THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 11, Number 1/2010, pp. 92–99

RECENT ADVANCES IN ROMANIAN LANGUAGE TEXT-TO-SPEECH SYNTHESIS

Dragoş BURILEANU[*, **], Cristian NEGRESCU[*], Mihai SURMEI[*]

[*] "Politehnica" University of Bucharest, Faculty of Electronics, Telecommunications and Information Technology
[**] Romanian Academy Center for Artificial Intelligence, Bucharest
Corresponding author: Dragos BURILEANU, E-mail: bdragos@mESsnet.pub.ro

Spoken language interfaces are currently playing an increasing role in the human-machine interaction, becoming a necessity for most of the new practical applications and services demanded by our modern and "mobile" world. This is mainly a consequence of the fact that communication networks may offer simple and inexpensive access to a large amount of diversified information and services, concurring to the development of many economic and social domains from the Information Society. In particular, machine's voice output is already largely required to expand the potential of the commercial applications, adding flexibility, speed and naturalness to existent interfaces. This paper describes our recent work in developing a high-quality text-to-speech synthesis system in Romanian language and presents a telecommunication platform based on a client-server architecture and standardized signaling protocols for accessing text information through communication channels.

Key words: Text-to-speech synthesis; Non-uniform unit selection; Harmonic plus Noise Model; Partial processing; Client-server architecture; RSS reader; Open protocols; MRCP; SIP/RTP.

1. INTRODUCTION

Two important characteristics of our Information Society can be easily noticed nowadays: *mobility*, expressed by the widespread use of portable electronic devices, and the steady users' demand for a more simple, natural and effective interaction with their portable systems, to get easier access to information, any time, and from everywhere. On the other hand, speech technology can offer the simplest and most natural interface to a computing environment, allowing for hands-free and eyes-free operation and for a wider access to information and services. As a result, and due to the significant advances in spoken language processing brought in the last two decades, lots of speech-enabled applications together with new families of intelligent and interactive services based on voice input/output became available [5, 8].

Particularly, machine's voice output is largely required today to access a variety of services in network-based applications. Speech synthesis technology can be a viable option for fast, easy and efficient access to text messages using communication networks. Certainly, a service that will facilitate the access to a web feed or to the messages stored on the e-mail server (for example) using the usual phone line and a portable device brings a supplementary value by its new mobility dimension.

However, reading aloud written messages such as news feed or e-mail/SMS encounters two major problems. First, in this kind of applications one cannot predict the message to be spoken, so the system must generate the speech from arbitrary texts (database records, e-mail/SMS messages, etc.). This task can be accomplished only by a complete text-to-speech (TTS) synthesis system, which must provide at least a very good intelligibility for the resulting speech to be helpful and accepted by the user [4]. Then, one must choose between two possible approaches: settling the whole TTS application on the mobile device, or using standard communication channels and a client-server architecture [1, 10]. Due to computing and memory resource constraints and cost and power consumption limitations of the complete in-device solution, the second approach is more often preferred in present.

Besides the difficulty of developing highly natural synthesis systems, industry speech application developers are making notable efforts to design and propose good quality TTS solutions; several companies already use TTS synthesis in providing diverse information to users over the telephone line.

These TTS commercial systems offer a good intelligibility and an acceptable naturalness of the synthetic speech; most of them are using unit concatenation techniques and large acoustic databases. At the same time, there are also many obvious limitations, the major unsolved task remaining the prosody. Although many systems show an acceptable prosodic contour, the naturalness of the synthetic voice is still far from human speech: incorrect accents, unnatural pauses, inconsistent melodic profile for long sentences, and no option to change the style of utterance [13]. Improving the extraction of meaningful linguistic information encoded in the input text and more efficient synthesis techniques are still open research areas.

The main purpose of this contribution is to outline our recent achievements in developing a high-quality TTS synthesis system in Romanian language, emphasizing the new signal processing stage, based on a non-uniform unit selection procedure and a Harmonic plus Noise Model of speech for segment concatenation and prosody generation. As a particular application, we present the design and implementation of a web feed reader platform for mobile phones, based on a client-server architecture, standardized signaling protocols, and TTS synthesis.

