Retrospective and Perspectives of TTS & STT Technology Development and Implementation for South Slavic Under-Resourced Languages

Milan Sečujski[1](✉), Branislav Popović[1], Darko Pekar[2], Nikša Jakovljević[1], Edvin Pakoci[2], Siniša Suzić[1], Tijana Nosek[1], Nikola Simić[1], Vuk Stanojev[1], and Vlado Delić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
secujski@uns.ac.rs
[2] AlfaNum Ltd, Novi Sad, Serbia

Abstract. Speech technologies such as text-to-speech (TTS) and speech-to-text (STT) are becoming increasingly applicable. Significant improvements in their quality are driven by advancements in deep machine learning. The ability of devices to deeply understand human speech and generate appropriate responses is a hallmark of AI capabilities. Developing speech technology requires extensive speech and language resources, which is why many languages with smaller speaker bases lag behind widely spoken languages in the development of speech technologys. Prior to the deep learning (DL) paradigm, hidden Markov models (HMM) and probabilistic approaches dominated speech technology development. This paper reviews the challenges and solutions in TTS and STT development for Serbian, highlighting the transition from HMM to DL. It also explores the future prospects of speech technology development for under-resourced languages and its role in preserving these languages.

Keywords: Speech Technology · Development and Implementation · Text-to-Speech · Speech-to-Text · Hidden Markov Models · Deep Neural Networks

1 Introduction

The development of speech technology, encompassing both text-to-speech (TTS) and speech-to-text (STT) systems, has revolutionized human-computer interaction and significantly impacted numerous fields, including accessibility, communication, and artificial intelligence. While the initial research focus was to address fundamental challenges in signal processing and language modeling, the last few decades have seen remarkable improvements, driven primarily by the emergence of machine learning algorithms and the vast availability of training data. With the advent of statistical methods and, more recently, deep learning (DL) techniques, modern TTS and STT systems have become more robust, adaptive, and capable of learning from large-scale datasets. These systems are now ubiquitous, powering virtual assistants, voice chatbots, transcription services, and accessibility tools that have become integral to everyday life [1].

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 23–42, 2025.
https://doi.org/10.1007/978-3-031-77961-9_2

The paper explores the historical evolution of speech technology, outlining the key milestones, technical challenges and innovations, with particular focus on under-resourced languages [2]. Namely, while both technologies have made significant strides in dominant languages like English, their application to under-resourced languages has historically lagged. Under-resourced languages often lack large-scale linguistic datasets, making the development of high-quality TTS and STT systems for these languages particularly challenging. The authors of the paper are members of a research team behind the development of first fully functional and commercially widely applied TTS and STT systems for Serbian and several other South Slavic languages, and besides giving a general historical review of speech technology, the authors will focus on the issues and challenges they have encountered in the development of speech technology for under-resourced languages.

The remainder of the paper is structured as follows. In Sect. 2 we will present a historical overview of speech technology, with emphasis on their language-dependent elements. Section 3 will focus on particular challenges encountered in the development of speech technology for Serbian and other kindred languages. Section 4 focuses on the issues related to under-resourced languages, and Sect. 5 will conclude the paper.

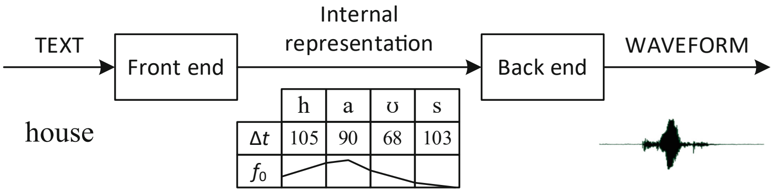
2 Evolution of Speech Technology

Technological advances in the field of artificial intelligence and machine learning have been followed by our perpetually changing perspective on speech technology. In their beginnings, both speech recognition and synthesis have been viewed as typical signal processing areas and focused on topics such as speech coding [1]. The development of first commercial TTS or STT systems required specialized knowledge of linguistics, turning speech technology into a prime example of interdisciplinary knowledge area, where the tasks of conversion of text into speech or vice versa are decomposed into smaller subtasks corresponding to different tasks in human speech recognition and production, requiring different knowledge and relying on different types of speech and language resources (speech corpora, text corpora, lexicons, rule lists, statistical models). However, the recent developments in artificial intelligence and machine learning have shifted the focus towards deep learning systems using sophisticated neural network architectures whose components exhibit little correspondence with particular human speech production or recognition tasks. As a result, both TTS ans STT are now viewed as typical instances of machine learning problems, whose success is due to the ability of neural networks to model the complexity of human language, learn from vast amounts of data, and generalize to unseen speech or text inputs. One of the most important advantage of this shift in perspective is the possibility of using transfer learning, which allows pre-trained models (typically trained on a large, resource-rich dataset in a major language) to be fine-tuned to new, often low-resource languages. This approach reduces the need for massive amounts of labelled data, which is often unavailable for under-resourced languages [2 – 4].

