DEVELOPMENT OF SPEECH TECHNOLOGY FOR SERBIAN AND ITS APPLICATIONS

Vlado D. Delić[1], Darko J. Pekar[2], Milan S. Sečujski[1], Branislav Z. Popović[1], Edvin T. Pakoci[2], Siniša B. Suzić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Trg D. Obradovića 6, 21000 Novi Sad, Serbia
e-mail: vlado.delic@uns.ac.rs, secujski@uns.ac.rs, bpopovic@uns.ac.rs, sinisa.suzic@uns.ac.rs
[2] AlfaNum doo, Bul. Vojvode Stepe 40, 21000 Novi Sad, Serbia
e-mail: darko.pekar@alfanum.co.rs, edvin.pakoci@alfanum.co.rs

Abstract:

The paper presents an overview of the progress in the development of speech technology, most notably automatic speech recognition and text-to-speech synthesis, for Serbian and kindred South Slavic languages, achieved through the cooperation of the Faculty of Technical Sciences in Novi Sad, as well as the company AlfaNum. A particular focus is given to the progress enabled by the introduction of deep learning, as a novel paradigm with tremendously increasing popularity in the last decade, showing potential to support highly natural human-machine interaction. In the second part of the paper a description of several most notable applications of the developed speech technology is presented, in terms of performance, potentials and limitations, including the system for automatic transcription of dictated medical findings, a speech enabled virtual assistant for mobile phones, as well as speech recognition and synthesis as assistive technologies integrated into efficient aids for people with a wide range of disabilities.

Keywords: deep neural networks, speech recognition, speech synthesis, speech technology applications

1. Introduction

Artificial intelligence (AI) is a cornerstone of speech communication between humans and machines, and it has shown great progress with the development of deep neural networks (DNN) and deep learning. This shift in the paradigm of the development of speech technology [1] and the perspectives of its application, particularly with respect to the Serbian language, are the principal focus of this paper.

Speech technology for Serbian and kindred South Slavic languages have been developed for over two decades within a cooperation between the Faculty of Technical Sciences of the University of Novi Sad, Serbia, and the company AlfaNum, also from Novi Sad, with occasional participation of partners from Serbia and other countries through a number of joint projects. The term speech technology refers principally to automatic speech recognition (ASR) and text-to-speech synthesis (TTS), but in a broader sense it can also refer to tasks such as identification of the speaker or his/her emotion from speech. In the first decade of the 21st century the principal mathematical foundation for the development of speech technology were statistical models, most notably hidden Markov models (HMM), but the focus has shifted to DNN as an AI paradigm which has enabled a breakthrough in the quality of speech technology, evidenced through an increase in the accuracy and robustness of ASR as well as intelligibility, naturalness and expressiveness of TTS. Namely, with the advent of machine learning, the focus of the research community in this field has shifted from achieving a

functional human-machine dialogue based on ASR and TTS, to the development of more sophisticated capabilities of cognitive systems, such as emotion recognition and expression [2]. This shift of focus was motivated by our understanding that not only human communication and reasoning are greatly affected by emotions, and that they actually assess their interaction with machines in a way analogous to their assessment of social interaction with other humans [3]. Indeed, the need for a dialogue system to create an impression that it actually has emotions was formally established in [4] as one of the four key factors of cognitive systems.

The rest of the paper is organized as follows. Section 2 presents a retrospective of the development of ASR and TTS with special emphasis on the progress achieved owing to the introduction of deep learning and the achieved performance and potential applications of speech technology for Serbian. Section 3 describes several applications of ASR and TTS which can be considered to be of national importance since they represent an actual and substantial contribution to the preservation of the Serbian language as an under-resourced language in the digital era, and prevent technological dependence from solutions developed for languages with larger speaker bases. Of particular importance are the development of speech recognition in critically important domains such as healthcare, law enforcement and state administration, as well as the development of high-quality TTS in the domains of media and education. Coupled together, ASR and TTS enable two-way speech communication between humans and machines, and it should also be noted that, as assistive technologies, they are of extreme importance for the realization of efficient aids for people with a wide range of disabilities. The descriptions of particular applications include brief overviews of their performance, as well as potentials and limitations presented in our recent publications.

2. AlfaNum ASR and TTS Development

In the development of speech and language technology for Serbian and other kindred South Slavic languages, the Faculty of Technical Sciences and AlfaNum have jointly developed or accumuilated a large number of speech and language resources (dictionaries, text corpora, speech corpora, text or speech annotation systems). The importance of such resources becomes more important than ever before with the advent of machine learning, having in mind the importance of data in the framework of machine learning and artificial intelligence.

