Journal of Balkan Libraries Union

ISSN 2148-077X

http://www.balkanlibraries.org/journal
http://bkdb.trakya.edu.tr

Design Considerations of an Interactive Robotic Agent for Public Libraries

Stelios M. Potirakis[a], Todor Ganchev[b], Gürkan Tuna[c,*], Nicolas-Alexander Tatlas[a] and Recep Zogo[d]

[a]Department of Electronics Engineering, Technological Education Institute of Piraeus, Aigaleo, Greece
[b]Department of Electronics, Technical University Varna, Bulgaria
[c]Department of Computer Programming, Edirne Vocational School of Technical Sciences, Trakya University, Edirne, Turkey
[d]Directorate of Library and Documentation, Trakya University, Edirne, Turkey
** Corresponding author. Tel.: +90-284-224-0283; fax: +90-284-224-0287; e-mail: gurkantuna@trakya.edu.tr*

Research Article

ARTICLE INFORMATION

Article history:
Received 17 June 2013
Received in revised form 29 August 2013
Accepted 9 September 2013
Available online 16 September 2013

Journal of Balkan Libraries Union
Vol. 1, No. 1, pp.1-6, 2013.

ABSTRACT

The role of public libraries has long been recognized, rendering them a timeless offered form of public service. Users of a public library should have easy access to catalogs and full text of printed and electronic versions of books, magazines, and periodicals, as well as to multimedia databases. Every day, most public libraries are in service for several hours, thus computer-based library service applications provide a valuable service supplement. In this study, a robotic agent which guides users in libraries is proposed. The robotic agent is equipped with sound acquisition and reproduction chains and is capable of understanding some specific commands and guiding the users. The agent is currently able to understand commands and respond in English. Therefore, it may be useful for public libraries visited or remotely used by foreign, English speaking, users. Future work consists of the implementation of language packages for Turkish and the evaluation of field tests that will be held at the library and documentation center of Trakya University, Edirne, Turkey.

Keywords: Interactive robotic agent, public libraries, automatic speech recognition, speech-to-text.

I. Introduction

Public libraries are libraries that are accessible by the general public and are operated by civil servants. They exist in many countries across the world and are often considered an essential part of having an educated and literate population (Rubin, 2010). In public libraries, users are allowed to take books and other materials off temporarily. In order to improve service quality of libraries, users should be guided well by the civil servants and computer-based applications (Hsieh, Chang and Lu, 2000).

Library service research has recently focus on the exploitation of different forms of concurrent technological facilities in order to improve the services offered to typical, distant, as well as to physically or visually impair users (Breitbach and Prieto, 2012), (Cho, Kim, and Cha, 2012), (Evans and Reichenbach, 2012), (Fassbender and Mamtora, 2013), (Hill, 2013), (Jonnalagadda, 2012), (Mairn, 2012), (Mallon, 2012), (Mikawa, Morimoto and Tanaka, 2010), (Singh and Moirangthem, 2010). These technological facilities include smart-phones, internet access, robots and voice interaction with the user.

Human-Robot Interaction (HRI) is an extensive research area covering several topics towards direct interaction with robots through gesture and speech (Kiesler and Hinds, 2004). It plays a key role in the development of socially-intelligent robots by enabling them to understand the requirements of their users (Yanco and Drury, 2004). It is used in several robotic applications such as teleoperation, robot hosting, and service robots (Werner, Oberzaucher and Werner, 2012). HRI-based applications are realized in three main steps: 1- awareness and acquisition, 2- interpretation and 3- execution.

In this study, a robotic agent for serving users is proposed. By recognizing speech, the agent determines tasks to be executed and is supposed to aid civil servants in guiding users. Supported by automatic speech recognition (ASR) and text to speech (TTS) synthesis systems, the agent can understand several commands and ask clarifying questions in case of ambiguity.

The rest of the paper is organized as follows. Section II presents the details of the robotic agent designed to operate in libraries. Section III gives the details of the speech interaction subsystems of the proposed system. Finally, the paper is concluded in Section IV.