2.1 Text-To-Speech Synthesis

The history of text-to-speech (TTS) technology spans several decades, evolving from early mechanical devices to today's sophisticated systems based on artificial intelligence. The first known speech synthesizer, VODER, was developed by Homer Dudley at Bell Labs in 1939 [5]. It could produce basic speech sounds using a keyboard but required manual operation. In 1961 IBM created the *IBM 704*, one of the earliest examples of computer-generated speech, which used formant synthesis to emulate human vocal tract shapes. The first full TTS system for English was introduced in 1968 by Teranishi and Umeda [6]. In the 1980s, digital signal processing advanced TTS with the development of DECtalk, relying on a source-filter algorithm [7], which provided a more natural-sounding synthesized voice.

The most common feature of the first commercial text-to-speech systems, able to convert any text in a given language (in this case English) into speech, was their internal structure, which was divided into parts charged with language processing (referred to as *front end*) and signal processing (referred to as *back end*) [8]. Until quite recently, this division, shown in Fig. 1, has represented the joint feature of all practically applicable text-to-speech architectures.



	h	a	ʊ	s
Δt	105	90	68	103
f_0				

Fig. 1. Internal structure of a classical TTS system.

Front End. The front end firstly converts raw text containing symbols such as numbers and abbreviations into their orthographic equivalents, which is often referred to as *text pre-processing* or *text normalization.* The front end then assigns phonetic transcriptions to each word, which is referred to as *text-to-phoneme* or *grapheme-to-phoneme* conversion. The text is then segmented out into prosodic units such as phrases and sentences. The output of the front end is the linguistic representation of the input text, including its phonetic transcription and prosodic information. The back end, also referred to as the *synthesis module,* subsequently converts the linguistic representation from its symbolic form into sound. In most TTS systems this task, which is practically language independent, includes the computation of the desired prosody features (phone durations, pitch contour), which are subsequently imposed on output speech waveforms [9].

Humans are able to perform tasks related to text normalization, grapheme-to-phoneme conversion and prosody generation automatically, owing to inference capabilities of their brains. In doing so, humans unconsciously exploit their entire linguistic competence, which is most notable in the task of prosody generation. Namely, although prosody is predominantly affected by lexis and syntax, it also appeals to higher levels of linguistic competence of the reader, including semantics and pragmatics. In machines, the modules for prosody generation are unavoidably simplistic, and are usually further segmented into morphological analysis, contextual analysis and syntactic-prosodic parsing. For most languages, the morphological analyzer proposes all possible part-of-speech (POS) categories for each word, the contextual analyzer considers each word in its context and reduces the list of its possible POS categories to a very small number of highly probable hypotheses, while the syntactic-prosodic parser examines the remaining search space relating it to the expected prosodic realization [8]. All front-end tasks are heavily language dependent, which means that a front end for each new language had to be developed anew. Modules for all these tasks could be implemented as either rule based or based on some form of machine learning, but in either case, they required the existence of a speech or language resource whose creation typically required a significant effort of experts (e.g. rule sets or labelled text or speech data). The advent of machine learning and neural networks has introduced many changes into this paradigm. Initially, neural networks were used as the method of choice for particular front-end tasks, most notably grapheme-to-phoneme conversion [10, 11]. They have achieved notable results in prosody generation from linguistic cues identified by the front end [12], and the recent advances in end-to-end TTS are aimed at completely eliminating the need for the front end as a separate module in a TTS pipeline.

Back End. The back end, also referred to as the *synthesis module,* converts the linguistic representation from its symbolic form into sound. The widespread usage of the TTS technology came with the introduction of concatenative synthesizers, with the idea of producing speech by concatenation of prerecorded speech segments. While some of the early systems used a fixed-size unit inventory for synthesis [13, 14], a true improvement in speech quality came with dynamic unit selection from large speech databases [15]. Although this approach, assuming a very large speech database is available, produces high-quality speech, there are still audible glitches at the concatenation points if the appropriate units cannot be found in the database. Furthermore, this approach is also extremely inflexible in terms of changing the speaking style or the voice of the speaker, which can be done only by recording and annotating a new speech database.

With the increasing popularity and demand for TTS, the demand has also grown for algorithms able to produce different voices and speaking styles from small data samples. The turn of the century saw the advent of statistical parametric speech synthesis, based on modelling the spectrum, fundamental frequency, and duration of speech by multispace probability distribution hidden Markov models (HMM) and multidimensional Gaussian distributions [16]. This approach enables the transformation of a speaker-independent system toward a target speaker using very small samples of speech data [17], creating expressive voices [18], as well as multilingual voices [19]. However, this method never achieved the naturalness of concatenative TTS, principally due to smoothness caused by modelling similar contexts with the same Gaussian mixtures, but also to the use of inferior vocoders, i.e. systems that produce speech waveforms from predicted acoustic features. A detailed review of HMM-based TTS can be found in [20]. Some approaches have combined parametric synthesis with unit selection, which is referred to as hybrid synthesis. The most common hybrid systems in general use parametric based models to drive unit selection [21, 22].