2. A TTS SYNTHESIS SYSTEM IN ROMANIAN LANGUAGE

Our research collective works from a number of years in the text-to-speech synthesis domain. Several versions of a Romanian language TTS system were built successively in order to improve the performance of different constituent modules and consequently enhance the quality of the system.

The architecture of the current version is depicted in Fig. 1. The system is based on acoustic segment concatenation and uses multiple instances of non-uniform speech units (*diphones* and *polyphones* – to solve a number of difficult vowel-semivowel transitions), labeled (off-line) according to contextual and phonetico-prosodic information from the recorded speech corpus. It consists of a natural language processing stage (input text analysis, letter-to-phone conversion, and prosody estimation), which provides the phonetic transcription of the input text, together with prosodic marks and auxiliary contextual and phonetic information, and a signal processing stage, which transforms the information received into speech. The system uses a two-stage unit selection procedure: first, the available instances are successively filtered, based on the comparison of stored linguistic information and similar data extracted (on-line) from the input text; then, a distance measure is used to find the optimal sequence that minimizes spectral distortions at unit boundaries. We want also to highlight the presence of an automatic diacritic restoration module, acting before text analysis, which carry out the automatic restoration of missing diacritics from texts in Romanian, if necessary; this feature is particularly useful in applications that need to process electronically stored texts.

Because most of the modules in the language processing stage have been described elsewhere ([2, 3], and [14]), the remainder of this section will focus on the signal processing stage in our TTS system and will discuss the synthesis technique used for segment concatenation and prosody generation.

The most popular technique used today in TTS synthesis is the TD-PSOLA (Time Domain – Pitch Synchronous Overlap Add), due to its simplicity in achieving time/frequency-scale modification associated with very low computational costs. The perceived quality of the speech signal generated by using this method is fairly good, but is strongly dependent on the precision in estimating the pitch period. Moreover, the fact that the concatenation of the acoustic units is performed in the time domain disables the possibility to maintain the phase coherence around the concatenation point. In addition, TD-PSOLA has no straightforward mechanism for adapting the spectral envelopes of the acoustic units, which is known to be a major drawback of this non-parametric method. On the other hand, the solutions developed in the frequency domain can cope with the situations described before, leading to better quality of the synthesized signal, while increasing the computational load of the model (to an acceptable level).

For the current version of our TTS system we chose to use a Harmonic plus Noise Model (HNM), which is a technique developed in the frequency domain, based on the more general concept of splitting the signal $s(t)$ into a sinusoidal part $s_h(t)$ (which contains the quasiperiodic components), and a stochastic part $s_n(t)$ (which accounts for the nonperiodic signal components) [6, 12]:

$$s(t) = s_h(t) + s_n(t)\,. \tag{1}$$

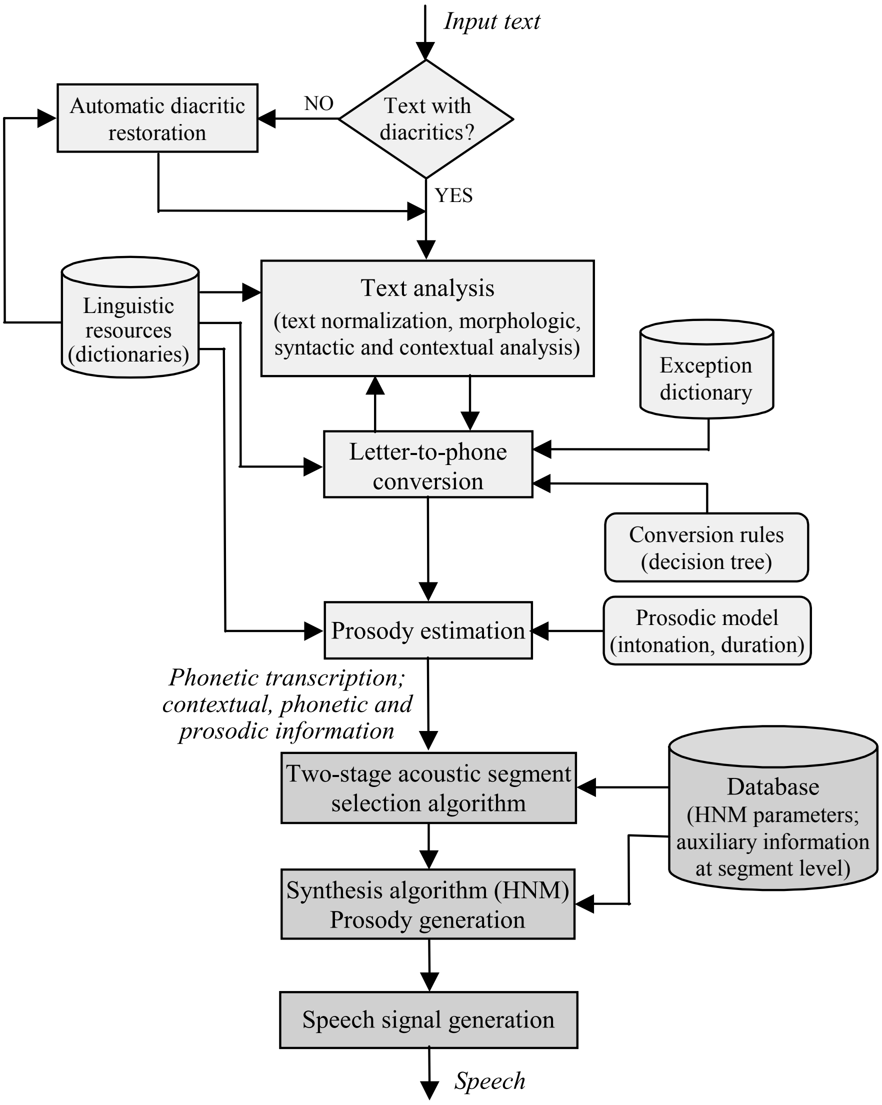
Input text
Text with diacritics?
NO
Automatic diacritic restoration
YES
Linguistic resources (dictionaries)
Text analysis
(text normalization, morphologic, syntactic and contextual analysis)
Exception dictionary
Letter-to-phone conversion
Conversion rules (decision tree)
Prosody estimation
Prosodic model (intonation, duration)
Phonetic transcription; contextual, phonetic and prosodic information
Two-stage acoustic segment selection algorithm
Database
(HNM parameters; auxiliary information at segment level)
Synthesis algorithm (HNM)
Prosody generation
Speech signal generation
Speech

Fig. 1 – The TTS system architecture.

This model assumes that the distances between the frequencies of the sinusoidal components are equal and are set in agreement with the estimated fundamental frequency. The model provides a suitable parameterization of the speech signal which is not only in agreement with the speech production mechanism, offering the possibility to reconstruct the original signal with high quality, but also allows a large flexibility in managing the characteristics of the signal and therefore helps speech transformations such as time/frequency-scale modification, which is fundamental for prosody generation in TTS synthesis.

The whole set of model parameters are divided into two categories, describing the harmonic part and respectively the noise part of the signal: $\mathbf{P}_h(t_{ha})$ and $\mathbf{P}_n(t_{na})$. Regarding the harmonic part of the signal, we adopted the well-known sinusoidal model, which assumes that the (harmonic part of the) signal actually is a sum of sinusoidal components (also known as *partials*) with continuously varying parameters. A partial is characterized completely by the triad of time-domain functions: the instantaneous amplitude, $A_i(t)$, the instantaneous frequency, $\omega_i(t)$, and the instantaneous phase, $\theta_i(t)$:

$$s_h(t) = \sum_{i=1}^{P(t)} A_i(t)\cos\left(\theta_i(t)\right), \tag{2}$$

where $\omega_i(t) = \frac{\mathrm{d}}{\mathrm{d}t}\theta_i(t)$.

Regarding the reconstruction of the noise part from the original speech signal, we also used a classic approach based on linear predictive analysis. The parameters $\mathbf{P}_n(t_{na})$ in this case are the LPC coefficients and the gain factors that will allow restoring the loudness of the original signal.

