Computer Science and Information Systems 18(1):231–250 https://doi.org/10.2298/CSIS200424036S

Spoken Notifications in Smart Environments Using Croatian Language

Renato Šoić, Marin Vuković, Gordan Ježić

Faculty of Electrical Engineering and Computing,
10000 Zagreb, Croatia
{renato.soic, marin.vukovic, gordan.jezic}@fer.hr

Abstract. Speech technologies have advanced significantly in the last decade, mostly due to rise in available computing power combined with novel approaches to natural language processing. As a result, speech-enabled systems have become popular commercial products, successfully integrated with various environments. However, this can be stated for English and a few other "big" languages. From the perspective of a minority language, such as Croatian, there are many challenges ahead to achieve comparable results. In this paper, we propose a model for natural language generation and speech synthesis in a smart environment using Croatian language. The model is evaluated on 27 users to estimate the quality of user experience. The evaluation goal was to determine what users perceive to be more important – generated speech quality or grammatical correctness of the spoken content. It is shown that most users perceived grammatically correct spoken texts as being of the highest quality.

Keywords: natural language processing, smart environment, speech synthesis, natural language generation.

1. Introduction

Speech enabled computer systems have become a common presence in the last several years. Their domain has extended from personal devices and specialized services to all types of smart human-inhabited environments. As a result, there are many examples of employing speech interfaces in smart home systems, industrial facilities, smart vehicles, public services, etc. Regardless of the operational domain, spoken communication between human users and computer systems needs to address specific features related to the wider context of the target environment [1]. This includes events and conditions in the environment, current and future activities related to users, and the linguistic context when interaction is in process.

In this paper, we describe a model for natural language generation and speech synthesis in a distributed environment. The goal is to notify human users about conditions, changes, and events in the environment with naturally sounding spoken notifications. In order to do so, we integrate research efforts from two research domains – speech technologies and Internet of Things (IoT). The primary motivation is to provide support for Croatian language in a real smart IoT environment. Unfortunately, Croatian language is technologically seriously underdeveloped, with sparse resources

available [2]. There are available services related to speech technologies in Croatian language provided by Google, Microsoft, Amazon, and Apple. However, these solutions are commercial products and therefore not available for research purposes. Furthermore, the performance of these systems is often not of desirable quality for a native Croatian speaker. In comparison, in case of English language, there are numerous solutions available, providing experience at a very high level of quality. Therefore, this paper proposes the process and resources required to design a functional system. The proposed system can operate in complex IoT environments consisting of numerous sensors and devices, while interacting with human users using a text-to-speech subsystem. The presented concept can be employed in different scenarios from various domains, such as smart homes, Industry 4.0, public transportation systems, etc.

Development of a system providing speech recognition and speech synthesis capabilities requires a vast amount of work and large volumes of data. From the perspective of an under-resourced minority language such as Croatian, the required efforts are even greater. The first challenge is related to the required resources. For speech synthesis, a large audio corpus from a limited number of different speakers is essential. The recorded speech should be of high quality, and it is desirable that recorded speakers are professionals [3]. In case of speech recognition, a large audio corpus from many different speakers is recommended, with variable sound quality, so the system can learn to handle interferences in a real-world scenario. In addition to speech corpora, flawless corresponding textual transcripts are also required. There are a few Croatian corpora available for research purposes, but their volumes and quality are not sufficient for employing data-driven methods. They are extracted from specific domains (e.g., weather forecast, radio shows, etc.) [4] or contain spontaneous speech in different dialects [5]. Therefore, these corpora are not convenient for the intended task as they do not adequately cover the required variety of linguistic features. In comparison, there are open corpora for English [6][7] and projects providing support for many other languages [8].

The second challenge is related to language modeling and cognitive processes. Unlike English, Croatian language is morphologically very rich [9]. This introduces many challenges related to speech recognition, but also cognitive analysis and synthesis, i.e. natural language understanding and generation. In case of natural language understanding, it is very important to recognize user intent correctly, as it results with an action being performed by the system. Similarly, when constructing a notification for the human user, the message content must be clear and precise.

In case of a distributed smart environment, a system which enables human-computer interaction needs to process all the events from the environment in real-time and provide feedback when necessary. In this case, understanding both user's and environmental context is essential. User's context is constructed based on their presence, physical and virtual (online) activities, requests, etc. Environmental context is related to information from the current state of the given environment, such as temperature, humidity, light, noise, etc. Additionally, it should also include future predictions, such as weather forecast, public events and happenings as well as personal planned events.

With all the described problems and challenges considered, the conclusion is that building such a complex system requires very careful system decomposition. Therefore, we adopted a modular approach in which each system module represents one major building block while relying on a specific set of services to accomplish its tasks. The

proposed system consists of speech synthesis and a natural language generation subsystem and is organized as an orchestration of those subsystems and their related services required for contextual text to speech synthesis. To achieve high-quality context-based speech synthesis, all the proposed subsystems and related services need to be available and perform their functions. However, since the defined subsystems might be distributed across the IoT network with sometimes limited resources and potential connectivity issues, it is possible that one or more of the subsystems is unable to provide the required processing in real time. In this paper we present the ideal conditions with high-quality synthesized speech, but also examine the cases when one or more subsystems fails to perform the task. Finally, we compare the user perception of quality of synthesized speech recordings when one or more subsystems failed to provide the results to see whether it is possible to still have understandable speech in limited connectivity / processing environment, such as large and complex IoT systems.

The evaluation was performed as a survey with 27 participants who were presented with a set of generated spoken notifications. The presented notifications were generated using two different speech synthesis solutions for Croatian language developed in scope of our research. The first was a statistical parametric speech synthesis system trained using a small self-recorded dataset, while the other was a WaveGlow implementation trained on 15 hours of carefully selected audio from Croatian radio shows. Survey participants were expected to grade the notifications in terms of intelligibility, grammatical correctness, and overall quality. The goal was to determine what users perceive to be more important – the generated speech quality or the grammatical correctness of the spoken content.

The next section describes the current state in research and development of speech-enabled systems and their applications in various domains. The third section describes the proposed model for spoken notifications in a smart environment, concepts related to context-awareness and required subsystems with their architectures. In the fourth section we evaluate the proposed system operating in a smart home environment and discuss the results. Finally, a conclusion is given, with plans for future research and development.

