ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **13**, Number 3, 2010, 310–320

Multi-Access Network-Based TTS Architectures

Dragos BURILEANU[1,2], Mihai SURMEI[1], Cristian NEGRESCU[1], Catalin UNGUREAN[1], Aurelian DERVIS[1]

[1] *Politehnica* University of Bucharest, Faculty of Electronics, Telecommunications and Information Technology

[2] Romanian Academy Center for Artificial Intelligence, Bucharest, Romania

E-mail: bdragos@mESsnet.pub.ro

Abstract. There are many areas were text-to-speech (TTS) synthesis could enhance the end-user experience or even change the behavior of using existing network services starting from simple notification applications to complex presence enabled instant messaging convergent services. For this to happen we may need to bring TTS service in the multi-access area of Internet Multimedia Subsystem (IMS) networks and implement network-based TTS architectures. This paper proposes a pre-IMS network architecture leveraging a Romanian TTS system and a reference framework for developing future convergent network services around TTS technology.

Key-words: Multi-access, network-based speech processing, IMS network, MRCP, presence, instant messaging.

1. Introduction

There are various technology areas were ease of use is the most important characteristic. Generic speech technology is a logical choice for such cases as long as the complexity of implementing synthesis, recognition and speaker identification is assumed and fully understood.

Domains where speech technology fits could be completely independent, covering database information retrieval, car navigation, health or public administration. Nevertheless, the needed speech services have the same requirements to fulfill. Our

proposal is to consider speech technology as part of the telecom service enabler class, together with other generic enablers such as call control, user status, presence or unified messaging.

Adopting several access technologies is a mandatory stage of developing and deploying rich multimedia session communication, being one of the topics of the migration towards Next Generation Networks (NGNs) from traditional public switched and packet networks. If speech processing technologies such as synthesis and recognition are included in the already deployed network services list, a more natural, intuitive and flexible communication mean will be available to the demanding end-user.

Because text-to-speech synthesis is the kind of technology able to dramatically improve the usability of existing terminal devices, mobile or fixed, by adding flexibility and naturalness [1], there are many live implementations mainly offered by telecom or services operators, realized in various architectures and placed in different parts of the telecom layers.

One could imagine different service types based on the TTS synthesis technology:

- *Notify/warning services*: news reading or location related traffic information is a good example. This type of service means pushing the information to the common (voice only) user terminal.

- *Legacy terminal adaptation*: means reading e-mails or SMS's from old POTS (Plain Old Telephony System) terminals. In this case the user is popping the information from the e-mail server or from the Service Center and it could be applied for any text source.

- *Dynamic cross-media communications:* ability to switch from voice to text during the same communication session depending on the traffic events such as *presence status* change.

- *Accessibility*: developing voiced services for people with disabilities, by pushing or popping the information from the relevant servers.

Another dimension to take into consideration is the manner to implement such services.

- *Embedded systems*: the user terminal performs the entire speech synthesis task. It is clear that for complete in-device solutions, the bandwidth requirements is very low as only text will be carried out between peers; in addition, some of the services (e.g., notify/warning) are very well suited for this approach. However, there are also important drawbacks: computing and memory resource constraints, robustness in adverse environments, cost and power consumption limitations are serious difficulties to deal with [3]. Besides, it is not possible to implement some services (e.g., legacy terminal adaptation) and finally, it is not easy to ensure intelligence property for the software implementation.

- *Networked systems*: the speech processing is centralized and maintained by a trusted entity, such as a fixed/mobile telephony operator or an independent

service operator. The main benefits of this approach can be easily revealed: all of the previously discussed services could be implemented using this approach, the intelligence property is granted, and a new service could be easily deployed using the operator's channels. The obvious limitation is that a voice channel between user and speech server must be established, leading to higher communication costs.