Some of the first attempts to use neural networks for TTS are reported in [23]. However, this approach has since gained popularity and eventually taken precedence over other approaches, mostly owing to recent development of computer hardware, particularly graphical processing units (GPUs). Deep neural networks (DNN) replaced decision trees and Gaussian mixture models in non-linear mapping of linguistic features to acoustic features [24]. They also represent a form of parametric synthesis, in that a model is used and trained on a large dataset, inferring values of parameters that will be used to synthesize speech at runtime. Intelligible and natural sounding synthetic speech could be produced even by relatively simple feedforward neural networks, and further improvements were achieved by using long short-term memory (LSTM) neurons [25], generative adversarial networks [26] and stacked bottleneck features [27].

Advanced TTS Features and Approaches. Deep neural networks have also introduced advanced possibilities such as flexible synthesis in different voices and speaking styles. Most methods for creating new voices using limited amounts of training data are based on multispeaker models, requiring a large database consisting of multiple speakers, with each speaker usually represented with less data than in case of single-speaker models [28]. In such models the variety of contextual information and better network generalization usually yield higher TTS quality. Different modalities for speaker representation have been used, including unique speaker vectors [29, 30] as well as the division of the neural network into parts shared across all speakers and speaker-specific parts [31].

The ability of a TTS to convey different emotional states or styles is a necessity for many applications, since it has been shown that emotion, mood, and sentiment affect attention, memory, performance, judgment, and decision-making in humans [32]. Initial approaches to emotional speech synthesis were focused on statistical modelling of speech parameters with HMMs [33, 34] and Gaussian mixture models [35], while more recent advances exploit deep neural networks [36, 37] and deep bi-directional LSTM (DBLSTM) [38, 39]. Further improvement in performance has been achieved with end-to-end neural network architectures [40, 41], while some of the most recent advances include synthesis of mixed emotions [42].

A significant advance in the quality of DNN TTS came with the WaveNet architecture [43], able to directly predict raw audio samples instead of using a vocoder, relying on predictive distributions dependent on previous audio samples. Conditioned on linguistic features derived from text and speaker identity, it significantly outperforms all other TTS systems, and its drawbacks related to extreme computational complexity were somewhat mitigated by the introduction of approaches such as Parallel WaveNet [44]. A similar model called DeepVoice [45], was based on replacing all parts of TTS pipeline by corresponding independently trained DNNs, but this resulted in a cumulative error in synthesized speech in the end.

As opposed to WaveNet and DeepVoice, which still use some form of front end and generate speech based on lexical features, there are systems which use raw orthographic text as input, such as Tacotron [46], Tacotron 2 [47], and Deep Voice 3 [48]. Tacotron outputs spectrograms that are transformed to speech samples using the Griffin-Lim algorithm, which also introduces artifacts in generated speech. On the other hand, Tacotron

2 system-generated spectrograms are used for conditioning standard WaveNet architecture, which generates speech samples. DeepVoice 3 can output spectrograms or other features which can be used as input to some waveform synthesis models.

Adaptation to new speakers has also been investigated in end-to-end systems [49, 50] as well as synthesis in different styles [40, 51]. Tacotron 2 offers high speech quality but can be slow and prone to issues like word skipping. FastSpeech [52] improves on this by using a Transformer network for faster, parallel mel-spectrogram generation, reducing word skipping and allowing smoother voice speed control. While FastSpeech relies on a complex teacher-student distillation process and suffers from inaccurate duration predictions and information loss, FastSpeech 2 [53] addresses these issues by training directly with ground-truth data and incorporating additional speech variations like pitch and energy, improving training speed and voice quality.

End-to-end systems, while eliminating the need for detailed labeling (such as prosody annotation), require vast amounts of data, which must typically be of high quality and often from the same speaker to achieve high-quality TTS. However, even in these conditions, such systems can struggle with certain aspects. One significant drawback is the lack of control over specific language-dependent features [54] or the exact output, which can lead to unwanted artifacts or distortions known as hallucinations.

One of the advanced approaches for TTS is VALL-E [55], a neural codec language model. Unlike previous methods, which treat TTS as continuous signal regression, VALL-E frames it as a conditional language modeling task. By treating TTS as a sequence generation problem, VALL-E leverages discrete neural audio codecs and a GPT-3-like architecture for its robust performance. Owing to in-context learning, it can synthesize high-quality, personalized speech from just a 3-s speech sample. VALL-E outperforms existing zero-shot TTS in naturalness and speaker similarity.

Extensions include VALL-E-X for cross-lingual zero-shot TTS [56], and VALL-E-R [57], which enhances speech generation robustness with phoneme monotonic alignment. VALL-E 2 further improves performance with repetition-aware sampling and grouped code modeling, achieving human-level parity on LibriSpeech and VCTK datasets. MELLE [58], another approach, generates mel-spectrograms directly from text, bypassing vector quantization. VALL-E offers superior zero-shot TTS performance, speaker adaptation, and control over diverse speech attributes but comes at the cost of higher computational requirements.