2. Speech Technologies in Smart Environments

From the human point of view, interaction with complex systems which constitute smart environments tends to be complicated. The complexity is due to presence of many different devices performing various functions across the physical environment. These devices communicate with each other and can influence each other's actions. Regardless of the application domain, smart systems are mostly distributed, provide different communication channels and users need to conform to specific user interfaces. The most used, but also most simple and straightforward method of interaction is a traditional graphical user interface, where a human user interacts with the system by using a client application which provides insight into various system features.

However, smart environments are still primarily supporting humans and the ultimate goal should be to achieve the most intuitive way of system interaction with human users. Speech has been recognized as the most natural and efficient method of communication for humans [10]. This requires that the system is capable of receiving spoken requests

and providing spoken feedback to human users. These functionalities require speech recognition and speech synthesis capabilities, respectively. However, in the context of a smart system, speech recognition and speech synthesis components represent interaction interfaces, with no cognitive processing of users' requests and system feedback. Therefore, additional components are required, the ones which could enable translation between natural language and language used by the system (e.g. system events, commands, etc.). These are typically represented as the natural language understanding and natural language generation component.

Human-computer interfaces which enable spoken interaction between a human user and the computer system have been successfully applied in various domains, from smart home systems to industrial facilities. However, their capabilities are usually limited to recognition of specific commands and reproduction of predefined spoken notifications. The rise of Intelligent Personal Assistants (IPAs) improved the situation significantly. IPAs have evolved into systems capable of leading conversations with human users, understanding linguistic and semantic context. They have also provided options for integration with various devices and external systems, making the IPAs very flexible and extensible [11]. All distinguished IPA platforms enable their users to develop new functionalities for the devices, thus creating a constantly growing ecosystem of services, but also introducing possible security and privacy threats [12]. Privacy is a serious issue in all notable IPA systems, since they constantly record and analyze private conversations due to their dependency on cloud infrastructure [13]. Despite all the mentioned advantages, the IPA approach hasn't yet achieved full integration with IoT smart platforms, in sense that it is fully capable of processing all events in the given environment and understanding its context.

IoT systems are mostly focused on enabling machine-to-machine interactions between devices, while typically relying on traditional Graphical User Interface (GUI) for interaction with human users. However, for humans, spoken interaction is more intuitive than learning all the options of a GUI or pushing buttons on specific controller devices. This has been recognized, resulting with significant advancements in speech recognition and natural language understanding in domain of smart environments. Spoken interaction in smart home applications with customizable devices has been introduced, thus expanding the system capabilities [14]. Furthermore, in addition to identifying spoken words, emotion recognition has been introduced with purpose of detecting user's mood [15]. The given examples did not focus on speech synthesis which would enable the IoT system to present basic information in spoken format back to the user. An example of a speech enabled IoT system with a combination of cloud services for speech recognition and speech synthesis is described in [16].

In case of Croatian language, there are numerous challenges. In the domain of speech technologies, Croatian language is seriously underdeveloped. Therefore, it is required to first develop usable speech recognition and speech synthesis systems which could perform adequately and provide satisfactory results, then they can be integrated with a system operating in a real-world environment. Our goal is to provide a framework which could be efficiently deployed into different smart environments and would have the ability to support a wide array of applications. Additionally, our long-term focus is the application subsystem that would enable spoken interaction between human user and the distributed computer system.

Spoken interaction between a human user and a computer requires both speech recognition and speech synthesis capabilities, along with respective cognitive processing components. Additionally, it is essential to understand user's and environmental contexts, as they represent a critical role in interaction methods which could be employed.

In scope of this work, we present the model of the complete system, but focus on requirements related to spoken notifications, which are natural language generation and speech synthesis.

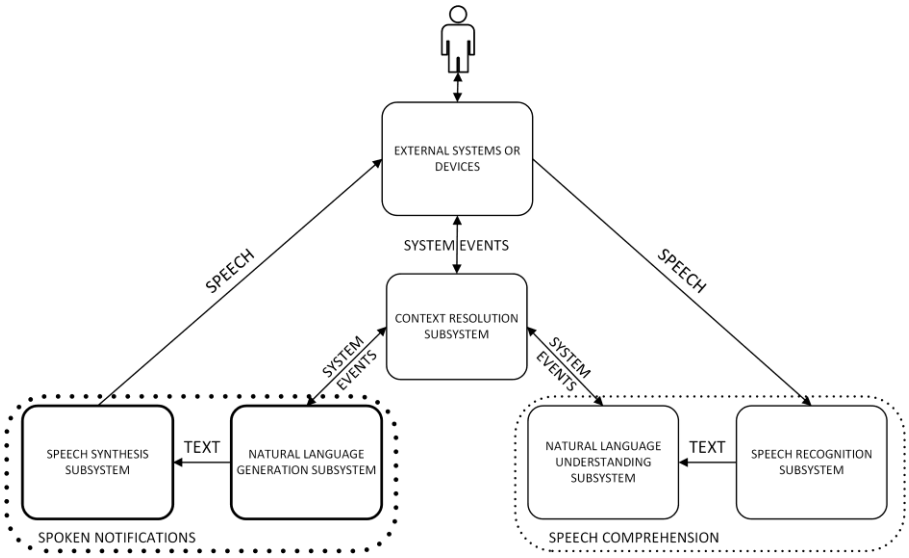

Fig. 1. Spoken interaction in smart environments

When examining related work and the presented requirements, a proposed model of the complete system should have the building blocks as presented in Fig.1. As depicted in the figure, all external events, including speech, are captured using various external devices and sensors. Depending on the captured event, an action or a notification (or both) may follow. This is decided by the context resolution subsystem, which calculates importance of the received event and creates a notification request if necessary. In case of a spoken command, the recorded audio is forwarded to speech recognition subsystem, which transforms it into the textual transcription of the spoken sequence of words. The natural language understanding (NLU) subsystem then translates the received text into a system event which can be processed by the context resolution subsystem. This subsystem performs several challenging assignments. First, it monitors all events in the system, thus building the environmental context. Second, it monitors user activities in order to derive user context. Based on available information, context resolution subsystem can make decisions related to the state of the environment (actions or notifications), regarding both the system and the user. From the perspective of the system, context resolution subsystem tends to keep the system (and the environment) in a stable state. From the users' perspective, it enables the user to interact with the smart environment in a convenient and meaningful way.