II. System Architecture

The architecture of the proposed system is shown in Fig. 1. The system translates the speech of the users into text and then carries out tasks given to. On the other hand, any feedback given by the system to the user is first generated in text form and then transformed to synthesized speech.

Each component of the proposed system is briefly described below.

- Audio stream: Utterances from users.
- Speech to text: Translates the utterances into text using an ASR subsystem.
- Dialog management: Manages dialog and resolves any ambiguities through clarifying questions.
- Task Management – Determines tasks to be executed.
- Audio feedback: Generates any feedback messages from the robotic agent in text form and then synthesizes speech using a TTS subsystem.

In applications, such as library robot agent the task completion rate criterion is adopted. A task is considered completed when the recognition engine provides a correct transcription of the speaker's intention. The speech understanding component parses every command phrase and translates it to the corresponding concept, which feeds the dialog generation module. The dialog generation flow operates on a conceptual level, which facilitates the handling of the variety of library user requests.

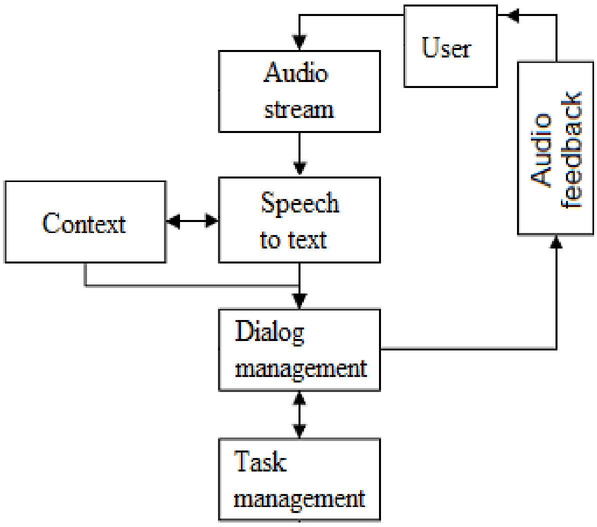


Fig. 1. An overview of the system architecture

III. Speech Interaction Subsystems

Recently, speech recognition and synthesis have become key techniques in intelligent HRI. Speech interaction, which is the fundamental way of communication among humans, is not just an easy way to interact with the machine; especially in the case of visually impaired people and in general people with physically disabilities that do not permit them to use mouse or keyboard as input devices this is probably the most effective way to communicate with the machine (Sharma and Wasson, 2012).

Command and Control Speech Recognition

In applications with speech-enabled HRI it is extremely important to guarantee reliable and error-free communication. This is the reason why the speech interface in robot agents is typically implemented as a *command and control speech recognition*. In such cases, the user interface makes use of a relatively small set of carefully chosen commands, which facilitates the use of strict grammars and boosts the speech recognition accuracy. In fact, in command and control speech recognition in indoor environments, it is typical to achieve nearly perfect speech recognition accuracy for vocabulary size of fifty, sixty or even larger number of command phrases (word combinations). Due to the use of strict grammars even when the word recognition rate is not perfect, the *task completion rate* is close to 100 %. These considerations motivated our choice of command and control speech recognition for the speech interface of the library robot.

The robotic agent ASR subsystem consists of two modules: (a) the word spotter and (b) the command and control speech recognition module. The word spotter is constantly seeking a specific word (e.g. *Agent*), which indicates that what follows is a command to the robotic agent. Detecting the word *Agent* activates the command and control speech recognition module. Thus, instructions which are not preceded by the word *Agent* are ignored. This rule follows the good practices of command and control speech interfaces, a safety measure so that the robot is not activated sporadically during discussion between people.

For the development of the word spotter and the command and control speech recognition module we made use of the Microsoft Speech API 5.3, which provides a high-level interface to a speech recognition engine (MSDN Microsoft, 2013a). Specifically, we set the speech recognizer in *shared recognition engine* mode, so that the word spotter and the command and control speech recognition use the same instance of the engine but with different grammars. This allows for more efficient use of the hardware resources, as the word spotter and the command and control speech recognition components do not work simultaneously, but are cascaded. The word spotter enables the command and control speech recognition module only if the robot name *Agent* is included in the command phrase.