Another advanced TTS system is YourTTS [59], a multilingual zero-shot end-to-end TTS. Built on the VITS framework, it allows for accent and style transfer, which means it can synthesize speech with the style of a specific speaker, even in a different language. Its zero-shot learning feature enables it to mimic new voices based on short audio samples. It is usually trained on a large multilingual dataset for multiple speakers and uses neural waveform generation methods such as HFG [60] or WaveGlow [61].

While YourTTS offers broad multilingual capabilities and zero-shot learning for new speakers, VALL-E focuses on high-fidelity voice cloning and adaptation to specific voices. YourTTS is more versatile across languages and accents, whereas VALL-E excels in replicating individual voices with high accuracy.

While many zero-shot multi-speaker TTS systems, like YourTTS and VALL-E-X, are limited to several high-resource languages, the XTTS system addresses this limitation by enabling multilingual training and improving voice cloning [62]. Building on the

Tortoise model, XTTS introduces novel modifications for faster training and inference, and it has been trained in 16 languages, achieving state-of-the-art results in most of them. This advancement significantly broadens the applicability of zero-shot TTS to include low and medium resource languages.

The development of Serbian TTS has kept pace with advancements in modern technology, ensuring that its quality level has always been on par with TTS systems for global languages. As early as, in the 2000s, the first Serbian concatenative TTS was developed at the Faculty of Technical Sciences in Novi Sad [63]. Within the collaboration with the company AlfaNum, founded in 2003, the system was continuously improved and initially applied as a screen reader for the visually impaired. The first HMM-based TTS for Serbian was created in 2012 [64], but its quality was not sufficient to replace the already high-quality concatenative TTS for practical applications. However, in 2017 a DNN-based TTS for Serbian was developed [65], which soon surpassed the quality of concatenative TTS, while also enabling flexibility in multi-speaker synthesis [66]. A further step, achieving synthesis quality nearly indistinguishable from human speech, was made possible with the use of HFG-based vocoders [67]. Today, work continues on further improvements, heading towards end-to-end systems [45].

2.2 Speech-To-Text (Speech Recognition)

The first electrical STT systems, developed in the 1950s and 1960s exploited formant energies to recognize isolated phonemes, syllables and digits [68, 69]. The common approach for the first generation of STT systems exploited knowledge of articulatory and acoustical phonetics.

One of the main issues in these first systems were the variations in the duration of the same acoustic unit, which was overcome in the 1970s with the introduction of dynamic programming [70, 71]. Another advance was that instead of formants, linear predictive coefficients (LPCs) were introduced [72], under the assumption that the vocal tract can be modeled as an all-pole system, which yielded a more precise acoustic representation as in this way the entire spectrum envelope was taken into consideration.

The development of digital electronics in the 1970s and the 1980s shifted the focus towards more complex STT tasks – STT systems were required to recognize entire sentences with vocabularies containing hundreds of different words [73, 74]. The task was split into 3 layers – acoustic, lexicon and grammar/language layer. The acoustic layer connected acoustic representations of phonemes with the phonemes or simpler recognition units. The lexicon layer connected these acoustic units with the words, and grammar/language layer defined possible sequences of words to reduce the complexity of the search space. At the same time, acoustical modelling began to be treated as a problem of sequence decoding in noisy telecommunication channels [75].

This statistical approach based on hidden Markov models (HMMs) [76–78], was the most prevalent approach for acoustic modeling until the early 2010s and the emergence of deep learning. HMM in combination with a Gaussian mixture model (GMM) was an effective way to model time (HMM) and acoustic (GMM) variability of phonemes. To model coarticulation, a context dependent phoneme (i.e. triphone) became a basic modeling unit [79], with triphones spanning 3 HMM states. As increasing the number of models with a fixed amount of available training data reduces the amount of data for the training of each model, different state tying procedures were proposed [80]. To increase the robustness to noise new features based on human perceptual model were introduced, such as mel-frequency cepstral coefficients (MFCCs) [81] and perceptual linear predictive coefficients (PLP) [82]. Since HMM in combination with GMM is a generative model, in order to reduce in-class variability different normalization methods were introduced. Various cepstral mean and variance normalization techniques were introduced to reduce the channel variability in case of MFCC [83–85] and PLP [86]. One of the reasons for the longevity of HMM-GMM is the efficient introduction of discriminative training criteria in model training (maximum mutual information [87, 88], minimum classification error [89] and minimum phoneme error [90]). However, discriminative models gain accuracy if the number of observations per parameter is sufficiently large [91, 92].

The knowledge of the relationships between words and their phonetic transcriptions is typically stored in lexicons, which are usually created manually. Initial language models represented manually created graphs allowing limited numbers of possible word sequences. As the number of possible words rises, it becomes impractical to create such models manually and statistical *n*-gram models were introduced [93, 94]. *N*-gram models calculate scores proportional to the probabilities of *n*-word-long sequences based on texts from newspapers, books and other documents rather than spontaneous language. Different methods have been applied to achieve *n*-gram smoothing [95, 96].