In order to validate the model, first we provide a block diagram (Fig. 2), which will describe only the analysis and synthesis procedures. Since the speech signal is a non-stationary process, the analysis block operates framewise, using a Hanning analysis window with duration of approximately twice the maximum searched pitch period, placed around the chosen analysis moments t_a. Each analysis frame is subjected to a harmonic analysis stage, in order to extract the optimum number of partials that best characterize the periodic part of the speech signal. In the next stage, considering that no time scaling is involved, the synthesized harmonic signal $\hat{s}_h(t)$ around synthesis moment $t_{hs} = t_{ha}$ is subtracted from the original signal and the noise part $s_n(t)$ of the speech signal is obtained. Once the parameters that model the noise part are computed, the noise synthesis procedure is initiated to produce an estimate for the noise part $s_n(t)$ of the original speech signal. Let us remark that usually, the analysis moments for harmonic part of the speech signal are different from the analysis moments for corresponding noise ($t_{ha} \neq t_{na}$).

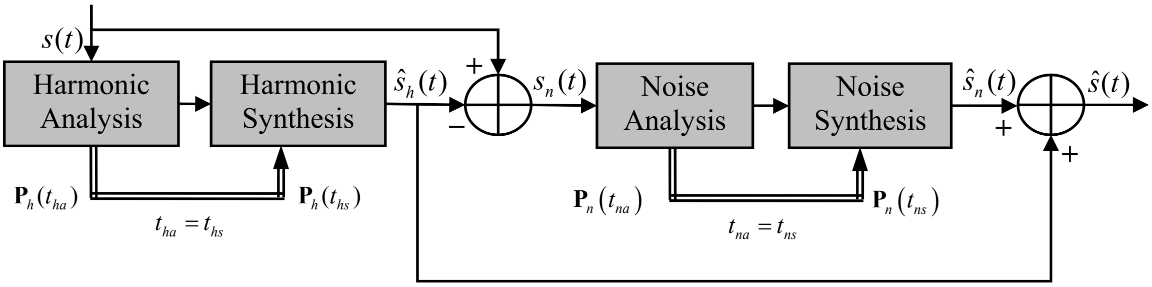

Fig. 2 – Analysis by synthesis based on HNM – block diagram.

The harmonic analysis routine addresses two very important tasks: initial partial extraction, followed by the selection of a subset of partials. The role of the first step in the analysis routine is to provide a set of partials for each original speech analysis frame $s(t)$. In our particular implementation, the process of partial extraction is based on the new concept of *instantaneous frequency* (IF) *attractors*, which provides more accurate results compared to the classical peak picking method [7]. During this process, a very large number of partials that characterize the whole speech signal (periodic plus noise), is extracted. The usage of this set of partials would allow producing a high quality synthesized signal, if we apply solely the classical sinusoidal model, but the associated computational costs would be extremely high. Moreover, such an approach would impair the perceived quality of the pitch/time-scaled signals, since the harmonic and noise parts should follow different processing paths. These two aspects account for the following step of our analysis routine, which is responsible with selecting the optimum number of partials corresponding to the harmonic part of the speech signal. This stage involves a series of operations performed on all the partials extracted at each analysis frame: tracking generation (creates links, if possible, between partials from successive frames), removing short tracks and/or isolated partials (refines the rough estimate of the initial tracks and also accentuates the harmonic part of the signal), fundamental frequency estimation, harmonic part separation (saves only the partials corresponding to the quasiperiodic part), completion of missing partials from harmonic part, frequency correction of the partials (alters the frequencies of the quasiperiodic components in order to match exactly multiples of the estimated fundamental frequency). The last two operations have a significant impact upon the prosody modification stage and grant us the possibility to shift the frequencies of harmonic components in a coherent manner, while keeping the distortions to a low level.