2.1 ASR Development

Automatic speech recognition is a process of recognition and translation of spoken language into text using computer algorithms. Research on speech recognition began in the 1950s when the first such system called „Audrey“ was devised to enable single-speaker digit recognition [5]. In the late 1960s, the use of hidden Markov models (HMMs) for speech recognition was proposed, which in the next two decades enabled recognition on dictionaries containing several thousand words [6]. The statistical approach based on HMMs and Gaussian mixture models (GMMs) was also used in the development of the first system for speech recognition in the Serbian language. [7] The system functioned under the assumption that the speech signal can be observed as a short-term stationary signal, i.e., frames are extracted every 10 ms and then transformed into appropriate features (sets of cepstral coefficients obtained by applying short-term Fourier transform and cosine transform for spectral correlation). Speech decoding in terms of determining the most likely spoken word sequence was made possible using Viterbi algorithm, while language modelling was conducted by applying the *n*-gram language model [8]. The first commercial systems in the Serbian language enabled speech recognition using dictionaries containing several thousand words.

The development of ASR systems also involved recording and processing different audio databases for acoustic modelling as well as processing various textual resources for language

model training. This required the exhaustive work of a large number of human labellers who manually marked the boundaries of every phoneme in the sentence. Today's system for automatic speech recognition in Serbian is based on deep neural networks, and the preparation of spoken language databases is now significantly accelerated [9]. Flat-start training is applied, which in addition to the audio database requires transcriptions in a textual format (phoneme boundaries are determined automatically, using an iterative algorithm based on HMM-GMM models, minimum phoneme error and speaker adaptive training). The system uses chain sub-sampled time-delay neural network (TDNN) models, with a 143-dimensional feature vector which typically contains 40 high-resolution Mel-frequency cepstral coefficients (HiRes MFCCs), 3 pitch-related features – the probability of voicing, log-pitch and delta-pitch, and a corresponding 100-dimensional iVector per frame. Depending on the domain of application, 5 to 15 layers with 512 or 1,024 neurons are trained. Less complex systems, such as voice assistants on mobile phones, require fewer layers and neurons to be able to recognize spoken commands in real-time [10]. Owing to the phonetic similarity of South Slavic languages (primarily Serbian and Croatian), it is possible to exploit resources in both Serbian and Croatian during the training of acoustic models, by applying transfer learning to Serbian or other appropriate domain databases, either by tuning the whole network or training the additional neural network layers [11]. In total, over 1,500 hours of speech were collected in both Serbian and Croatian, including audiobook recordings recorded in a studio environment and spoken by professional speakers, radio talk show recordings and mobile phone recordings. The data is further augmented by noise addition, speed and pitch perturbance.

N-gram language models approximate the probability of a word sequence based on its occurrence in the training corpus. The biggest problem of *n*-gram models is the problem of data sparsity, where a large number of higher *n*-grams do not appear in any of the corpora, while the vast majority of them only appear rarely. This problem is especially pronounced in the case of highly inflected languages, such as Serbian. An additional problem is the impossibility of modelling longer contexts, since the width of the context is determined by the order of n-grams, and increasing the order exacerbates the problem of data sparsity. With this in mind, state-of-the-art ASR systems for Serbian use class-based recurrent neural network language models (RNNLMs), which project each word into a compact continuous vector space that can be described by a limited set of parameters and are also able to model sequences of arbitrary length using their recurrent connections [12]. Additional incorporation of morphological features in the process of language model training has enabled increased accuracy of such systems, which in certain cases, such as the system for dictating medical findings in Serbian, reaches an accuracy of over 98% on dictionaries containing about 250,000 words [13]. The architecture of this system is shown in Fig. 1.

2.2 TTS Development

The development of TTS systems for Serbian has been following general trends in TTS research community. The earliest commercially widely used TTS systems were based on concatenation of speech segments of different lengths from a prerecorded speech database [14]. The first Serbian TTS system was also based on the concatenative approach and is presented in [15]. Although speech generated with concatenative approach has a high level of intelligibility and naturalness, there are audible artefacts at many segment concatenation points. The flexibility of such a TTS, which refers to the possibility of changing speech features that may carry information related to speaker identity or speech style, is also low with this approach. Namely, in order to obtain a reasonable degree of flexibility, either in terms of speaker identity or speech style, one would often have to use prohibitively large and complex speech corpora.