2. Reference framework for network-based TTS

The effort of building up combined services using speech enablers could be minimized and more structure could be added in the development process if a reference architecture for speech synthesis (and recognition) in Romanian language [2] is defined, starting from the following basic characteristics:

- Using of open protocols.
- Ensure the interworking with other elements from the network by following latest developments in the telecom field.
- Possibility of smooth implementation of sub-architecture by implementing part of the functionalities.
- Possibility to build and manage the speech loop: *recognition* $\leftrightarrow$ *synthesis*.
- The processing will rely on streaming; any deferred messaging or file manipulation method should be avoided. The consequence of this will be the real-time nature of the resulting combined services.

Continuing the view of the reference architecture it is natural to imagine a pure reference service to be executed on such architecture.

The reference service will have the following characteristics:

- Any terminal able to sustain a secure connection towards internet:
  - cellular phones, capable to sustain a GPRS (General Packet Radio Service)/EDGE (Enhanced Data for GSM Evolution)/3G (Third Generation)/HSDPA (High-Speed Downlink Packet Access) or LTE (Long Term Evolution) connection;
  - fixed soft-clients running on PC;
  - soft-clients running on Wi-Fi enabled terminals.
- The terminal will send text and receive the corresponding voice channel with the TTS conversion result in the same session and in real-time.
- Underlying network will consist of existing GSM/WCDMA or fixed infrastructure with an independent SIP (Session Initiated Protocol) signaling layer on top of it. This so called *overlay network* (the terminal does not support natively SIP

based technology) will provide the TTS functionality. If the operator already has a SIP infrastructure, this could be reused as well.

As one can notice in the above definition of the service, the communication model *text-voice-text* involves two different media types mixed together in the same session and cross-media transcoding is needed. The most suited network able to cope with these requirements in a standardized way is IMS (Internet Multimedia Subsystem).

The reference architecture and any particular instance of the reference service will be realized using publicly available software stacks (e.g. Unimrcp) and development frameworks (e.g. Apache APT), running on off-the-shelf hardware.

3. Pre-IMS TTS network architecture

The scope of our research is to have a proper test network to develop on by integrating the existing IMS functionalities, not to implement a full-fledged IMS network. Furthermore, the non-real-time applications are still required by the market and this is the reason we have developed in the first phase two such services, as will be next described.

The architecture depicted in Fig. 1 includes a SIP server which has two roles in a generic SIP system: the *Registrar* and the *Application Server* (AS) [4]. The GGSN (Gateway GPRS Support Node) is the base network for GPRS/3G. The SIP signaling is granted through the IP infrastructure connecting terminals and network nodes (GGSN, SIP AS/Registrar, MRCP (Media Resource Control Protocol) client and MRCP server serving the TTS engine) present in the conversation [6]. The voice payload is transported using RTP (Real-Time Protocol) [5].

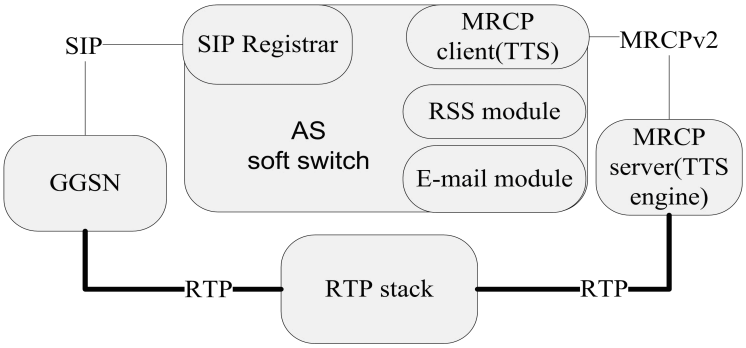

Fig. 1. TTS enhanced VoIP architecture.

We developed two TTS-based services on the classic VoIP (Voice over IP) architecture: an *e-mail* and an *RSS reader*.