For the command and control speech recognition module, we established a simple grammar, which can be summarized as:

Agent → (please) → **{verb}** → (a | an | the) → **{item}**

where the set of allowed actions is **{verb}** = find | bring | go | show | … and the set of allowed objects is **{item}**= article | journal | book |... Any word which does not fit this grammar is ignored. For safety reasons, when the confidence level of the speech recognizer is low the robotic agent asks (through the text-to-speech module) the command to be repeated.

The Text to Speech Subsystem

Speech synthesis is the artificial production of human speech, as natural and intelligible as possible. In the special case that a speech synthesis subsystem accepts text as input and narrates it to speech, this subsystem is referred to as a text-to-speech synthesizer or TTS for short. TTS is the function that is complementary to ASR; a robot having both ASR and TTS subsystems can provide a full speech-based HRI (Holmes and Holmes, 2001).

Fig. 2 provides the block diagram of a generic TTS module. In a higher level this can be divided to two parts (Van Santen et al., 1997), the "front-end" providing a translation of the input text to convert the input text to sound intermediate linguistic representation, and the "back-end" producing the synthesized speech waveform based on the obtained linguistic representation. The front-end first processes the provided character strings, e.g., identifies numbers, abbreviations, resolves different spellings for a word, etc., to identify the spoken words corresponding to the provided text. Then, then assigns phonetic transcriptions to each word, and finally, the linguistic analysis follows which reveals linguistic parameters, like the phonetic values of the parts of each word, divides the text into prosodic units, like phrases, clauses, and sentences and finally defines prosody characteristics, i.e., duration, intonation (involving fundamental frequency F_0 , i.e. pitch, information) and intensity patterns of speech for the sequence of syllables, words and phrases (Van Santen *et al.*, 1997), (Reddy and Rao, 2013).

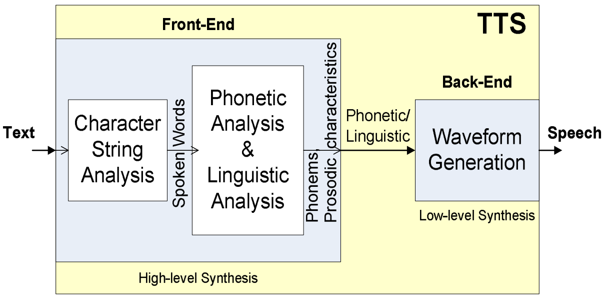


Fig. 2. Block diagram of a generic TTS system

The quality of TTS subsystem is determined by the similarity of the synthesized voice to human voice (naturalness), and the extent to which the synthesized voice is understood (intelligibility). Provided that a successful front-end has been designed for the language of interest, the quality of the TTS subsystem is determined by the success of the back-end, i.e., the method of waveform synthesis. Therefore, it has been a subject of intense research during last decades.

The early TTS back-ends (until the 1990's) were using articulatory or sinewave formant synthesis (e.g. (Rubin, 1982), (Remez *at al.*, 1981)). However, due to the low naturalness of the synthesized voice, scientists turned to concatenative speech synthesis which is based on the concatenation of segments of recorded speech. These segments may vary from sentences and phrases to diphones and phones (Olive, 1997). This kind of synthesis generally provides better naturalness and intelligibility, however calls for a large recorded speech database (corpus-based) and greatly depends on the segment type (phonetic unit) employed. Although, concatenative synthesis is still active and continuously improving, a more efficient (at least in the required database point of view) formant synthesis method emerged in mid-1990's, the corpus-based statistical parametric synthesis. This method became quite popular during the 2000's and is usually found as the HMM-based speech synthesis (Dutoit, 1997), (Ling, 2012). The main advantage of this approach is its flexibility in changing speaker identities, emotions, and speaking styles (Tokuda *et al.*, 2013). It is comprised of two parts: (i) the corpus-based HMM training part, which is very similar to the corresponding HMM ASR training, and (ii) the synthesis part that takes into account the training part and the front-end linguistic parameters to synthesize speech corresponding to the input text.