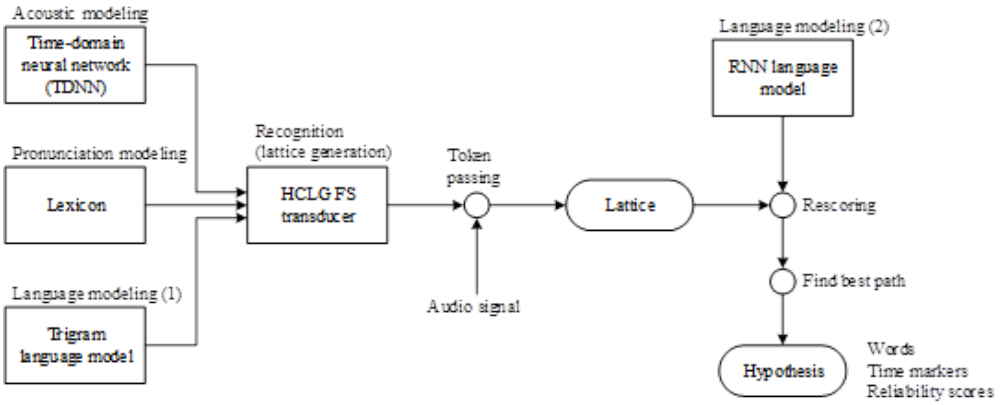

Fig. 1. Block diagram of the speech recognition system for dictation of medical findings.

In early 2000s the focus of TTS research community shifted to parametric approaches, among which the Hidden Markov model (HMM) method was most dominant [16]. HMM synthesis is based on creating models which, based on input linguistic features, could generate acoustic features which are converted to the speech samples by an appropriate vocoder. In general, the size of the speech corpora needed to train such systems is smaller compared to the corpora needed for the concatenative approach and they enable more flexibility in changing speech characteristics. Despite these advantages, parametric systems were not used extensively for commercial speech synthesis since generated speech, although free of concatenation artefacts, sounded muffled and buzzy. The Serbian HMM system was introduced in [17], while a comparison between Serbian concatenative and HMM systems is given in [18].

The rapid development of TTS based on the usage of deep neural networks has started in the early 2010s. In the early stage of using DNNs, the TTS module for modelling speech parameters based on input linguistic features using HMMs (also referred to as *back-end*) were replaced by the different types of neural networks [19], while standard vocoders were still used. The first such system for the Serbian language is presented in [20]. An in-depth analysis of the system and its comparison with respect to the HMM based one is given in [21]. This system was also successfully applied in generating personalized TTS based on amateur data [22], and it was later extended to support expressive speech synthesis in different voices [23], even when no training data of the target speaker in the target speech style is available [24]. The model proposed in [23] is based on two neural networks, one predicting phone durations and the other predicting acoustic features, using information about the speaker and speaking style in an embedded form, as shown in Fig. 2. In this way the model itself has the opportunity to establish similarities and differences between particular speakers and speaking styles, and learns to generalize more easily.

The most recent trends in TTS for Serbian are based on creating so-called neural vocoders, which directly generate speech samples conditioned on some specific input [25]. An initial system utilizing efficient neural vocoder in Serbian is presented in [26].

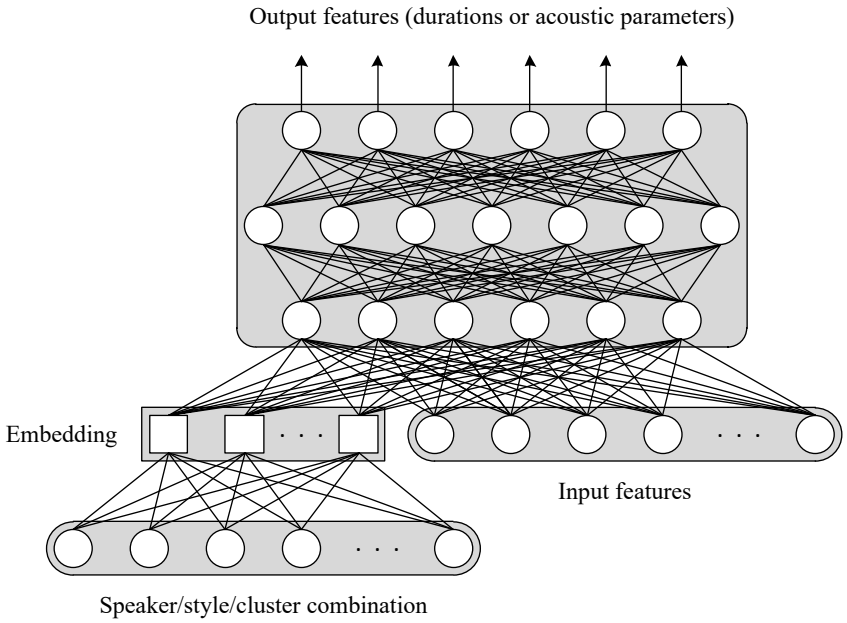

Fig. 2. The architecture of the model proposed in [23] for two neural networks that predict phonetic segment durations or acoustic features of synthesized speech respectively.