The stage performing the harmonic analysis is the most challenging part of the model, since it has a significant impact on the requirements and performance for the rest of the processing scheme, on the computational load of the whole algorithm, and of course, on the overall quality of the final synthesized speech signal. A poor estimation of the initial set of partial’s parameters or inappropriate cross-connections during the tracking algorithm can lead to an unfortunate separation between the harmonic and noise part, which can further compromise the algorithms for time/frequency-scale modification.

The synthesis of the harmonic signal follows the same ideas formulated by the classic sinusoidal model. It uses linear interpolation for generating the amplitude functions and cubic interpolation for the phase functions [6]. It should be noted that in this stage a partial is associated with a harmonic controlled oscillator, characterized by the same triad of parameters. This type of synthesis eliminates by default the discontinuities around the concatenation points between successive acoustic segments. The noise analysis and synthesis are performed using known algorithms and no special attention is required for these operations.

Once the original speech signal is decomposed and accurately modeled using the two subsets of parameters, we are able to perform the desired task of time/frequency-scale modification, using the system presented in Fig. 3 and according to the designed prosodic model [3].

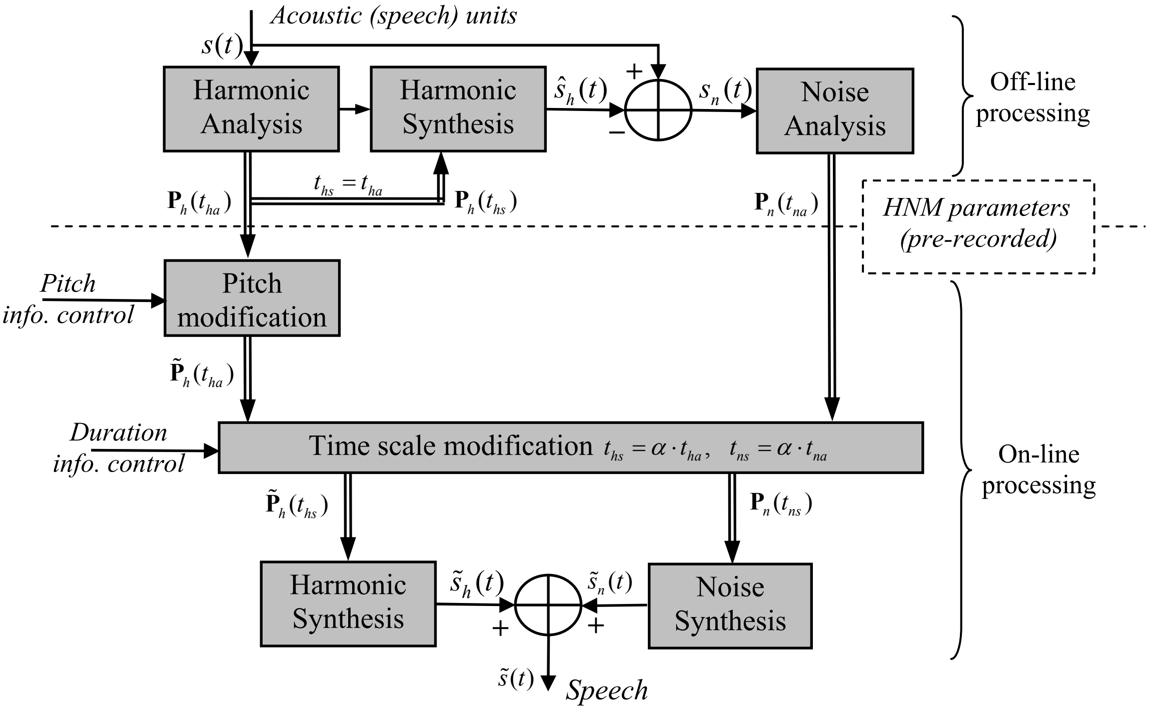

Fig. 3 – HNM – the core of the signal processing stage in the TTS system.

When pitch modification is activated, we change only the frequencies of the partials, keeping unmodified the noise part of the signal (which caries no information about pitch). Furthermore, if the desired effect is the modification of the acoustic segment duration in the synthesized speech, the time-scale modification is involved for both harmonic and synthesis parts, by changing only the values of the synthesis moments (t_{hs} and t_{ns}) while the actual sets of parameters remains unchanged [12]. The combination of the two-presented mechanisms gives us a separate control of the duration and pitch for the reconstructed speech.