A number of different TTS products are already commercially available for different languages, while others other TTS engines are free-accessible, some of them being multilingual providing a platform for the development of TTS for different languages. Of course, all operating systems, as well as internet applications provide speech synthesis features.

In this context, Microsoft Speech Application Programming Interface (SAPI) is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. It is possible for a 3rd-party company to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with SAPI (Chungurski, Arsenovski and Gjorgjevikj, 2012), (Microsoft Research, 2013), (MSDN Microsoft, 2013a). The System.Speech.Synthesis namespace can be used to access the SAPI synthesizer engine to render text into speech using an installed voice, such as Microsoft Anna (Sharma and Wasson, 2012), (MSDN Microsoft, 2013b).

For the purposes of this research we have used a Matlab-based TTS engine that employs SAPI. An example of the obtained speech from an indicative response message that the proposed system should produce after the selection of the requested book by a person that uses the library services is presented in Figs 3 and 4. For the blue parts of the synthesized waveform (Fig. 3, middle panel) that have been identified as possibly voiced (Fig. 3, top panel), the corresponding blue parts of the bottom panel of Fig. 3 show the estimated pitch (fundamental frequency, F_0) (Gonzalez and Brookes, 2011), while Fig. 4 shows the spectrogram of the synthesized voice.

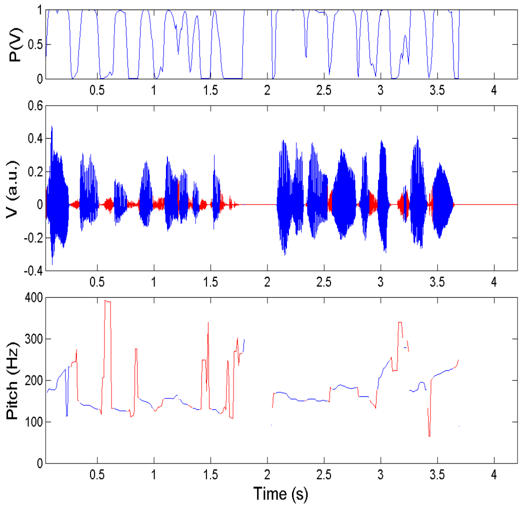


Fig. 3. The voice produced by the proposed system corresponding to the response text message "Your selection has been registered; we will bring you the book shortly". Probability of being voiced for each specific part of the synthesized waveform (top panel), the produced voice waveform (middle panel) and pitch analysis of the produced waveform (bottom panel).

The design of a Turkish language TTS has already been a research topic (e.g. (Bozkurt and Dutoit, 2001), (Bicil, 2010), (Tekindal and Arik, 2012), (Uslu, Ilk and Yilmaz, 2013), (Uslu *et al.*, 2013), (Yurtay *et al.*, 2012)), while a relevant (logotom) database has already been designed by The Scientific and Technological Research Council of Turkey (TÜBİTAK), the National Research Institute of Electronics and Cryptology (UEKAE) and the Izmir Institute of Technology (İYTE) (Bicil, 2010), (Tekindal and Arik, 2012), (Uslu, Ilk and Yilmaz, 2013), (Uslu *et al.*, 2013), (Yurtay *et al.*, 2012). In the next steps of our research we also intend to contribute to this direction.

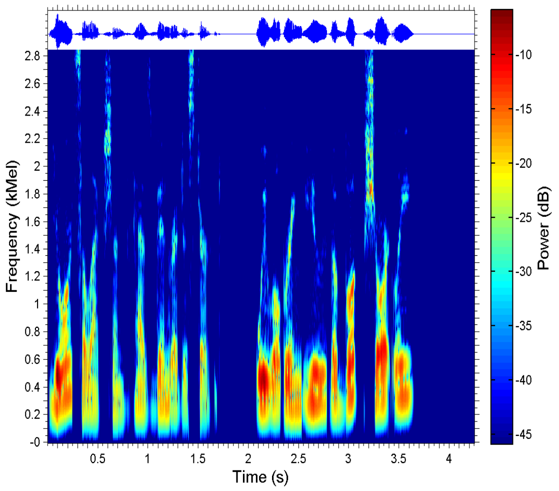


Fig. 4. The voice produced by the proposed system corresponding to the response text message "Your selection has been registered; we will bring you the book shortly". The produced voice waveform (top panel) and spectrogram analysis of the produced waveform (bottom panel).