3. ASR and TTS applications

Early applications of AlfaNum ASR were constrained to the recognition of small and medium vocabularies, because of the technological limitations and inability to accurately recognize spoken text within a large vocabulary. Some of the applications included enhancing IVRs (*Interactive Voice Response Systems*) with recognition of numbers (e.g. sums of money), cities or personal names, certain commands, or some combinations. In some use cases it was very helpful, since the alternative was choosing options with telephone keyboard. When smartphones appeared, AlfaNum developed its own Voice Assistant application, called Axon, which enabled users to call specific contact or a number by just saying appropriate command followed by contact name (e.g. "*Call Peter Smith on mobile"*). Since it also used limited vocabulary, it was quite accurate and it worked locally on the device (i.e. did not require internet access at all times).

Recent advancements in large vocabulary ASR paved the way for some other applications, such as dictation or transcription of prerecorded speech. AlfaNum is currently focused on developing customized ASR for dictation of medical findings [13], since it is quite specific, includes Latin and many other uncommon words and formulations, and thus is poorly covered by "general purpose" speech recognition systems (such as Google or Microsoft ASRs). Moreover, AlfaNum offers an in-home solution, which is very important for clients who demand privacy and do not accept solutions that rely on speech recognition services offered by a third party. For the same reasons, a system for dictation of legal documents is being developed along the same principles, expected to find its use within legal institutions and state administration.

Initial applications of TTS were mainly intended for the people with disabilities, most notably the visually impaired [27]. Coupled with so-called screen readers, AlfaNum TTS enabled them to work with their PCs completely independently from any other person. As smartphones emerged, AlfaNum developed a TTS application for Android, which provided the visually impaired with similar functionality. Other early applications were also in call centres, where AlfaNum TTS was able to relieve live agents of some of their burden and thus enhance the capacity of the call centre.

As the quality and flexibility of TTS matured, it started to be used for voice enabling of web sites, in chat bots, for entertainment purposes and more. The current technology also enables quick development of new voices based on very small recording sample, so-called voice cloning [22, 23]. This has recently been used on web site of national television, where each news item can be read out aloud in the (cloned) voice of one of their most popular speakers.

4. Conclusions

The paper has presented an overview of the progress in the development of speech technology for Serbian and kindred South Slavic languages, brought about principally by the shift of the paradigm in the research community from conventional methods to machine learning and artificial intelligence. These novel and increasingly popular approaches have been extremely successful in improving the quality of human-machine interaction and bringing it closer to natural – not only in terms of the accuracy and robustness of ASR and the intelligibility and naturalness of the TTS, but in terms of apparent consciousness of the system and its ability to establish complex social bonds with its collocutor through its capability to recognize emotion as well as to produce expressive speech. As it is more comfortable to perceive an artificial system as a real person than to think about all the implications of a communicating machine, our innate tendency to behave naturally in the interaction with machines should only be supported by the improvement of the naturalness of the way machines interact with us.

Acknowledgement: This research was supported by the Science Fund of the Republic of Serbia, through the project grant agreement No. 6524560: “Speaker/Style Adaptation for Digital Voice Assistants Based on Image Processing Methods (AI-S-ADAPT)”. Speech corpora used in the research were provided by Speech Morphing Systems Inc., Campbell, CA, United States of America, for research purposes.

References

[1] Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., Simić, N., Suzić, S., Delić, T., Speech Technology Progress Based on New Machine Learning Paradigm, Computational Intelligence and Neuroscience, Article ID 4368036, 19 pages, 2019.

[2] Picard, R. W. Affective computing. MIT Press, Cambridge, MA, 1995.

[3] Nass, C. I., Yen, C. The man who lied to his laptop: what machines teach us about human relationships. Current Trade Penguin Group, New York, NY, 2010.

[4] Picard, R. W.: What does it mean for a computer to “have” emotions. In: Trappl,. R, Petta, P., Payr, S. (eds) Emotions in humans and artifacts. MIT Press, Cambridge, MA, pp. 213-235, 2003.

[5] AbouElhasan, N., Elboraee, T., Mohamed, H., Adel, N., Eid, M.M., Survey on Automatic Speech Recognition, Journal of Computer Science and Information Systems, 11, 2020.

[6] Juang, B.H., Rabiner, L.R., Hidden Markov Models for Speech Recognition, Technometrics, 33(3), pp. 251-272, 1991.