To decrease the computational requirements, by reviewing the overall processing scheme for the TTS system (Fig. 1), one can remark that some of the processing stages can be pre-performed. Indeed, when the set of acoustic units is given, the harmonic analysis together with the corresponding synthesis, the noise part computing and also the noise analysis are performed in advance (only once), and the obtained HNM parameters are stored in the database. In the online synthesis procedure, the prerecorded HNM parameters are used directly as input parameters for the requested acoustic units.

A final consideration is devoted to the concatenation issue, which usually raises difficult problems to all TTS systems based on speech unit concatenation. When the input signal belongs to the same unit, the way in which the time and pitch-scale are performed assures that no significant audio artifacts appear. When the input signal belongs to different units, the native interpolation performed inside both harmonic and noise synthesis blocks in our approach maintains also a very low level of artifacts and therefore a very high quality of the synthesized speech is obtained [7].

3. AN RSS READER APPLICATION BASED ON TTS SYNTHESIS

As discussed in the introductory section, integration of voice-enabled solutions in the interactive services delivered over telephone line is at present an important concern for communication network developers. As an example, one of the most dynamic and useful source of information and news are represented by a family of web feeds commonly known as RSS (*Really Simple Syndication*, or *Rich Site Summary*) feeds, normally used by information or mass media professionals. Therefore having a service combining RSS, text-to-speech synthesis and telecom network integration makes perfect sense and brings value to the existing RSS information channels.

Starting from the Romanian language TTS system described in the previous section, our collective developed an RSS reader platform based on a client-server architecture and standard communication protocols. The rationale behind these two options is briefly explained below.

Unlike the complete embedded solution, a centralized platform offers several advantages: the possibility of granting the reliability of the service; the end-user is able to access the basic functions of the terminal services, leading to small costs, and the alternative of deploying new services without influencing the user. Also, we chose to use *open protocols* because this approach guarantees the interoperability of the particular product with any other application platform that implements the same industry standards. Actually, this is the current trend followed by most of the telecom industry developers.

To better address the last issue, the *Media Resource Control Protocol version 2* (MRCP v2) has been recently proposed [11]. Basically, a system using MRCPv2 consists of a *client* that requires the generation and/or consumption of media streams and a *media resource server* that has the resources or “engines” to process these streams as input or generate them as output. Some of these media processing resources could be speech synthesis, automatic speech recognition, or speaker verification/identification engines. The protocol requirements dictate that the client should be capable of reaching a media processing server and setting up communication channels to the media resources, and of sending and receiving control messages and media streams to/from the server. The *Session Initiating Protocol* (SIP) [9] is the signaling protocol that meets these requirements; in fact, we must emphasize that SIP is fast becoming a global standard for any kind of media signaling. MRCPv2 leverages these capabilities by building upon SIP and the *Session Description Protocol* (SDP). MRCPv2 uses SIP to setup and tear down media and control sessions with the server, and SDP to describe the parameters of the media sessions associated with the dialogue between the client and server [5]. In conclusion, MRCPv2 uses SIP and SDP to create the client/server dialogue and set up the media channels to the server.

The proposed application architecture is illustrated in Fig. 4. The main functional entities are the following:

1) *SIP overlay network*, consisting of three main units:
   a. The *softswitch*, which implements:
      - SIP signaling for registration, call setup and DTMF (*Dual-Tone Multi Frequency*): the SIP Registrar and SIP AS (*Application Server*)
      - RTP (*Real Time Protocol*) stack for voice payload
      - TTS client with MRCPv2 protocol stack
      - HTTP-based RSS module for fetching the web streams

   b. TTS server for Romanian language, 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) *2.5G/3G mobile network*: any of the existing GPRS (*General Packet Radio Service*) networks could be used (for demonstration purposes only) to support the communication between the mobile phone and the SIP AS from the SIP overlay network.
3) *Mobile phone*: currently we are using Windows Mobile based *smartphones*, with SSL VPN and VoIP (*Voice over IP*) clients.
4) *Internet*: place where RSS feeds lives.