Recent years brought a significant shift in paradigm – statistical based systems have been replaced with systems based on artificial neural networks, or more precisely, deep neural networks (DNN). Although there were successful experiments with STT based on neural networks in the 1980s and the 1990s [97, 98], their low speed was a significant problem. The first paper reporting comparable performance between neural network based STT systems and conventional ones was [99], and 3 years later DNN were reported to outperform state-of-the-art HMM-GMM systems by a wide margin [100]. A big step towards end-to-end models was made by introducing connectionist temporal classification (CTC), which allows DNN training for a sequence labelling task with unknown input-output alignment [101]. Several years later, end-to-end recurrent neural network (RNN) with beam search reached the performance of benchmark systems [102], eliminating much of the complex infrastructure of modern STT systems. State-of-the-art systems of today are based on transformers [103, 104]. The introduction of transformers trained in a semi-supervised manner has overcome problems related to training STT models for low-resource languages [105].

Although self-supervised models such as wav2vec [106] or wav2vec-S [105] can learn speech representations, they require adaptation for specific tasks such as STT. On the other hand, OpenAI Whisper [107] demonstrated the ability to perform STT (and tasks such as language recognition or translation) without additional fine-tuning, but with a requirement for 680,000 h of multilingual data. Whisper supports 99 languages, but with a huge disproportion in the amount of data for each language (65% of training data is English), which is reflected in higher WER for low-resource languages. Initial efforts in fine-tuning Whisper to Serbian, based on large existing datasets for Serbian and Croatian [108], are described in the following chapter.

3 Implementation Challenges and First Applications in Serbian

In its treatment of the issue of under-resourced languages in the development of speech technology, the paper focuses specifically on Serbian and related South Slavic languages. The early development has been driven by a collaborative research effort between the Faculty of Technical Sciences in Novi Sad, Serbia, and the company AlfaNum. As the only team consistently working on speech technology in the region, they had to create the first speech and language resources for several South Slavic languages, develop tools for speech annotation, design application programming interfaces (APIs), and provide support for various operating systems and platforms.

AlfaNum TTS [109] is a leading text-to-speech synthesis system that offers versions in Serbian, Croatian, Bosnian, and Montenegrin, incorporating natural intonation elements. The system delivers near-human voice quality through built-in intonation and accentuation features, significantly enhancing the naturalness of the generated speech. Additionally, it allows for adaptation to a specific speaker's voice with minimal speech data and can generate expressive speech for various applications.

AlfaNum STT [110] is an advanced continuous speech recognition system designed for Serbian, Bosnian, Croatian, and Montenegrin. Specialized language and acoustic models are employed as part of leading regional cloud-based and on-premise automatic speech recognition solutions, including commercial applications for medical and legal dictation, as well as voice assistant mobile applications.

As will be presented in more detail in the following sections, a wide range of applications of AlfaNum's speech technologies – both TTS and STT – have already been developed and deployed in Serbia or elsewhere in the region (Fig. 2).

3.1 TTS Applications

The first TTS application developed using AlfaNum TTS for people with disabilities was anReader, developed for the visually impaired [111], enabling them to use computers and smartphones equipped with screen readers such as JAWS or NVDA. To facilitate the use of anReader, it was necessary to develop a speech API. AnReader is officially recognized as an assistive tool for the visually impaired in Serbia, but its use has extended throughout the western Balkans.

AlfaNum TTS aids individuals with dyslexia by enhancing their reading speed and supports those with congenital or acquired speech disorders. Individuals with speech impairments can type their intended messages, which TTS can then vocalize. Owing to voice conversion, laryngectomized users can use speech synthesizers to replicate their own voices using only several minutes of their earlier speech recordings.

Augmentative alternative communication (AAC) aids support those with limitations in producing or comprehending spoken or written language, including conditions such as cerebral palsy, autism, and intellectual disability. AAC devices range from simple aids, such as picture boards for requesting food or assistance, to sophisticated speech-generating devices. A notable multilingual AAC application, cBoard, employs AlfaNum TTS for several South Slavic languages.

The most basic commercial applications of AlfaNum TTS are used for voice announcements in public transportation. In these applications, TTS with remote access

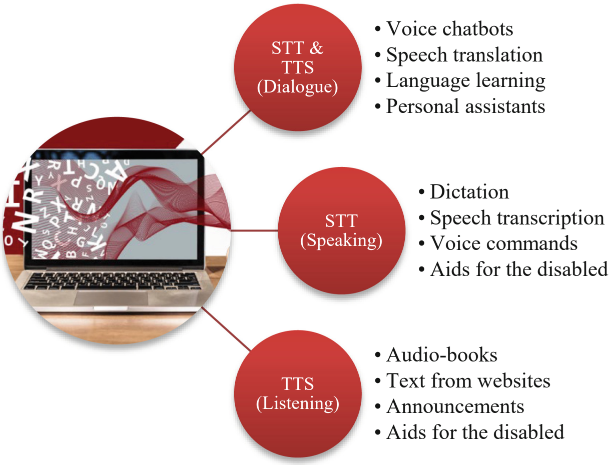


Fig. 2. TTS and STT applications developed and deployed in Serbia and neighboring countries.