The context resolution subsystem can decide when it is important to inform the user about something and initiate interaction. Furthermore, it provides an additional layer of understanding regarding user's spoken requests. For instance, it will identify what the user's context was when a spoken command was issued and deduce to which device(s) it refers. Understanding of context is crucial for human-computer interaction in smart environments as it can enable the computer system to independently decide when, how, and why to initiate interaction with the user [17]. Context is comprised of information collected from the environment through use of various sensors. This information then needs to be processed and specific context variables need to be inferred from the available data. As a result, a context descriptor is constructed, and all system components can use the newly available information. Depending on specific conditions in the target environment, a specific method of interaction may be more appropriate and more efficient than others. For instance, in a loud environment (e.g., factory setting) there is no point in using speech interface for interaction with the computer system. However, employing the visual approach by using a traditional graphical user interface makes an efficient interaction method. Additionally, in a smart home environment it is desirable to provide a method of interaction which allows users to issue requests in the most intuitive way, by using speech. However, environmental conditions are dynamic in both examples, therefore the employed interaction method needs to be adjusted according to environmental context [18].

Understanding context is essential in case of human-computer spoken interaction and presence of multi-modal interaction methods. However, this will not be discussed in detail here, since the model proposed in this paper is focused on spoken notifications from computer system to a human user.

3. Proposed Model for Spoken Notifications

Spoken notifications represent an interaction channel which enables human users to receive information from the smart IoT system in form of computer-generated speech. This functionality relies exclusively on natural language generation and speech synthesis subsystems, which are described in the following subsections.

3.1. Speech Synthesis Subsystem

Enabling speech synthesis for a given language requires multiple components performing specific functions. These components transform text from its initial form to a form enriched with information required for creation of its spoken counterpart. Speech synthesis systems have evolved significantly over the last two decades, with development of more sophisticated techniques for language [19] and acoustic modeling, and new methods for generation of audio signal [20]. These improvements were mostly related to application of novel machine learning techniques applied on large text and speech corpora.

Regarding Croatian language, speech technologies are underdeveloped, there are no commercially available systems which are designed and implemented exclusively for

Croatian language. Some available systems are solutions adapted from similar languages, such as Newton Dictate [21], a speech recognition system designed primarily for Czech, and AlfaNum TTS [22], a speech synthesis system designed primarily for Serbian, but these solutions do not fully comply with Croatian prosody and morphology. In academic circles, there have been initiatives and projects dealing with speech recognition [23] and speech synthesis [24], but without usable results from the consumer's point of view.

In the group of Slavic languages, there have been successful research efforts which have resulted with commercially available systems for both speech recognition and speech synthesis. In case of Czech and Serbian language, novel methods have been successfully applied in the field and there are commercially available systems based on recent research achievements. Even though there are many similarities between Croatian language and Czech or Serbian language, there are still specific challenges which need to be addressed. For example, Czech language is accentuated in written form, while Croatian is not. This makes speech synthesis more complicated, because in case of Croatian language pronunciation accent needs to be modelled either by employing rule-based approach or learning from available corpora. Compared to Serbian language, Croatian has different accentuation rules, essential for producing high-quality synthesized voice. Additionally, in Croatian, foreign names and words in textual form are written in original language, which makes speech synthesis more challenging.

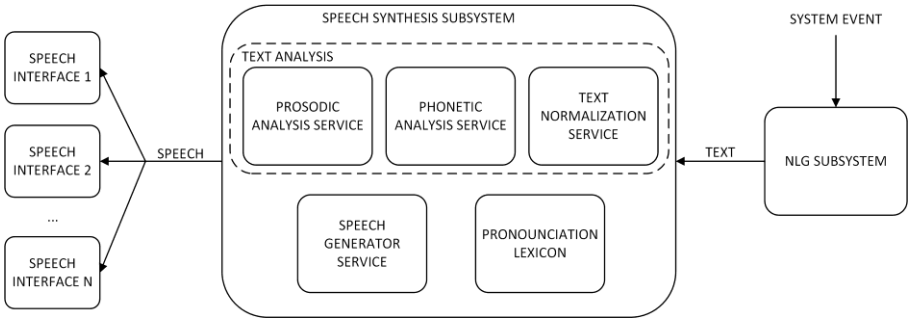

Fig. 2. Speech synthesis subsystem and related services

The proposed speech synthesis subsystem enables real-time generation of spoken notifications intended for the human user. Its purpose is to generate and reproduce spoken notifications using the convenient interaction channels. In this process, several specific tasks need to be performed. As shown in Fig. 2, these tasks can be represented with corresponding specialized services. The input text is typically received from the natural language generation subsystem, while the resulting synthesized speech can be reproduced on any available device (i.e., speech interface).

Text Analysis

Producing speech from the given text requires text analysis and transformation before the actual speech is generated. The text analysis process consists of three major steps:

text normalization, phonetic analysis and prosodic analysis. In the text normalization step, the entire text needs to be transformed into pronounceable units. This means that all non-standard words need to be identified and replaced with their corresponding fully extended counterparts. This includes numbers, time and date, acronyms, abbreviations and symbols (e.g., "7" becomes "seven", "ºC" becomes "degree Celsius", etc.). Text normalization is handled by employing a rule-based approach combined with a predefined dictionary. Text normalization in Croatian language is significantly more challenging than in English, because each fully extended form needs to be correct in terms of gender, case and number [25]. For this reason, once the initial normalization process is done, the result is validated by the spell-checker service which can detect and correct possible errors.

Completely normalized text is required for the next step – the phonetic analysis. In this step the text currently in form of sequence of graphemes transforms into a sequence of phonemes, which represent the content which will be pronounced. In case of native Croatian words, this transformation is straightforward because Croatian orthography is phonemic, i.e., phonetic representation of a word corresponds with its written form [26]. In English or German language this procedure is much more complicated because there are special rules required. However, Croatian language is very challenging in case of words coming from foreign languages. Regarding orthography, foreign words remain in their original form. This means that such words should be transformed to phonemes according to pronunciation rules in their original language. In this case, a pronunciation lexicon which contains phonetic transcriptions of foreign words is typically used, because foreign words are mostly names.