IV. Conclusion

Generally, public libraries are in service for several hours. Even if it seldom happens, foreign users may want to use the public libraries. In this situation, an English-speaking public servant is needed to help them. In this study, a robotic agent is proposed to deal with this situation.

The robotic agent consists of an automatic speech recognition system and a text to speech subsystem. A Matlab-based text to speech engine employing Microsoft Speech Application Programming Interface has been used to test the applicability of the proposed robotic agent.

Future work of this study consists of the implementation of Turkish language packages for field tests that will be held at the library and documentation center of Trakya University, Edirne, Turkey.

Acknowledgements

An earlier version of this paper has been accepted for publication in the proceedings of the 7th International Advanced Technologies Symposium (IATS'13).

The authors wish to thank the anonymous reviewers for their valuable suggestions.

References

Rubin, R. E. (2010) *Foundations of Library and Information Science (3rd ed)*, New York, Neal-Schuman Publishers.

Hsieh, P.–N., Chang, P.–L. and Lu, K.–H. (2000) 'Quality Management Approaches in Libraries and Information Sciences', *Libri*, 50, 191-201.

Breitbach, W. and Prieto, A. G. (2012) 'Text reference via Google Voice: a pilot study', *Library Review*, 61(3), 188-198, DOI: 10.1108/00242531211259319.

Cho, H.-Y., Kim, B.-I., and Cha, S.-J. (2012) 'A Study on the Improvement in Statistical Indicators of Libraries for the Disabled', *Journal of the Korean Society for Library and Information Science*, 46(1), 141-162, DOI: 10.4275/KSLIS.2012.46.1.141.

Evans, D. A. and Reichenbach, J. (2012) 'Need for Automatically Generated Narration', *Proc. of CIKM 2012* (Conference on Information and Knowledge Management), 21-24, DOI: 10.1145/2390116.2390130.

Fassbender, E. and Mamtora, J. (2013) 'A Workflow for Managing Information for Research Using the iPad, Sente and Dragon Dictate: A Collaboration Between an Academic and a Research Librarian', The Australian Library Journal, 62(1), 53-60, DOI: 10.1080/00049670.2013.768520.

Hill, H. (2013) 'Disability and Accessibility in the Library and Information Science Literature: A Content Analysis', *Library & Information Science Research*, 35(2), 137-142, DOI: 10.1016/j.lisr.2012.11.002.

Jonnalagadda, S. (2012) *Android Application for Library Resource Access*, Master's Thesis, San Diego State University.

Mairn, C. (2012) 'Three Things You Can Do Today to Get Your Library Ready for the Mobile Experience', The Reference Librarian, 53, 263-269, DOI: 10.1080/02763877.2012.678245.

Mallon, M. (2012) 'The New Distance Learners: Providing Customized Online Research Assistance to Urban Students on the Go', Urban Library Journal, 18(1), 4.

Mikawa, M., Morimoto, Y. and Tanaka, K. (2010) 'Guidance method using laser pointer and gestures for librarian robot', *Proc. of IEEE RO-MAN 2010*, 373-378, DOI: 10.1109/ROMAN.2010.5598714.

Singh, K.P. and Moirangthem, E. (2010) 'Are Indian Libraries VIP-Friendly? Information Use and Information Seeking Behaviour of Visually Impaired People in Delhi Libraries', *Library Philosophy and Practice*, 2010, 374.

Kiesler, S. and Hinds, P. (2004) 'Introduction to this Special Issue on Human-Robot Interaction', *Human Computer Interaction*, 19(1), 1-8.