Today, e-mail is maybe the most common form of electronic communication. However, the convenience of this particular form of sharing information is limited by the necessity of an Internet connected computer. A service that allows for accessing e-mail by phone enhances the user experience, since information can be retrieved practically

anytime and from everywhere. TTS technology can play an important role in this network-based service.

Furthermore, the communication network developers are interested in the integration of voice-enabled solutions in the services offered over the telephone line. For example, the information or mass media professionals, and not only, use as source of information and news in a very dynamic mode a family of web feeds known as RSS (Really Simple Syndication, or Rich Site Summary) feeds. Thereby the existing RSS information channels could become more valuable by combining RSS and text-to-speech synthesis integrated in a telecom network service.

In the proposed application architecture (briefly described in Fig. 1), we have the following functional entities:

1. SIP overlay network, consisting of three main units:
   (a) The *soft switch*, which implements:
      i. SIP signaling for registration, call setup and DTMF (Dual-Tone Multi Frequency): the *SIP Registrar* and SIP *AS* (Application Server);
      ii. *RTP stack* for voice payload;
      iii. *MRCP client* acting as TTS client;
      iv. HTTP-based *RSS module* for fetching the web streams;
      v. *E-mail module*, capable of both IMAP and POP3 protocols. Building-up modules written in a variety of languages could extend the soft switch used. This particular e-mail client was implemented using Perl.
   (b) *MRCP server* exposing the TTS engine (our TTS synthesis system for Romanian language, partially described in [2]), which implements the server side of the MRCPv2 protocol.
   (c) SSL (Secure Sockets Layer) VPN (Virtual Private Network) server used to tunnel the TCP (Transmission Control Protocol) and UDP (User Datagram Protocol) traffic from the SIP overlay network directly to the mobile phone. For our particular need the tunnel is based on UDP in order to improve the responsiveness of the system by eliminating the TCP overhead.
2. Mobile network: the communication between the mobile phone and the SIP AS from SIP overlay network could be supported by any of the existing GPRS networks.
3. Mobile phone: currently we are using Windows Mobile based smart phones, with SSL VPN and VoIP clients.
4. Internet: to access e-mail or RSS servers.

The e-mail and RSS services consist of the following stages:

1. The WM6.1 VPN client must be initially installed on the smart phone so the user can initiate in the first stage an SSL tunnel.

2. In the second stage, with the connection established and having an IP belonging to the SIP overlay network on the smart phone, will be possible routing to the SIP Registar and SIP AS.

3. The end-user starts the SIP client previously installed on the smart phone. Automatically it will perform the SIP registration. At the end of the process, the SIP client will be authorized to perform or receive calls within the SIP overlay realm.

4. By dialing a predefined extension, the SIP client will send an INVITE to the SIP AS that will trigger the RSS reading or the e-mail reading processes.

5. The RSS module fetches the latest news for the preconfigured feeds, or the e-mail reader will connect and fetch the status of the email account (folders and new messages).

6. SIP AS initiates a MRCP dialogue with TTS server, extracting the relevant text depending on the service used.

7. The TTS server opens the first RTP leg towards the soft switch RTP port (the codec used for this stream is G.711 PCM A-law 8 kHz);

8. The SIP AS opens the second RTP leg towards the smart phone, using GSM codec; then it will transcode and relay the stream from the first RTP leg.

This platform permits concurrent client accesses, as each of the network functional entities are multi-threaded. As long as a MRCP client is available, it is possible to choose from various SIP soft switches by implementing the MRCP protocol.

In terms of performance, the following settings had an impact of the quality of the voice:

- SSL tunnel based on UDP;
- ciphering was set to "off";
- the RTP stream from smart phone to SIP overlay was disabled;
- the internal TTS audio coding of the voice was recoded as G711 PCM A-law 8 kHz.

4. IMS-based TTS

The pre-IMS architecture does not allow us to properly mix the needed service capabilities. Only using an IMS architecture makes possible to combine in a structured way several network capabilities [9, 10].