Prosodic analysis defines prosodic characteristics for the given content which will be synthesized. Prosodic features are duration, pitch, and intonation. Each feature is typically represented with its own model. Duration model is based on decision trees, which are used to determine phrase and sentence breaks. Pitch accent is typically applied when a word or a phrase in a sentence is emphasized. Intonation is related to variations in fundamental frequency in the given sentence. Pitch and intonation are modeled by employing machine learning algorithms on the available corpora.

Pronunciation Lexicon Service

Foreign words, names and phrases used in Croatian language retain their original orthography in the written form. However, when attempting to synthesize such content, the result is most often unrecognizable, as it does not conform to pronunciation rules present in Croatian language. For this reason, pronunciation lexicon service contains a collection of most common foreign words, names, and phrases and their phonetic transcriptions.

The lexicon approach proved itself to be the most viable option, since the alternative would be implementing language-specific pronunciation rules, which would introduce many additional challenges, such as language detection and support for language specific orthography.

Speech Generator Service

Selecting the most appropriate method for generating speech depends mostly on available resources and amount of work required. In our case, two approaches were considered – statistical parametric speech synthesis and generative neural network approach. In both cases, the speech generator relies on learning from the available data.

Statistical parametric approach is capable of building a generative model based on relatively small corpora. In the training phase, parametric representations of speech are extracted from a speech database and then modeled by using a set of generative models, typically Hidden Markov Models. During this process, linguistic units and corresponding parameters required for generating the waveform are identified. Additionally, a probabilistic model is built, whose purpose is to create parameters which were not present in the training dataset [27]. Statistical parametric speech generator was trained using a self-recorded speech corpus consisting of 654 sentences, summing up to two and a half hours of recorded speech. However, all sentences were recorded by a professional speaker. Regarding achieved results, this method provided satisfactory speech synthesis for Croatian language. Generated speech was intelligible, but not natural-sounding, with transitions between linguistic units often sounding abrupt.

Motivated by the significant improvements in quality of synthesized speech achieved by using deep neural networks for audio generation, we have decided to explore that approach. The generative approach using deep neural networks was first published in 2015 by DeepMind and has established itself as the most authentic speech synthesis method, based on English and Chinese (Mandarin) language [28]. The initial WaveNet implementation was not suitable for real-time applications because it was too computationally intensive. For instance, it would take approximately 50 seconds to generate 1 second of audio. The initial model and implementation were significantly improved in a very short amount of time, reducing the generation process to 50 milliseconds for 1 second of raw audio [29]. The generative deep neural network approach has been explored in various implementations, some of notable examples being Tacotron [30] and WaveGlow [31]. These approaches differ in underlying architectures of employed neural networks. Current results show that deep neural networks can successfully generate speech comparable to human speech in terms of intelligibility and naturalness. In case of Croatian language, there are still no notable research results or commercial systems which can produce speech of such high quality.

The main resource for developing a speech synthesis system based on generative neural networks is a large text and speech dataset (i.e., corpus) from which the system can be trained. In our case, the corpus consists of radio show recordings and corresponding transcripts obtained from Croatian Radio Television. The available corpora are radio shows in their original form, partially consisting from low-quality segments. Therefore, additional processing was required to create a dataset which could be used for training the neural network. For this purpose, additional services were developed, as displayed in Fig. 3.

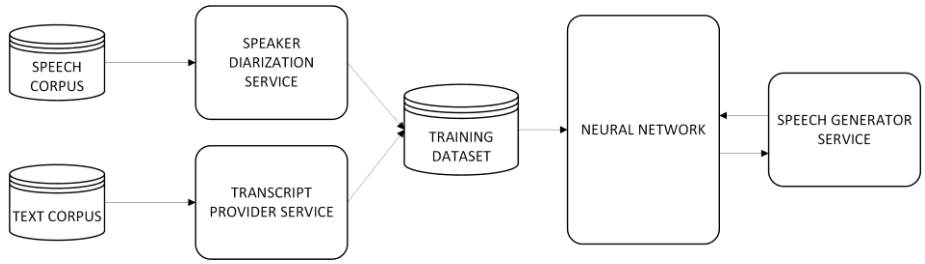

Fig. 3. Training the Speech Synthesis Subsystem

The speaker diarization service performs processing of audio files and extracts segments belonging to a designated speaker. This process is essential, because the available audio data consists of speech belonging to professional speakers (i.e., radio show hosts) and guests. Additionally, there were also segments of low-quality speech which originated from reporting by telephone. The results of speaker diarization process were audio segments organized according to speakers. The diarization is performed using adapted method provided by LIUM open source diarization toolkit [32].

The transcript provider service retrieves the text segments corresponding to the extracted audio segments. Not all available transcriptions had the timestamps which could be used for identifying the required segment, and there were also some erroneous entries. Therefore, the transcript provider service needs to combine transcripts with the original audio files in order to find correct text boundaries for the given extracted audio segments. This was achieved by analyzing spectrograms of the raw audio segments and recognizing word boundaries from the corresponding raw transcripts. The final quality of processed transcripts was satisfactory, but still lacking important content, such as some interpunction symbols like comma and colon. While they are not essential for training the generative model, these symbols are important for learning prosodic features (speed, intonation, pauses, etc.) which should be correctly reproduced from the input text.

The final step in preparation of the training dataset is removal of entries which contain foreign words and/or names. As previously explained, in Croatian language foreign names keep their original written form. Therefore, they introduce noise in the training of the neural network, as they do not conform to general rules valid in case of native words. This was performed by matching foreign names from a predefined list to the maximum extent possible and by further inspection of the prepared transcripts.

For the purpose of this research, we used approximately 15 hours of segmented audio from the radio shows with their corresponding transcripts. For speech generation, a WaveGlow implementation is used and the first results are promising, having in mind the relatively small training dataset. Evaluation of the resulting generated speech is presented in the next section.

3.2. Natural Language Generation Subsystem

Producing a spoken notification from an event which occurred in the environment requires a component capable of constructing a text sentence in human language, which

human users could fully comprehend. There are several possible approaches to generating text notifications adapted for human users. They differ in terms of which input data is used as the knowledge source and the methods which are employed to construct the output text. We are using template-based approach, which is quite straightforward and the most convenient when dealing with structured input data which typically occurs in a smart environment [33]. In this context, template-based approach means that system event data (i.e. well-known elements and attributes) is extracted and organized into a raw textual representation which will be improved with help of additional language processing services.