Yanco, H. and Drury, J. (2004) 'Classifying Human-Robot Interaction: An Updated Taxonomy', *Proc. of the IEEE SMC 2004 International Conference on Systems, Man and Cybernetics*, 2841-2846.

Werner, K., Oberzaucher, J. and Werner, F. (2012) 'Evaluation of Human Robot Interaction Factors of a Socially Assistive Robot Together with Older People', *Proc. of the 2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, 455-460.

Sharma, F.R. and Wasson, S.G. (2012) 'Speech Recognition and Synthesis Tool: Assistive Technology for Physically Disabled Persons', *Int. J. Comp. Sc. Telecom.*, 3(4), 86-91.

MSDN Microsoft (2013), Microsoft Speech API (SAPI) 5.3, Available at: http://msdn.microsoft.com/en-us/library/ms723627(v=VS.85).aspx [Accessed 11 June 2013].

Holmes, J. and Holmes, W. (2001) *Speech Synthesis and Recognition (2nd ed.)*, CRC Press.

Van Santen, J.P.H., Sproat, R., W., Olive, J.P. and Hirschberg, J. (1997) *Progress in Speech Synthesis*, New York, Springer.

Reddy, R. and Rao, K.S. (2013) 'Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis', *Computer Speech and Language*, 27, 1105–1126.

Rubin, P.E. (1982) 'Sinewave synthesis', Internal memorandum, New Haven, Haskins Laboratories.

Remez, R., Rubin, P., Pisoni, D. and Carrell, T. (1981) 'Speech perception without traditional speech cues', *Science*, 212(4497), 947–949. DOI:10.1126/science.7233191.

Olive, J.P. (1997) 'Concatenative Syllables', in *Progress in Speech Synthesis*, 261-262, New York, Springer.

Dutoit, T. (1997) *An Introduction to Text-to-Speech Synthesis*, Dordrecht, The Netherlands, Kluwer.

Ling, Z.-H. Microsoft Research (2012) 'HMM-based Speech Synthesis: Fundamentals and Its Recent Advances', Available at: http://research.microsoft.com/apps/video/dl.aspx?id=174749 [Accessed 21 July 2013]

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K. (2013) 'Speech Synthesis Based on Hidden Markov Models', *Proc. IEEE*, 101(5), 1234-1252.

Chungurski, S., Arsenovski, S. and Gjorgjevikj, D. (2012) 'Development overview of TTS-MK speech synthesizer for Macedonian language, and its application', *Proc. of ICT Innovations 2012*, 599-604.

Microsoft Research (2013) SAPI: Speech Application Programming Interface Development Toolkit, Available at: http://research.microsoft.com/en-us/projects/sapi/default.aspx [Accessed 11 June 2013]

MSDN Microsoft (2013) System.Speech.Synthesis Namespace, Available at: http://msdn.microsoft.com/en-us/library/system.speech.synthesis.aspx [Accessed 11 June 2013]

Gonzalez, S. and Brookes, M. (2011) 'A pitch estimation filter robust to high levels of noise (PEFAC)', *Proc EUSIPCO*.

Bozkurt B. and Dutoit, T. (2001) 'Implementation of Two Diphone-Based Synthesizers for Turkish', *Proc. Quatriemes Rencontres Jeunes Chercheurs en Parole*, 38-41.

Bicil, Y. (2010) *Turkish text-to-speech synthesis*, Master's Thesis, Sakarya University.

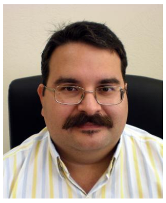
Tekindal, B. and Arik, G. (2012) 'Görme Engelliler için Türkçe Metinden Konuşma Sentezleme Yazılımı Geliştirilmesi ("Development of Speech Synthesis Software From Turkish Text for the Visually Impaired")', *BİLİŞİM TEKNOLOJİLERİ DERGİSİ*, 5, 9-18.