The architecture presented in Fig. 2 illustrates the main functional roles of an IMS network: SBC (Session Border Control) to sustain the mobile terminal connectivity

through the operator's firewall layers, HSS (Home Subscriber Server) used for subscriber authentication and authorization, CSCF (Call Session Control Function) to route the calls and the AS (Application Server) for the execution of the service itself, together with other telephony services, MRFC (Media Resource Function Control) and MRFP (Media Resource Function Processor) that will implement for instance the old IVR (Interactive Voice Response) or Voicemail functionalities, similar to the classical GSM or PSTN networks.

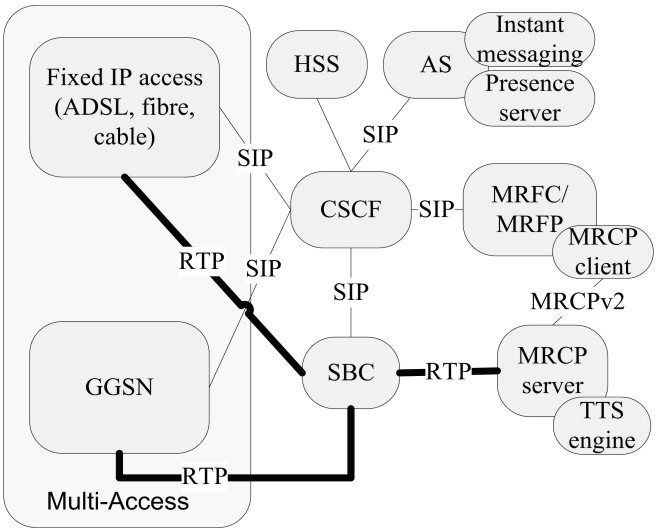

Fig. 2. TTS service in IMS context.

The TTS functionality is by definition a hybrid functional node between signaling and payload. Therefore is closely located to the MRFP/MRFC pair. The telecom environment is trying to standardize protocols in speech technology (both synthesis and recognition). The most advanced one is the MRCPv2 protocol referred by 3GPP documents (TR 24.880).

Although most of the services based on TTS follow the client-server architecture, they belong to one of the next two categories:

- are not real-time, such as in e-mail or RSS reader applications;
- ask for supplementary pieces of software (for example proprietary interfaces) at the client-side, and are usually tailored for a limited set of devices [7].

Reading e-mails or RSS's is very useful, but they are not real-time and therefore limited. The true challenge consists in real-time services and their strict design requirements.

Using proprietary interfaces leads to difficulties in integrating with the rest of communication network and poses a distinct threat for the owner of the network to get locked-in on a particular technology. Furthermore, observing the overwhelming

evolution towards an "all IP" ecosystem, any solution based on legacy technologies such as TDM (Time Division Multiplex) to convey the resulted voice could be already considered obsolete.

In order to avoid the above mentioned limitations one could decide to embed the speech technology into the end-user terminals [1]. However, besides the increased processing power and storage-needed capabilities, another drawback of this approach is the over-specialization of the devices, transforming them in niche terminals. Furthermore, an automated update mechanism is needed for the TTS-enabled terminals in order to have the latest TTS algorithm installed, the update server being centralized and available via internet.

Our approach consist of treating the text-to-speech conversion (together with a voice recognition process), as a fundamental component in the telecom service space and enrich the core network with speech capabilities. This approach will give structure and will lead to a canonical way of building innovative telecom services [3, 8].

5. IM-Presence-TTS convergent service

The way of communicating is shifting towards a multi-session convergent experience and IMS networks expose the needed services to make this change possible. Our team target is to develop a specific multimedia telecom service by mixing voice, presence and text-to-speech synthesis on the previously proposed reference architecture.

This real-time service allows changing an already established text chat session between two users into a hybrid communication instance: one of the users will continue chatting, the other will hear and reply using voice, and everything will happen without session interruption. The trigger to change the session from text-based to voice-based with cross-media transcoding is for instance the presence status of one of the users.