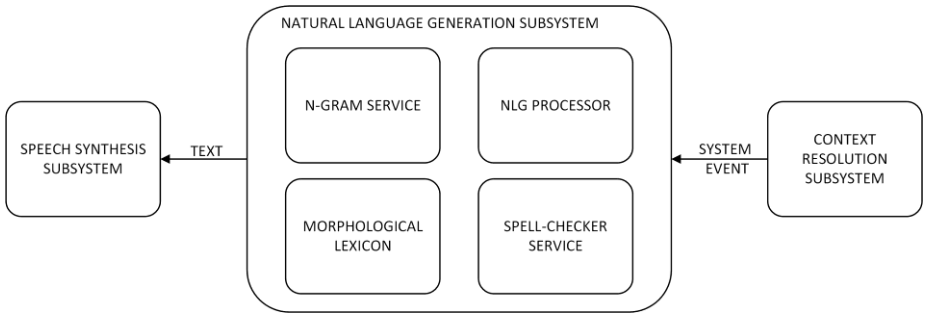

Fig. 4. The natural language generation subsystem and related services

The natural language generation subsystem consists of several services performing specific tasks, as displayed in Fig. 4. Based on system event received from the context resolution subsystem, the NLG subsystem produces a textual notification which is then passed on to speech synthesis subsystem.

NLG Processor Service

The entire process of generating textual notifications is orchestrated by the NLG processor service. This service performs the following steps required to produce a natural sentence from synthetic event: text structuring, lexicalization and linguistic realization.

In the first step the initial notification content is generated by employing the template-based approach. The structured system event is transformed to a raw textual notification. In its initial form, the notification is certainly not correct from the linguistic perspective. In the lexicalization step, the notification is extended by substituting certain tokens with appropriate phrases. Finally, in the linguistic realization step, the notification needs to be transformed into a well-formed sentence. This includes choosing the right morphological forms, verb tenses and conjugations, and punctuation marks.

While performing the previously described steps in generating the notification, NLG processor service relies on several specialized services. In the lexicalization step, n-gram service and semantic lexicon service are used to improve the raw notification with appropriate phrases. In the linguistic realization step, morphological lexicon service is used to rectify morphological forms, which is essential for Croatian language, while

spell-checker service performs final validation of the notification as a whole. These services are explained in more detail in the following section.

N-gram Service

Language models are essential components in both speech recognition and speech synthesis systems. Their purpose is to provide probability of a certain word occurring next in a given word sequence. In speech recognition, this is important for prediction of a next word which may be spoken by the human user, but also for correction of the erroneously recognized words, based on linguistic context. In speech synthesis, prediction of the following word enables better preparation of prosodic features of the generated speech.

Generally, n-grams are word sequences comprised of n words. An n-gram-based language model is constructed from a large text corpus, from which the word sequences are extracted. As a result, the model contains information about word transitions as they occur in real-life language usages. The n-gram model described in this paper was constructed by using an n-gram database obtained from a large general vocabulary text corpus collected by Hascheck, a Croatian spell-checker service. Currently, the n-gram service is based on 3-gram system, which contains word sequences comprised of three words. However, Hascheck contains n-gram collections of different lengths, with $2 <= n <= 7$ [34].

An n-gram system can be represented by a directed graph, where words are represented by nodes, while connections in the graph represent the existence of a linguistic connection, with the associated probability. N-gram based language models are very useful in case of morphologically rich languages, such as Croatian. They provide information about exact word forms in the given linguistic context. For this purpose, we developed a 3-gram model consisting of approximately 2 million n-grams which was implemented into neo4j graph database. The n-grams were extracted from Hascheck's collection. This provided us the ability to query the database in order to predict the most probable following word based on the given word or a sequence of two words. Additionally, this approach enabled us to determine the probability of words preceding the given word.

Morphological Lexicon Service

Morphology is a linguistic discipline in which the smallest meaningful linguistic units (morphemes) are analyzed, including their form and their transformation depending on linguistic context. Croatian is a morphologically rich language, which means that words typically occur in various forms when used in a sentence. For instance, nouns are determined by their gender, number and case; verbs by verb tense, aspect, mood, and voice.

Morphological lexicon enables identification of word's morphological characteristics, such as word types, tenses, cases, etc. Additionally, it can provide information about the base form for any given word. In case of speech comprehension, this allows for better resolution of understanding user's intent. In scope of spoken notifications, it enables

more correct and credible generated notifications, because they will sound more naturally when reproduced if the spoken words are in appropriate forms.

In domain of natural language processing, morphology is important in all related disciplines. In the process of translation from machine-generated events to natural language and vice-versa, understanding morphology plays the most important role regarding the quality of results. This makes the process of natural language understanding especially challenging, because minor errors in some word forms can cause potentially big differences in the meaning of a sentence.

Morphological lexicon was constructed using a segment of the morphological database available in Hascheck [34], summing up to approximately 700 000 entries. The implemented service provides the morphological descriptor for the given word, thus determining its exact morphological form. Additionally, it can provide the entire morphological tree for the given word.

4. Evaluation in a Smart Home Environment

The system described in previous sections was evaluated in a simulated smart home environment. The evaluation setup was designed to test how the users perceive spoken notifications of different levels of quality. For this purpose, several scenarios were examined with different proposed system components being involved. In this sense, we wanted to determine how certain components involved in the entire proposed process can affect the final result – synthesized spoken notifications.

The simulated smart-home environment consisted of three remote nodes equipped with standard sensors (temperature, humidity, microphone), interaction devices (i.e. speakers) and a central server which was collecting and processing system events. The entire process of generating textual and reproducing spoken notification starts with a specific system event received by the natural language subsystem. For example, when a temperature sensor reads a temperature above predefined threshold, it triggers an event that should result in a spoken notification to the user. We defined several templates for notifications according to the events, and the appropriate template for the examined event of temperature above threshold is:

`[`*`at`* `$timestamp, $source` *`in`* `$location` *`is`* `$value]` (in English)
`[`*`u`* `$timestamp, $source` *`u`* `$location` *`je`* `$value]` (in Croatian)

When dealing with Croatian and similar morphologically rich languages, the process of obtaining grammatically correct notification from such template becomes much more complex due to many grammatical rules. This is done by previously described services employed in the natural language generation process.

