, sometimes exploiting video gateway facilities which are provided by a number of vendors. The scenario is far more encouraging considering the IMS (IP Multimedia Subsystem) network where multimedia contents will be delivered over a packet switched networks.

The IMS is a standard architecture for next generation networking that has aroused strong interest from telecom operators wishing to centralize mobile and fixed multimedia services. The IMS architecture uses VoIP for voice transport with a pervasive use of SIP as the basic protocol for signaling, RTP for the transport, and that runs over IP and exploits open standards defined by IETF to convey ITU defined codec for audio and video. Originally conceived for mobile networks only, IMS has recently acknowledged the need for fixed/mobile convergence and now supports in its framework also fixed networks.

A functional and strongly simplified view of an IMS architecture is depicted in Fig. 1 where only those components relevant to the delivery of multimedia services are represented.

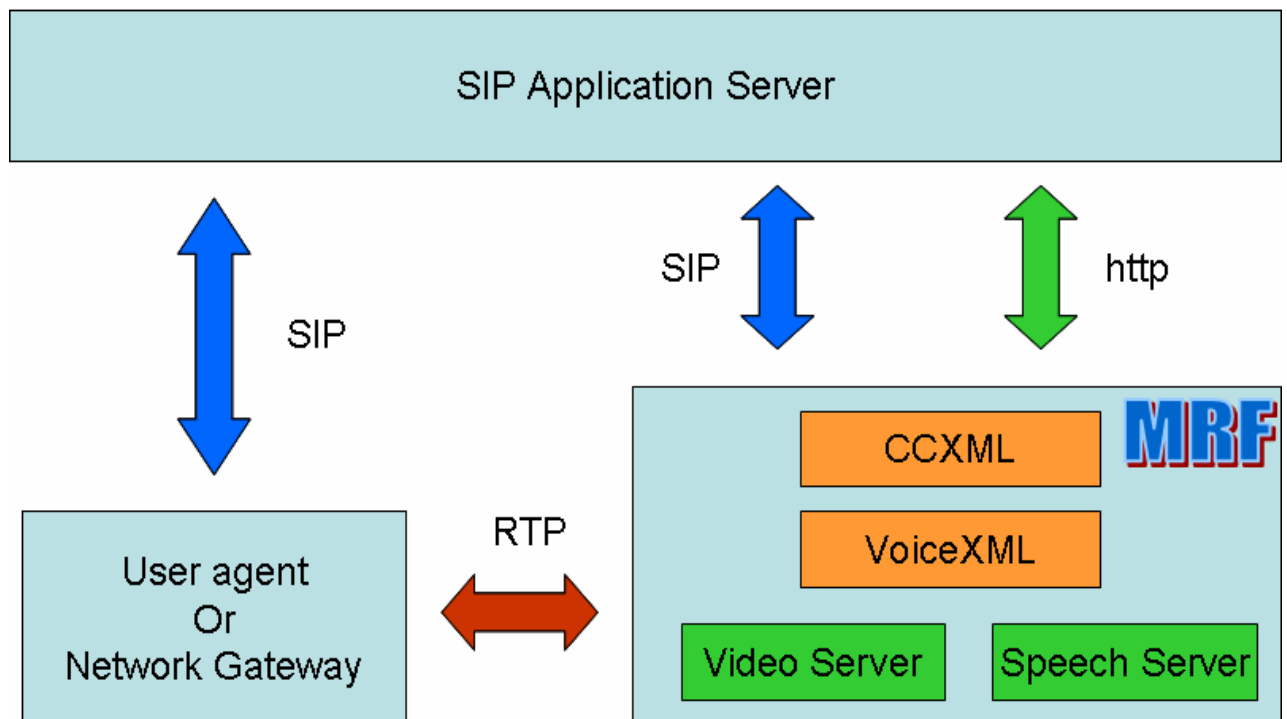


Fig. 1 – Simplified block diagram of IMS architecture

The SIP Application Server plays the essential role of hosting and executing services. Users can access these services directly as user-agents (for instance with a VoIP terminal either fixed or mobile) or through a gateway (when making the call from a circuit switched network). The Media Resource Function (MRF) is responsible for offering the media resources to provide multimedia automated services which extend the ancillary services of PSTN networks. These include for instance: playing multimedia announcements with or without TTS, automated services using ASR and/or SV, real-time transcoding of video streaming, multimedia conferencing, etc.

In consideration of the assumption to base interactions between components on open standards, the MRF includes a CCXML server, a VoiceXML server, a Speech Resource server and a Video Resource server. Services are therefore described in CCXML for call control and in VoiceXML for multimedia dialogs. CCXML and VoiceXML pages, as well as grammars and audio files, are fetched from the Application Server where the service logic is retained.

5. Conclusions

We believe the deployment of IMS, combined with the availability of UMTS bandwidth, will create new opportunities to develop the technology market in the area of automated multimedia services. Our platform solution is evolving to address this opportunity. Our CCXML server is fully compliant with the current version of W3C, and our certified VoiceXML interpreter has been extended to deal with multimedia content without requiring additional knowledge and expertise from our customers when developing vocal VoiceXML services. In particular a few elements and properties have been added to manage generic media content, to include sophisticated graphic presentations and to deal with push-to-talk possibilities. Moreover our basic technologies (ASR and TTS) are improving in performance and functionalities particularly useful for multimedia scenarios, like greater noise sensitivity also achieved by exploiting DSR, larger audio bandwidth representation and improved embedded performance.

6. Acronyms

3G	Third Generation (Mobile Communication System)
3G-324M	Umbrella specification for multimedia communication over circuit switched based on ITU-T H.324 and adopted by 3GPP
3GPP	3 rd Generation Partnership Project
AJAX	Asynchronous JAVaScript and XML
AS	Application Server
ASR	Automatic Speech Recognition
CCXML	W3C Voice Browser Call Control (http://www.w3.org/TR/ccxml/)
DSR	Distributed Speech Recognition (e.g. ETSI Aurora)
DTMF	Dual-Tone MultiFrequency
ETSI	European Telecommunications Standards Institute
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IP	Internet Protocol
ITU	International Telecommunication Union
IVR	Interactive Voice Response
IVVR	Interactive Video-Voice Response
MRF	Media Resource Function
PSTN	Public Switched Telephone Network
RTP	Realtime Transport Protocol
SCE	Service Creation Environment
SIP	Session Initiation Protocol
SISR	W3C Semantic Interpretation for Speech Recognition (http://www.w3.org/TR/semantic-interpretation/)
SMIL	W3C Synchronized Multimedia Integration Language (http://www.w3.org/TR/SMIL2/)
SRGS	W3C Speech Recognition Grammar Specification (http://www.w3.org/TR/speech-grammar/)
SSML	W3C Speech Synthesis Markup Language (http://www.w3.org/TR/speech-synthesis/)
SV	Speaker Verification
TTS	Text-To-Speech
UMTS	Universal Mobile Telecommunication System
VoiceXML	Voice Extensible Markup Language (for version 2.0: http://www.w3.org/TR/voicexml20/ , for version 2.1: http://www.w3.org/TR/voicexml21/)
VoIP	Voice over IP
VUI	Voice User Interface
W3C	World Wide Web Consortium