(either cloud or on-premise) can be used, as well as the MS-SAPI5 interface if the main application is written for the Windows operating system.

On the other hand, the most commonly used TTS application is the web service "*Read me*", enabling users to listen to news articles in the background while performing other tasks. This feature is widely used on the websites of public media services and various government and public institutions in the region. Some media services, such as the Radio Television of Serbia, even use cloned voices of their own presenters. Implementation of such services faces additional challenges, since AlfaNum provides TTS functionality but does not have access to the internal organization of the web site, which is usually handled by another company. The common practice is to provide high-level libraries for accessing TTS (PHP, Java, Python), and implement the service in collaboration with this company.

The first audio library for the visually impaired was established at the Library of the Union of the Blind in Belgrade. It operates as a client-server system, allowing visually impaired users to access a large database of books via a local network or the internet. The system provides audible output without the need for a separate screen reader, and enables navigation through chapters, paragraphs, and bookmarks. To protect copyright, books are encrypted on the client side and can only be accessed in the audio format. Such a service is also beneficial for individuals who cannot hold books due to physical disabilities, but has become increasingly popular among those who simply prefer audiobooks. The Audio Library of the University of Novi Sad was also developed based on the same idea, and there is also fruitful collaboration with publishers of textbooks for elementary and high school education in Serbia based on TTS [112].

3.2 STT Applications

Even a small vocabulary STT system integrated into smart home technologies can significantly enhance accessibility for the disabled by allowing them to control devices such as lights and appliances through voice commands. On the other hand, large vocabulary STT systems are suitable for speech transcription and online dictation. There are many potential users of such services, including media outlets, medical or legal practitioners, government agencies etc. GPU-based computers can transcribe speech several times faster than real-time. Some of the existing STT solutions for Serbian, developed within the collaboration of the Faculty of Technical Sciences and AlfaNum are specifically tailored for on-premise users, such as medical and legal institutions.

The MEDICTA and IURISDICTA systems are advanced dictation tools designed to enhance the efficiency of medical and legal professionals by converting dictated speech into text [113]. MEDICTA is tailored for medical findings and operates in real-time on standard computers with regular microphones. It accurately interprets acronyms, punctuation, and initial capital letters, and even allows code-switching between Latin and Serbian. In contrast, IURISDICTA is specifically designed for legal document dictation, effectively recognizing legal terminology and acronyms. Both systems achieve WER below 2%, support user-defined commands, and allow for efficient manual correction of misrecognized words. They also support the use of templates to streamline the dictation of frequently repeated sections, further increasing efficiency.

TRANSCRIPTA is an advanced transcription system that converts recorded speech into text using the open-source Whisper model [107]. It generates transcripts from various audio sources, including TV shows, meetings, conferences, and court hearings. It accurately recognizes natural speech from multiple speakers (WER below 10%, CER below 5%). This level of accuracy was achieved by fine-tuning Whisper with datasets developed for Serbian and Croatian, allowing the system to transcribe Serbian with remarkable precision. TRANSCRIPTA also incorporates a diarization option that distinguishes between different speakers in the audio. Combined with time markers embedded in the transcripts, this enables quick searches through audio and video archives and enhances the efficiency of listening and manual correction of transcripts.

3.3 STT&TTS Applications

Joint applications of STT and TTS facilitate two-way human-machine communication. The development and implementation of these systems in the western Balkans are still in the early stages. AlfaNum's first personal assistant, Axon Voice Assistant, was created for mobile phones, allowing users to make calls, send messages, and perform voice dialing of contacts, addressing complex morphology of Serbian names [114]. However, adaptation to a variety of phone models and operating systems posed a significant challenge for a small company such as AlfaNum, which is why the system was never commercially released. Personal voice assistants are now being integrated not only into smartphones but also into robots, smart speakers, and smart home systems. The future of speech technology in personal assistants looks promising, as advancements in AI enable machines to not just recognize words but also to identify speakers and interpret their moods and intentions.

AlfaNum's TTS systems for Serbian, Montenegrin, Bosnian, and Croatian are currently being integrated into a mobile speech translator that supports over 60 languages. The process of aligning protocols and APIs necessary for accessing AlfaNum's STT and

TTS components is underway, as well as ensuring high throughput during peak times and minimal response latency.

Finally, following the remarkable progress of chatbots like ChatGPT, the next steps involve developing voice chatbots that use STT and TTS, alongside NLP. They will offer functionalities similar to those of call centers, as most calls will be managed automatically by chatbots, either in place of or in conjunction with a smaller number of human operators. We are currently at a stage where providers of end-to-end voice chatbot solutions are expanding into Serbia and other countries where AlfaNum offers advanced TTS and STT capabilities. Again, supporting standard APIs, high throughput and low latency is crucial for high-quality voice chatbots. For TTS, it is usually expected to have a latency of less than 0.5 s, while for STT, it can be somewhat higher since the system requires the entire user's query in order to respond, rather than just the first word. TTS is expected to support multiple speakers and styles, while STT is adaptable to a specific dictionary and language model that best suits the user's needs.