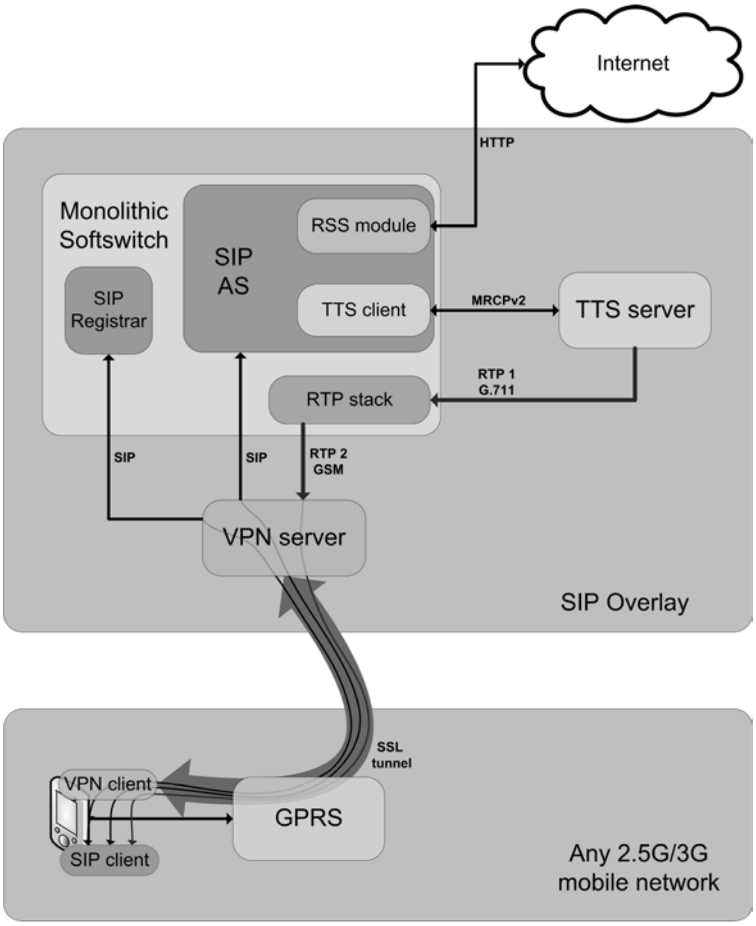

Fig. 4 – An RSS reader platform based on TTS synthesis in Romanian.

The RSS reader service consists of the following steps:

- The end-user initiate an SSL tunnel using the WM6.1 VPN client previously installed on the smartphone.
- After the connection is established, the smarphone will have an IP belonging to the SIP overlay network, therefore routing to the SIP Registrar and SIP AS will be possible.
- The end-user starts the SIP client previously installed on the smartphone. 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.
- By dialing a predefined extension (e.g. 9333), the SIP client will send an INVITE to the SIP AS that will trigger the RSS reading process:
  - the RSS module fetch the latest news for the preconfigured feeds;
  - SIP AS initiate a MRCP dialogue with TTS server, extracting the relevant text from the feeds;
  - the TTS server opens the first RTP leg towards the softswitch RTP port (the codec used for this stream is G.711 PCM A-law 8 kHz);
  - the SIP AS opens the second RTP leg towards the smartphone, using GSM codec; then it will transcode and relay the stream from the first RTP leg.
- Anytime during the call it is possible to choose the feed, skip to the next news, or revert to the previous one by sending DTMF codes: 1 to 9 to change the feeds from the main menu, 6 and 4 to skip or replay the news in the feed menu. The DTMF codes are conveyed by the SIP INFO method from the smartphone to the SIP AS.

This RRS reader platform permits concurrent client accesses, as all the network functional entities are multi-threaded. Furthermore, by implementing MRCP protocol it is possible to choose from a variety of SIP softswitches as long as a MRCP client is available.