The described service is to be consumed using mobile terminals (or to a lesser extent via PC clients). Designing an all-purpose IMS mobile terminal client is beyond the scope of our research and is definitely a bleeding edge technology of the moment. Nevertheless, several clients covering VoIP, IM and presence already exists and they are running on Symbian or Windows mobile smart phones. Besides that, there are several free developing platforms that allow building custom IMS clients for the mobile operating systems mentioned before.

This service uses three distinct telecom services during the same session: *instant messaging*, *voice* and *TTS conversion*.

The following steps describe the most important phases of the proposed service:

1. First we suppose both users are already registered in the HSS. The registration is a mandatory mechanism needed to support the routing of the calls between users.

2. The presence status of both users is defined for instance as "Walking" and defined as such in the application server that will play the role of a presence server.

3. Users A and B alternatively send text messages using the IMS clients installed on the mobile phones. Because we are developing a conversational service, using the *SIP Message* (RFC 3428) is not enough, the alternative being a SIP extension for session oriented messaging: the *Message Session Relay Protocol* (MSRP, RFC 4975).
4. User A changes the status to "Driving".
5. When user B sends a text message, the AS will reroute the text to the MRCP client, based on the new presence state of user A.
6. MRCP client queries the TTS engine, hosted on the MRCP server, using MRCP protocol for:
   (a) converting the text message;
   (b) opening an RTP flow towards the called subscriber. The RTP flow does not belong to the overlay network, as it is not signaling; it will represent the content of the conversation.
7. Additionally, the flow could be continued and the loop closed using a speech recognition engine, implemented on the same platform as the TTS engine, being part of the overall MRCP proposal. Therefore if user A wants to answer back, he will do this by speaking to the terminal, therefore opening another RTP flow up to the MRCP server. The ASR engine will route the resulted text message to user B, via the AS.

Figure 3 shows the call flow for this convergent service, in the general case of using both TTS and ASR technologies on the same MRCP Server.

6. Conclusions

Our first implementation of a network-based TTS platform is a pre-IMS realization, using a minimal number of support nodes: a SIP application server collocated with the registrar and the RTP stack for the payload, therefore a monolithic approach. Nevertheless, the existing implementation is a multi-threaded platform and permits several concurrent calls to take place. The technology used is open source: Unimrcp for the MRCP stack and FreeSwitch for the soft switch.

The services currently running on the pre-IMS platform are RSS and e-mail reader. These non real-time services are still demanded by the users and speech technology makes them even more appealing. Our target is to evolve to a full IMS environment and be able to orchestrate several services such as instant messaging, presence and TTS in one converged service.

An important aspect of convergent services is the access. We are considering all the major access types such as mobile, fixed or broadband based, not only mobile access with smart phones terminals. Even if the IMS environment is by excellence access agnostic, the service orchestration could bring the need of developing an appropriate client for a particular type of access.