4 Paradigm Shifts in the Development and Perspectives of Speech Technology for Under-Resourced Languages

Advancements in artificial intelligence and natural language processing have profoundly influenced our interactions with technology. TTS and STT systems are among the most prominent technologies that have emerged from these advancements. Although the implementation and evolution of TTS and STT technology have been rapid for many widely spoken languages, the adoption and effectiveness of these technologies face considerable challenges when addressing under-resourced languages [2, 115]. This section examines unique challenges encountered in the deployment of TTS and STT systems for such languages, again, taking Serbian as an illustrative example.

For widely spoken languages, TTS and STT technologies have undergone extensive development and integration into various applications. These languages benefit from the existence of large and diverse datasets necessary to train high-performance DNN-based STT and TTS systems. For instance, TTS systems in widely spoken languages are tapically able to produce voices with various accents, regional dialects, and speaking styles [116], while under-resourced languages face various challenges that impede even the basic functionality of TTS [117] and STT [118] systems.

Open-source initiatives provide essential resources and tools for developing TTS and STT systems for under-resourced languages, promoting collaboration and innovation by granting open access to technology and data for the industrial and academic community [46, 107, 119]. The emergence of open-source solutions has guided local companies toward developing products for specialized domains. By leveraging efficient and adaptable open-source solutions, local stakeholders can create products tailored to specific user needs. This approach reduces costs and allows for customization but also benefits from community support, accelerating development and ensuring they remain competitive and relevant. The development of speech technology for under-resourced languages was significantly facilitated by the use of transfer learning [120]. By adapting large pre-trained models, the existing general knowledge can be leveraged and the models tuned for a specific language or its regional variant to enhance the performance of TTS and STT while reducing the need for extensive new datasets. Multilingual models are particularly useful, providing dialectical variation by training on corpora from several different languages.

Involving local experts and stakeholders in the development process enhances the accuracy and relevance of TTS and STT technology. This approach ensures that the technology is tailored to local dialects, cultural preferences, and specific user needs, leading to more effective and broadly adopted TTS and STT solutions. In the case of the Serbian language, the collaboration between the Faculty of Technical Sciences in Novi Sad and the company Alfanum resulted in the development of a diverse range of speech resources and speech technology applications for the Serbian language [121]. Initially, the production of these resources required a significant amount of manual labeling, which was labor-intensive and time-consuming. As the project advanced, the adoption of state-of-the-art technologies enabled automatic transcription, significantly reducing the need for expert supervision. This transition, combined with the emergence of publicly available tools for developing speech models, accelerated the development and improved the scalability and efficiency of creating and updating language resources, allowing more rapid adjustments and refinements, and leading to more robust and comprehensive TTS and STT applications.

5 Conclusion

The paper discussed the paradigm shift in the development of text-to-speech (TTS) and speech-to-text (STT) technologies, highlighting the transition from hidden Markov models (HMMs) to deep learning (DL) models. It also explored future perspectives on speech technology applications for under-resourced languages, offering a historical overview and addressing the specific implementation challenges encountered in developing speech technology for Serbian and kindred South Slavic languages.

The case study of Serbian illustrates not only the challenges and solutions in the development of speech technology for under-resourced languages but also the specifics of implementation and exploitation in limited markets. These topics are analyzed and compared across the HMM and DL paradigms. The shift from HMM to DL has facilitated the development of speech technology for under-resourced languages. However, achieving greater independence from global AI giants requires systematic efforts to create speech and language resources for each language. This is why Serbia has established a National Program for Language Technology Development for Serbian as part of its broader AI development strategy. The program aims to create a comprehensive framework for developing speech recognition and synthesis, natural language processing, and other linguistic technologies. It focuses on resource and application development, research and innovation, and training and education, all intended to significantly enhance the capabilities of speech and language technologies in the region while fostering economic growth as well as cultural preservation.

6 Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments. This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7449, Multimodal multilingual human-machine speech communication, AI-SPEAK, and by the Ministry of Science, Technological Development and Innovation (Contract No. 451–03-65/2024–03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01–3394/1).