4. CONCLUSIONS

The paper discussed first the major role played by speech synthesis technology in what is already called "Human-Device Interaction", as part of voice-based multimedia user experience. From this perspective, we pointed out the benefits and also the challenges of network-based speech synthesis solutions for mobile devices.

We then described our advances in developing a high-quality TTS synthesis system in Romanian language, highlighting the philosophy and implementation of the new synthesis technique based on the Harmonic plus Noise Model, used for acoustic segment concatenation and prosody modification. Informal perceptual listening tests performed so far show a very good speech quality of the synthesized speech and confirm the validity of the complete speech generation strategy.

As one of the most dynamic and comprehensive text news source is represented by RSS feeds, we developed an RSS reader multi-thread platform using our TTS system in Romanian, involving 2.5G/3G WM6.1 terminals to consume the service, SIP/RTP overlay network to convey signaling and payload, and MRCP enabled TTS server as the heart of the service, all of these running on off-the-shelf hardware.

The RSS reader platform was used as well to test end-user subjective reaction to a specific test-to-speech conversion for Romanian language. Even if its main targets were to verify the network-based concepts and also to test our TTS system in a completely new environment, the platform is in present fully functional.

ACKNOWLEDGEMENT

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

REFERENCES

1. BAGEIN, M., PIETQUIN, O., RIS, C., WILFART, G., *Enabling Speech Based Access to Information Management Systems over Wireless Network*. Proceedings of the 3rd workshop on Applications and Services in Wireless Networks, Berne, 2003.
2. BURILEANU, D., *Basic Research and Implementation Decisions for a Text-to-Speech Synthesis System in Romanian*, International Journal of Speech Technology, Kluwer Academic Publishers, Dordrecht, **5**, *3*, pp. 211–225, 2002.
3. BURILEANU, D., NEGRESCU, C., *Prosody Modeling for an Embedded TTS System Implementation*. Proceedings of the 14th European Signal Processing Conference EUSIPCO 2006, Florence, pp. 715–718, 2006.
4. BURILEANU, D., *Spoken Language Interfaces for Embedded Applications*, Human Factors and Voice Interactive Systems (D. Gardner-Bonneau and H. Blanchard – Eds.), 2nd Edition, Springer, New York, 2008, pp. 135–161.
5. BURKE, D., *Speech Processing for IP Networks: Media Resource Control Protocol (MRCP)*, John Wiley & Sons Inc., New Jersey, 2007.
6. McAULAY, R.J., QUATIERI, T.F., *Speech Analysis/Synthesis Based on a Sinusoidal Representation,* IEEE Transactions on Acoustics, Speech and Signal Processing, **34**, pp. 744–754, 1986.
7. NEGRESCU, C., CIOBANU, A., BURILEANU D., STANOMIR D., *An Improved Hybrid Time-Frequency Algorithm for Time-Scale Modification of Speech/Audio Signals.* Advances in Spoken Language Technology, The Publishing House of the Romanian Academy, Bucharest, 2007, pp. 147–162.
8. 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.
9. ROSENBERG, J., SCHULZRINNE, H., CAMARILLO, G., JOHNSTON, A., PETERSON, J., SPARKS, R., HANDLEY, M., SCHOOLER, E., *SIP: Session Initiation Protocol*, RFC 3261, June 2002.
10. 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, 2006, pp. 213–216.
11. SHANMUGHAM, S., BURNETT, D., *Media Resource Control Protocol Version 2 (MRCPv2)*, Draft-ietf-speechsc-mrcpv2-12, March 2007.
12. STYLIANOU, Y., *Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis,* IEEE Transactions on Acoustics, Speech and Signal Processing, **9**, pp. 21–29, 2001.
13. TATHAM, M., MORTON, K., *Developments in Speech Synthesis*, John Wiley & Sons Ltd, Chichester-West Sussex, 2005.
14. UNGUREAN, C., BURILEANU, D., POPESCU, V., NEGRESCU, C., DERVIS, A., *Automatic Diacritic Restoration for a TTS-based E-mail Reader Application*, UPB Scientific Bulletin, Series C, **70**, *4*, Politehnica Press, Bucharest, pp. 3–12, 2008.

Received September 11, 2009