The timestamp value is transformed before it is inserted into the raw notification, as well as the Boolean type values, if any, while other values remain unchanged. A simple example notification in Croatian is given in Table 1. Each row represents a case where some or all of the NLG services are used, in order to illustrate the effect of each service on the resulting textual notification. The order of invoking the services is irrelevant

since each service tackles a different task regarding the raw notification. The labels (RAW, RAW+, COR) in the table are used for easier discussion of evaluation results.

Table 1. Evaluation - NLG related services and their impact on resulting notifications

Label	Service in use?			Resulting NLG notification
	N-gram	Morphological lexicon	Spell-checker	
RAW	NO	NO	NO	U 18:35 temperatura zraka u dnevna soba je 28 stupanj Celzijev.
RAW+	NO	YES	YES	U 18:35 temperatura zraka u dnevna soba je 28 **stupnjeva Celzijevih**.
RAW+	YES	NO	YES	U 18:35 temperatura zraka u **dnevnoj sobi** je 28 stupanj Celzijev.
RAW+	YES	YES	NO	U 18:35 temperatura zraka u **dnevnoj sobi** je 28 **stupnjeva Celzijevih**.
COR	YES	YES	YES	U 18:35 temperatura zraka u **dnevnoj sobi** je 28 **stupnjeva Celzijevih**.

Notification labeled RAW is the original notification generated by the event while the label COR corresponds to the correct and final form of notification, when all the required services have processed the RAW notification. Notifications labeled as RAW+ represent cases where one of the required NLG services failed to process the notification. The differences in grammatical forms as opposed to initial RAW form are emphasized with bolded words. In this sense, the evaluation was used to determine whether the wrong grammatical form affects the intelligibility of the notification. This, in turn, shows us whether it is important to have all the required NLG services proposed in this paper.

After forming the notification, it is then passed on to the speech synthesis subsystem. Prior to this step, numerical tokens (e.g. time in Table 1) are substituted with their lexical counterparts by the text normalization service. Finally, the notification is transformed to a sequence of phonemes, prosodic features are added, and a waveform is generated.

4.1. Results and Discussion

The evaluation survey was performed on 27 participants, 20 males and 7 females. Most of the participants were in their 20's (19 users), with 6 participants from 30-40, one participant over 40 and one participant over 60 years of age. In the survey, participants were required to evaluate synthesized speech samples produced by the previously

described system. Speech samples were graded on a 5-grade scale, with intelligibility, grammatical correctness, and overall quality as separate grading criteria. The goal of this evaluation is to provide feedback regarding the course of research in the future.

The evaluation dataset consisted of notifications produced by the NLG subsystem, generated from synthetic system events. The notification set contained three variations of the same message, as presented in Table 1. The first notification corresponds to the raw result generated from template (RAW). The second notification had only slight grammatical errors, corresponding to the case when one of the NLG services failed to process the notification (RAW+). The third notification was grammatically completely correct, corresponding to the case when all services were working (COR).

Furthermore, we wanted to evaluate whether grammatical correctness of the notification corresponds to the user perceived quality of the synthesized audio notification. This was done in order to see how important it is to have all the required services working and if the users would be willing to trade grammatically incorrect notifications for more human-like notifications.

For this purpose, two explored speech generation approaches were used: one based on statistical parametric speech synthesis (S1) and the other based on generative neural network approach (S2), which produced more natural, human-like speech. All notifications were synthesized using a female voice.

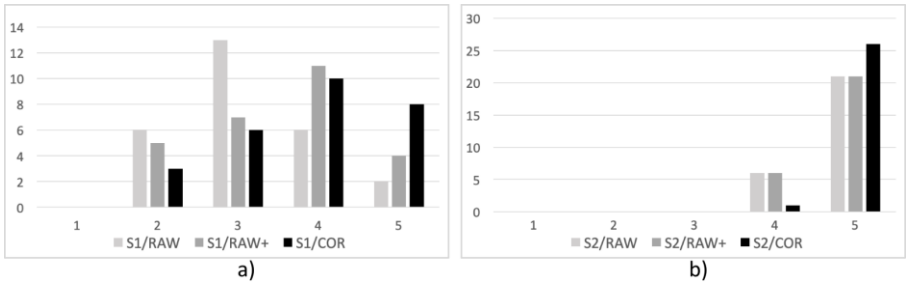

Fig. 5. Intelligibility – survey results comparing S1 (a) and S2 (b)

Evaluation results regarding intelligibility are shown in Fig. 5. The obvious conclusion is that perceived intelligibility is greater in case of high-quality speech synthesis method (S2). However, it is interesting to notice that perceived intelligibility is greater in case of S2 even when grammatical errors were present. This draws us to conclusion that users value quality of synthesized speech over grammatical correctness. This is best shown in comparison of S1/RAW and S2/RAW examples, where the lowest grade for S2 was 4, while S1 received more grades in range from 2 to 3.

Results related to grammatical correctness are displayed in Fig. 6. Despite the fact that differences in grammatical correctness were recognized (rising grades between RAW, RAW+ and COR) for both speech synthesis methods, users graded the high-quality synthesizer (S2) as more grammatically correct. Interestingly, this was not the case since both synthesizers had exactly the same inputs. This is a clear indication that quality of synthesized speech is more important for the user perceived quality.

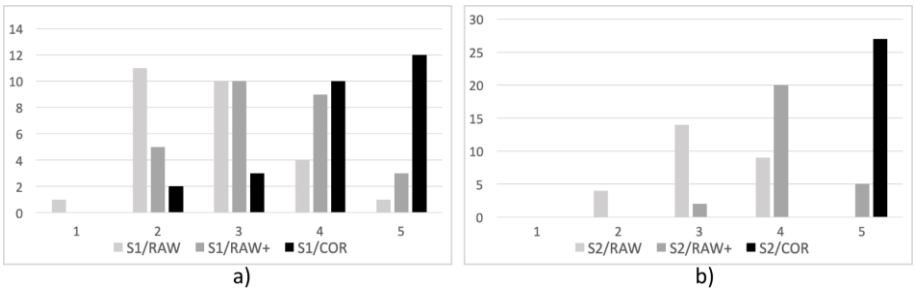

Fig. 6. Grammatical correctness - survey results comparing S1 (a) and S2 (b)