[7] Popović, B., Janev, M., Pekar, D., Jakovljević, N., Gnjatović, M., Sečujski, M., Delić, V., A Novel Split-and-Merge Algorithm for Hierarchical Clustering of Gaussian Mixture Models, Applied Intelligence, Vol. 37, No. 3, pp. 377-389, 2012.

[8] Delić, V., Sečujski, M., Pekar, D., Jakovljević, N., Mišković, D., A Review of AlfaNum Speech Technologies for Serbian, Croatian and Macedonian, in Proc. Int. Language Technologies Conference IS-LTC, Vol. 6, pp. 257-260, 2006.

[9] Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., Delić, V., Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolkit, in Proc. of the 17th SPECOM, Speech and Computer, Athens, Greece, pp. 186-192, 2015.

[10] Popović, B., Pakoci, E., Jakovljević, N., Kočiš, G., Pekar, D., Voice Assistant Application for the Serbian Language, 23rd Telecommunications Forum TELFOR, Belgrade, Serbia, pp. 858-861, 2015.
[11] Popović, B., Pakoci, E., Pekar, D., Transfer Learning for Domain and Environment Adaptation in Serbian ASR, TELFOR Journal, Vol. 12(2), pp. 110-115, 2020.
[12] Pakoci, E., Popović, B., Methods for Using Class Based N-gram Language Models in the Kaldi Toolkit, in Proc. of the 23rd SPECOM, Speech and Computer, St. Petersburg, Russia, 27-30 September 2021, Springer, LNCS, Vol. 12997, pp. 492-503, 2021.
[13] Popović, B., Pakoci, E., Pekar, D., Automatic Speech Recognition System for Dictating Medical Findings, 7th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN, Belgrade, Serbia, 28-30 September 2020, pp. 12-17.
[14] Hunt, A.J., Black, A.W., Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database, in Proc. ICASSP 1996, Vol. 1, pp. 373–376, 1996.
[15] Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., Delić, V., AlfaNum System for Speech Synthesis in Serbian Language, in Proc. Text, Speech and Dialogue, Brno, Czech Republic, pp. 237-244, 2002.
[16] Zen, H., Tokuda, K., Black, A.W., Statistical Parametric Speech Synthesis, Speech Communication, Vol. 51(11), pp. 1039-1064, 2009.
[17] Pakoci, E., Mak, R., HMM-based Speech Synthesis for the Serbian Language, 56th ETRAN, Zlatibor, Serbia, Vol. TE4, pp. 1-4, 2012.
[18] Pakoci, E., Mak, R., Ostrogonac, S., Subjective Assessment of Text to Speech Synthesis Systems for the Serbian Language, 20th Telecommunications Forum TELFOR, Belgrade, Serbia, pp. 732-735, 2012.
[19] Zen, H., Senior, A., Schuster, M., Statistical Parametric Speech Synthesis Using Deep Neural Networks, in Proc. IEEE ICASSP 2013, Vancouver, Canada, pp. 7962–7966, 2013.
[20] Delić, T., Sečujski M., Speech Synthesis in Serbian Based on Artificial Neural Networks, 24th Telecommunications Forum TELFOR, pp. 1-4, IEEE, 2016.
[21] Delić, T., Sečujski M., Suzić, S., A Review of Serbian Parametric Speech Synthesis Based on Deep Neural Networks, Telfor Journal, Vol. 9(1), pp. 32-37, 2017.
[22] Delić, T., Suzić S., Sečujski M., Ostojić V., Deep Neural Network Speech Synthesis Based on Adaptation to Amateur Data, IcETRAN 2018, Palić, Serbia, pp. 1249-1252, 2018.
[23] Sečujski, M., Pekar, D., Suzić, S., Smirnov, A., and Nosek, T: "Speaker/style-dependent neural network speech synthesis based on speaker/style embedding", Journal of Universal Computer Science, Vol. 26(4), pp. 434-453, 2020.
[24] Suzić, S., Delić, T., Pekar, D., Delić, V., and Sečujski, M. Style Transplantation in Neural Network-based Speech Synthesis, Acta Polytechnica Hungarica, Vol. 16(6), pp. 171-189, 2019.
[25] Van Den Oord, A. et al., WaveNet: A Generative Model for Raw Audio, in Proc. SSW 2016, Vol. 125, No. 2, 2016.
[26] Suzić, S., Pekar, D., Sečujski, M., Nosek, T., Delić, V., HiFi-GAN Based Text-to-Speech Synthesis in Serbian (under review)
[27] Delić, V., Sečujski, M., Jakovljević, N., and Mišković, D. Assistive character of speech technology, Speech and Language, Belgrade, Serbia, pp. 18-26, 2017.