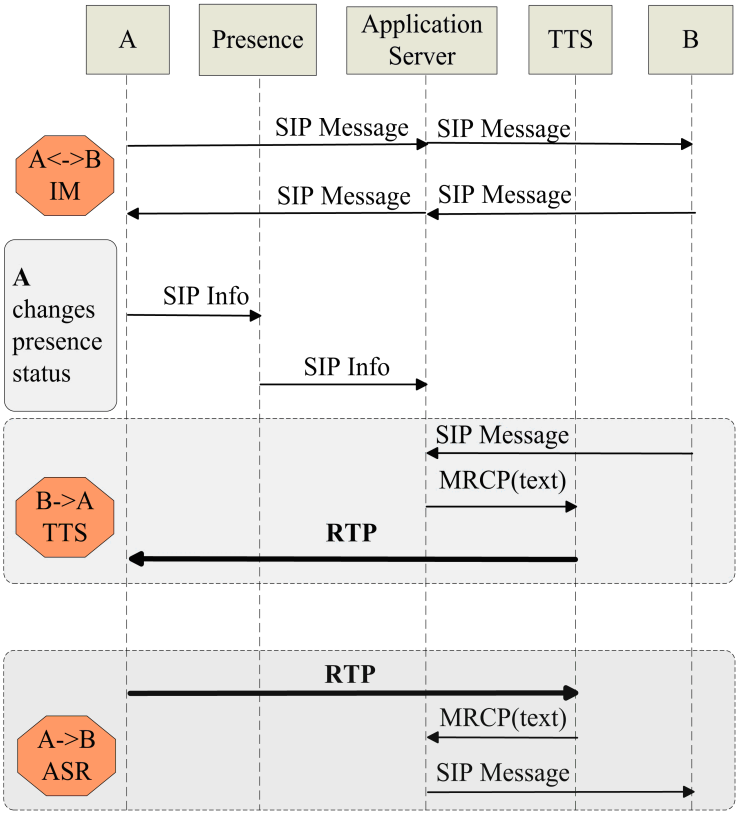

Fig. 3. IM-Presence-TTS call flow.

Acknowledgements. The research reported in this paper was funded by the Romanian National Research Authority CNCSIS, Grant "IDEI" no. 782/2007.

References

[1] Burileanu D., *Spoken Language Interfaces for Embedded Applications*, in *Human Factors and Voice Interactive Systems* (D. Gardner-Bonneau and H. Blanchard – Eds.), 2^{nd} Edition, Springer US, New York, pp. 135–161, 2008.

[2] Burileanu D., Negrescu C., Surmei M., *Recent Advances in Romanian Language Text-to-Speech Synthesis*, Proceedings of the Romanian Academy, Series A – Mathematics, Physics, Technical Sciences, Information Science, vol. **11**, no. 1/2010, Publishing House of the Romanian Academy, Bucharest, pp. 92–99, 2010.

[3] Rondel S., Pattabhiraman P.T., Ganapathiraju A., Khademi P., Rondel J., *Strategic Importance of Speech Technology for NGNs*, White Paper, Conversational Computing Inc., Redmond, WA, 2007.

[4] Rosenberg J., Schulzrinne H., Camarillo G., Johnston A., Peterson J., Sparks R., Handley M., Schooler E., *SIP: Session Initiation Protocol*, RFC 3261, June 2002.

[5] Schulzrinne H., Casner S., Frederick R., Jacobson V., *RTP: A Transport Protocol for Real-Time Applications*, STD 64, RFC 3550, July 2003.

[6] Shanmugham S., Burnett D., *Media Resource Control Protocol Version 2 (MRCPv2)*, Draft-ietf-speechsc-mrcpv2-12, March 2007.

[7] Shimizu T., Ashikari Y., Sumita E., Kashioka H., Nakamura S., *Development of Client-Server Speech Translation System on a Multi-Lingual Speech Communication Platform*, Proceedings of the International Workshop on Spoken Language Translation, Kyoto, pp. 213–216, 2006.

[8] Surmei M., Burileanu D., Negrescu C., Pirvu R., Ungurean C., Dervis A., *Text-to-Speech Engines as Telecom Service Enablers*, in *Advances in Spoken Language Technology*, Publishing House of the Romanian Academy, Bucharest, pp. 89–98, 2007.

[9] Svendsen T., Egeberg A., Holter T., Skogstad T., *VOCALS – Voice Centric User Interfaces for Location Based Services*, Proceedings of the 6^{th} Nordic Signal Processing Symposium – NORSIG'05, Stavanger, Sept. 2005.

[10] TalafovÁ R., Rozinaj G., epko J., Vrabec J., *Multimedia SMS Reading in Mobile Phone*, International Journal of Mathematics and Computers in Simulation, vol. **1**, issue 1, pp. 12–17, 2007.