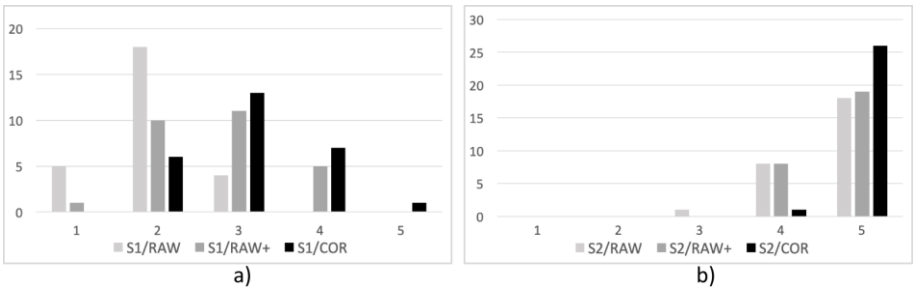

Fig. 7. Overall quality - survey results comparing S1 (a) and S2 (b)

Overall quality results are shown in Fig. 7. As expected, users regarded S2 as a far better speech synthesis method in terms of overall quality. Even in case of grammatically incorrect notification (S2/RAW), there were almost no grades below 4, which is unexpected and confirms the conclusion that users are willing to trade grammatical correctness for a more natural-sounding (or human-like) synthesized speech. On the other hand, grades regarding low-quality speech synthesis method (S1) are distributed across the entire range, with average perceived quality grade for best case (S1/COR) being 3.85, compared to 5.0 for high-quality synthesis (S2/COR).

The summary results for evaluated speech synthesis methods and notification variations are displayed in Table 2. The most interesting information is the grade difference between the two synthesis methods.

Table 2. Evaluation - NLG related services and impact on perceived quality of spoken notifications

	Intelligibility			**Grammatical Correctness**			**Overall Quality**		
	RAW	RAW+	COR	RAW	RAW+	COR	RAW	RAW+	COR
S1	3.15	3.52	3.85	2.74	3.37	4.19	1.96	2.74	3.11
S2	4.78	4.78	4.96	3.19	4.11	5.0	4.63	4.70	4.96
Diff	1.63	1.26	1.11	0.44	0.74	0.81	2.67	1.96	1.85

The intelligibility criterion directly reflects the quality of synthesized speech. Here, a considerable grade discrepancy between synthesizer S1 and S2 is present, as expected. It is interesting to notice that the smallest difference regarding intelligibility is related to the grammatically most correct notification form. This points us to a conclusion that grammatical correctness is important for intelligibility. For example, if it was not possible to develop a speech synthesizer capable of producing human-like speech for complex languages, it would definitely be required to ensure the grammatical correctness for spoken notifications, which would increase the quality perceived by users by approximately one grade.

The grammatical correctness criterion confirms that, despite low quality in case of speech synthesis method S1, the grade in its totally correct grammatical form (S1/COR) is one grade higher than the grade achieved by speech synthesis method S2 in its totally incorrect grammatical form (S2/RAW). This confirms that users noticed grammatical mistakes and corrections in case of low-quality speech synthesis method. From this we can conclude that all services described in the previous text should be present and functional when constructing the notification, regardless of the speech synthesis method which will be employed.

The differences between evaluated speech synthesis methods regarding overall quality reveal some interesting discoveries, as well. For instance, the grade range for low-quality speech synthesis method (S1) is from 1.96 to 3.11 (grade span of 1.15), while high-quality synthesis method grade span was 0.33. This again confirms that grammatical correctness affects users' quality of experience, even though in lesser extent than intelligibility.

The conclusion based on analyzed results is that both overall quality and grammatical correctness significantly affect user experience regarding evaluated spoken notifications. However, speech synthesis quality has somewhat greater influence. In case of morphologically rich languages, such as Croatian, we can conclude that grammatical correctness is especially important if there is no high-quality speech synthesis method available, since it ensures better user satisfaction in extent of an entire grade.

5. Conclusion

From the perspective of a minority language such as Croatian, designing and developing a system which can enable spoken interaction in a smart environment is a challenging endeavor. A lot of resources are required, yet there are few resources available. Additionally, it is inevitable to tackle with modeling of complex cognitive processes which are also language-related. In this paper we proposed an approach which enables us to develop a system as an orchestration of independent subsystems and their related service ecosystem. The initial results are not comparable to commercial products available for English language, but show promise.

Survey results showed that overall quality (i.e. naturalness, similarity to human speech) was the dominant factor regarding users' quality of experience. The generative neural network approach provided more natural sounding results and received better grades in all presented cases. According to grade comparison, we assume that morphological errors were perceived clearer for the better speech synthesis method, as

well. Therefore, the proposed natural language generation services have an important role regarding quality of experience when differences between proper and improper morphological forms can be perceived.

Regarding future work, implementation of an improved speech synthesis service based on deep neural networks is of highest priority. The evaluated speech synthesis subsystem was created for test purposes using limited corpora and therefore has limitations regarding a broader range of applications. Future work will be focused on improving models by using larger corpora suitable for deep learning and evaluation with more example notifications and participants. After developing a fully functional system for generating natural sounding speech, the next step would be addition of speech recognition and natural language understanding subsystems to enable spoken interaction between a human user and the system. This addition would also require significant extensions to context resolution and natural language generation subsystems in order to support more meaningful discourse.

Acknowledgment. This work has been supported in part by Croatian Science Foundation under the project 6917 “High-Quality Speech Synthesis for Croatian language” (HR-SYNTH).