Uslu, I.B., Ilk, H.G. and Yilmaz, A.E. (2013) 'A Rule Based Prosody Model for Turkish Text-To-Speech Synthesis', *Tehnički vjesnik*, 20(2), 217-223.

Uslu, B., Demir, N., Ilk, H.G. and Yılmaz, A.E. (2013) ‘Bilgisayar Bir Metni Vurgulu Okuyabilir mi? (“Can Computers Read a Text with Stress?”)’, *Bilig*, 65,165-176.

Yurtay, N., Çelebi, S., Gunduz, B.A. and Bicil, Y. (2013) ‘A Mobile Product Recognition System for Visually Impaired People with IPhone 4’, *AWERProcedia Information Technology & Computer Science*, 3, 204-211.

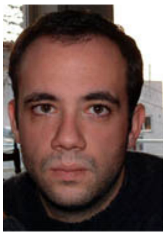
Authors’ Biographies



Stelios M. Potirakis currently serves as an Assistant Professor at the Department of Electronics of the Technological Education Institute of Piraeus. He received his B.S. degree in Physics with First Class Honors (1993), an M.Sc. degree in Electronics and Communications (1995), and a Ph.D. in Physics (2002), all from University of Athens. Dr. Potirakis has authored several papers in international conference proceedings and refereed journals and has been serving as a reviewer for international journals and conferences. His current research interests include analog electronics, analog circuit analysis, systems’ modeling, complex systems’ signal analysis, electroacoustics, applied acoustics, acoustic ecology, distributed sensor networks for environmental monitoring, electronic measurements, characterization of smart/ multi-function materials, and educational technologies. E-mail: spoti@teipir.gr

Todor Ganchev received the Diploma Engineer degree in electrical engineering from the Technical University Varna, Bulgaria, in 1993 and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2005. From February 1994 to August 2000, he consequently held engineering, research, and teaching staff positions at the Technical University Varna. Between September 2000 and September 2012, he has been a Researcher at the Wire Communications Laboratory, University of Patras, Greece. Since October 2012 he is a Faculty member of the Department of Electronics at the Technical University Varna, Bulgaria. Todor Ganchev authored/co-authored over 120 scientific publications in the areas of bioacoustics, signal processing, speech and audio processing, pattern recognition, and signal processing applications. Todor Ganchev is a Member of the EURASIP and a Senior Member of the IEEE. E-mail: tganchev@tu-varna.bg

Gürkan Tuna serves as an Asst. Prof. at Trakya University, Edirne, Turkey. He received his B.S. degree in computer engineering from Yildiz Technical University, Istanbul, Turkey, in 1998, and his M.S. degree in computer engineering from Trakya University, Edirne, Turkey, in 2009. He received his Ph.D. degree in electrical engineering from Yildiz Technical University, Istanbul, Turkey, in 2012. Tuna has authored several papers in international conference proceedings and refereed journals, and two book chapters. He has been serving as a reviewer for international journals and conferences. His current research interests include smart grid, ad hoc and sensor networks, robotic sensor networks, multisensor fusion, energy harvesting, and energy-aware routing.



Nicolas-Alexander Tatlas received an engineering degree in 2001 and a Ph.D. degree in 2006, both from the Department of Electrical and Computer Engineering, University of Patras, Greece. His research focused mainly on digital audio delivery over PANs and WLANs, digital audio amplification, and direct digital electroacoustic transduction. He was also involved in a number of international and national projects in digital audio technology. He has authored and presented more than 20 papers in scientific journals and international conferences. Dr. Tatlas is a member of the Technical Chamber of Greece, an associate member of the Audio Engineering Society, and a committee member of the AES Greek Section. Dr. Tatlas currently serves as a Lecturer at the Dept. of Electronics Engineering, Technological Education Institute of Piraeus, Egaleo, Greece. E-mail: ntatlas@teipir.gr

Recep Zogo received his master's degree in public administration from Trakya University, Edirne, Turkey, in 2009. He currently serves as the Head of Library and Documentation Directorate, Trakya University, Edirne, Turkey. E-mail: recepzogo@trakya.edu.tr