References

1. Alexakis, G., Panagiotakis, S., Fragkakis, A., Markakis, E., Vassilakis, K.: Control of Smart Home Operations Using Natural Language Processing, Voice Recognition and IoT Technologies in a Multi-Tier Architecture. Designs 3(3), 32. (2019)
2. Tadić, M., Brozović-Rončević, D., Kapetanović, A.: Hrvatski jezik u digitalnom dobu. META-NET White Paper Series. Springer, Heidelberg etc. (2012)
3. Cooper, E., Li, E.: Characteristics of text-to-speech and other corpora. In International Conference on Speech Prosody. 690–694. (2018)
4. Martinčić Ipšić, S., Matešić, M., Ipšić, I.: Korpus hrvatskoga govora. Govor, Vol. 21, No. 2, 135-150. (2004)
5. Hržica, G., Kuvač Kraljević, J.: Croatian adult spoken language corpus (HrAL). FLUMINENSIA: časopis za filološka istraživanja, Vol. 28, No. 2, 87-102. (2016)
6. Kominek, J., Black, A.W.: The CMU Arctic speech databases. In Fifth ISCA workshop on speech synthesis. (2004)
7. Ito, K.: The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/. (2020)
8. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. and Weber, G.; Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), ELRA, 4218–4222. (2020)
9. Vasić, D., Brajković, E.: Development and Evaluation of Word Embeddings for Morphologically Rich Languages. In 2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM). IEEE, 1-5. (2018)
10. Cohen, P.R., Oviatt, S.L.: The role of voice input for human-machine communication. In Proceedings of the National Academy of Sciences, 92(22), 9921-9927. (1995)
11. Berdasco, A., López, G., Diaz, I., Quesada, L., Guerrero, L.A.: User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. In Multidisciplinary Digital Publishing Institute Proceedings, Vol. 31, No. 1, 51-59. (2019)
12. Edu, J.S., Such, J.M., Suarez-Tangil, G.: Smart Home Personal Assistants: A Security and Privacy Review. arXiv:1903.05593. (2019)

13. Ford, M., Palmer, W.: Alexa, are you listening to me? An analysis of Alexa voice service network traffic. Personal and Ubiquitous Computing, Vol. 23, No. 1, 67-79. (2019)
14. Hamdan, O., Shanableh, H., Zaki, I., Al-Ali, A.R., Shanableh, T.: IoT-based interactive dual mode smart home automation. In IEEE International Conference on Consumer Electronics (ICCE). IEEE, 1-2. (2019)
15. Fedotov, D., Matsuda, Y., Minker, W.: From Smart to Personal Environment: Integrating Emotion Recognition into Smart Houses. In IEEE International Conference on Pervasive Computing and Communications Workshops. IEEE, 943-948. (2019)
16. Petnik, J., Vanus, J.: Design of smart home implementation within IoT with natural language interface. IFAC-PapersOnLine, Vol. 51, No. 6, 174-179. (2018)
17. Lovrek, I: Context Awareness in Mobile Software Agent Network. RAD, Croatian Academy of Sciences and Arts. Technical Sciences. Vol. 513, 7-28. (2012)
18. Soic, R., Skocir, P., Jezic, G.: Agent-based system for context-aware human-computer interaction. In KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications. Springer, Cham., 34-43. (2018)
19. Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling. arXiv:1602.02410. (2016)
20. Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L., Yamagishi, J.: A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4804-4808. (2018)
21. Newton Dictate, https://www.newtontech.net/en/newton-dictate/, accessed in September 2020.
22. AlfaNum, http://www.alfanum.co.rs/, accessed in September 2020.
23. Martinčić-Ipšić, S., Pobar, M., Ipšić, I.: Croatian large vocabulary automatic speech recognition. Automatika, Vol. 52, No. 2, 147-57. (2011)
24. Pobar, M., Ipšić, I.: Development of Croatian unit selection and statistical parametric speech synthesis. In Proceedings of the 34th International Convention MIPRO. IEEE, 913-918. (2011)
25. Beliga, S., Martinčić-Ipšić, S.: Text normalization for croatian speech synthesis. In Proceedings of the 34th International Convention MIPRO. IEEE, 1664-1669. (2011)
26. Načinović, L., Pobar, M., Ipšić, I., Martinčić-Ipšić, S.: Grapheme-to-Phoneme Conversion for Croatian Speech Synthesis. In 32nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2009), 318-323. (2009)
27. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication, Vol. 51, No. 11, 1039-1064. (2009)
28. Oord, A.V., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv:1609.03499. (2016)
29. Oord, A.V., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G.V., Lockhart, E., Cobo, L.C., Stimberg, F., Casagrande, N.: Parallel wavenet: Fast high-fidelity speech synthesis. arXiv:1711.10433. (2017)
30. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.: Tacotron: Towards end-to-end speech synthesis. arXiv:1703.10135. (2017)
31. Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flow-based generative network for speech synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3617-3621. (2019)
32. Rouvier, M., Meignier, S.: A global optimization framework for speaker diarization. In Odyssey 2012 - The Speaker and Language Recognition Workshop. (2012)

33. Bates, M.: Models of natural language understanding. In Proceedings of the National Academy of Sciences, Vol. 92, No. 22, 9977-9982. (1995)
34. Gledec, G., Šoić, R., Dembitz, Š.: Dynamic N-Gram System Based on an Online Croatian Spellchecking Service. IEEE Access 7, 149988-149995. (2019)

Renato Šoić is a research assistant at the Department of Telecommunications of the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He received the master's degree from the University of Zagreb, in 2010. He participated in many industrial projects from different domains, including monitoring and control systems in satellite industry, mobile payment services and large-scale analytics and recommendation systems. Renato Šoić has co-authored seven conference articles and three journal articles. His research interests include speech technologies and human–computer interaction in smart environments.

Gordan Ježić is a professor at the Department of Telecommunications of the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He received the Ph.D. from the University of Zagreb in 2003. He actively participates in numerous international conferences as a paper author, speaker, member of organizing and program committees or reviewer. Gordan Ježić co-authored over 100 scientific and professional papers, book chapters and articles in journals and conference proceedings. His research interest includes telecommunication networks and services focusing on parallel and distributed systems, Machine-to-Machine (M2M) and Internet of Things (IoT) systems, mobile software agents and multi-agent systems. He is a senior member of IEEE Communication Society, IEEE FIPA, KES International, member of technical committee of IEEE SMC on Computational Collective Intelligence, and leader of technical committee of KES Focus Group on Agent and Multi-agent Systems.

Marin Vuković is an assistant professor at the Department of Telecommunications of the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He received the Ph.D. from the University of Zagreb in 2011. Marin Vuković has co-authored over 40 journal and conference papers and reviewed a number of papers for international conferences and journals. He actively participated in panels, round tables and held invited lectures with the goal of popularization of science and profession. He is a co-author of the patent at the Croatian Institute for Intellectual Property. Marin Vuković is a deputy director of "Laboratory for Security and Privacy (SPL)" and "Laboratory for Assistive Technology and Alternative and Augmentative Communication (ICT-AAC)" at the University of Zagreb, Faculty of Electrical Engineering and Computing. He is a senior member of IEEE Communications Society.

Received: April 24, 2020; Accepted: October 06, 2020