References

1. Delić, V., et al.: Speech technology progress based on new machine learning paradigm. Computational Intelligensce and Neuroscience, Wiley, Article 4368036, 19 pages (2019)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: a survey. Speech Commun. **56**, 85–100 (2014)
3. Swietojanski, P., Ghoshal, A., Renals, S.: Unsupervised crosslingual knowledge transfer in DNN-based LVCSR. In: Workshop SLT, pp. 246–251. IEEE, Miami, FL, USA (2012)
4. Tan, X., Qin, T., Soong, F., Liu, T.Y.: A Survey on Neural Speech Synthesis. arXiv preprint arXiv:2106.15561 (2021)
5. Dutoit, T.: High Quality Text-To-Speech Synthesis of the French Language. Ph.D. dissertation. Supervised by Prof. Henri Leich. Faculté Polytechnique de Mons. (1993)
6. Teranishi R., Umeda N.: Use of pronouncing dictionary in speech synthesis experiments. In: Reports of the Sixth International Congress on Acoustics, vol. 2, pp. 155–158 (1968)
7. Hallahan, W.I.: DECtalk Software: text-to-speech technology and implementation. Digit. Tech. J. **7**(4), 5–19 (1995)
8. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, Boston, London (1999)
9. Van Santen, J.: Assignment of segmental duration in text-to-speech synthesis. Comput. Speech Lang. **8**(2), 95–128 (1994)
10. Sejnowski, T., Rosenberg, C.R.: Parallel networks that learn to pronounce English text. Complex Syst.1, 145–168 (1987)
11. McCulloch, N., Bedworth, M., Bridle J.: NETspeak – a re-implementation of NETtalk. Comput. Speech Lang. **2**, 289–301 (1987)
12. Ronanki, S.: Prosody Generation for Text-to-Speech Synthesis. Ph.D. thesis, University of Edinburgh (2019)
13. Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K.: ATR v-TALK speech synthesis system. In: Proceedings of International Conference on Spoken Language Processing, pp. 483–486 (1992)
14. Donovan, R.E., Eide, E.: The IBM trainable speech synthesis system. In: Proceedings of 5th International Conference on Spoken Language Processing (ICSLP 98), p. 4, ISCA, Sydney, Australia (1998)
15. Hunt A.J., Black A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP, pp. 373–376. IEEE, Atlanta, GA, USA (1996)
16. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, In: Proceedings of the 6th EUROSPEECH, pp. 2347–2350. Budapest, Hungary (1999)
17. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. **17**(s1), 66–83 (2009)

18. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: Proceedings of the 10th EUROSPEECH, pp. 2461–2464. Geneva, Switzerland (2003)
19. Qian, Y., Liang, H., Soong, F.K.: A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1231–1239 (2009)
20. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. Proc. IEEE **101**(5), 1234–1252 (2013)
21. Yan, Z.-J., Qian, Y., Soong, F.K.: Rich-context unit selection (RUS) approach to high quality TTS. In: Proceedings of ICASSP, pp. 4798–4801. IEEE (2010)
22. Qian, Y., Soong, F.K., Yan, Z.J.: A unified trajectory tiling approach to high quality speech rendering. IEEE Trans. Audio Speech Lang. Process. **21**(2), 280–290 (2013)
23. Weijters, T., Thole, J.: Speech synthesis with artificial neural networks. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1764–1769, San Francisco, CA, USA (1993)
24. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the ICASSP, pp. 7962–7966. IEEE (2013)
25. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Proceedings of 15th INTERSPEECH, pp. 1964–1968. ISCA, Singapore (2014)
26. Saito, Y., Takamichi, S., Saruwatari, H.: Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 84–96 (2018)
27. Wu, Z., King, S.: Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1255–1265 (2016)
28. Fan, Y., Qian, Y., Soong, F.K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: Proceedings of ICASSP, pp. 4475–4479. IEEE (2015)
29. Wu, Z., Swietojanski, P., Veaux, C., Renals, S., King, S.: A study of speaker adaptation for DNN-based speech synthesis. In: Proceedings of the 16th INTERSPEECH, pp. 879–883, Dresden (2015)
30. Hojo, N., Ijima, Y., Mizuno, H.: An investigation of DNN-based speech synthesis using speaker codes. In: Proceedings of the 17th INTERSPEECH 2016, pp. 2278–2282. San Francisco, USA (2016)
31. Fan, Y., Qian, Y., Soong, F.K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: Proceedings of ICASSP, pp. 4475–4479. Brisbane, Australia (2015)
32. Brave, S., Nass, C.: Emotion in human-computer interaction. In: Sears, A., Jacko, J.A. (eds.) Human-Computer Interaction Fundamentals, pp. 53–68, CRC, Boca Raton, USA (2009)
33. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: 8th EUROSPEECH, Geneva, Switzerland (2003)
34. Eyben, F., et al.: Unsupervised clustering of emotion and voice styles for expressive TTS. In: Proceedings of ICASSP, pp. 4009–4012. IEEE (2012)
35. Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y.: GMM-based emotional voice conversion using spectrum and prosody features. Am. J. Signal Process. **2**(5), 134–138 (2012)
36. Lorenzo-Trueba, J., Henter, G.E., Takaki, S., Yamagishi, J., Morino, Y., Ochiai, Y.: Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. Speech Commun. **99**, 135–143 (2018)

37. Luo, Z., Chen, J., Takiguchi, T., Ariki, Y.: Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data. In: Proceedings of the 18th INTERSPEECH, pp. 3399–3403. ISCA (2017)
38. Ming, H., Huang, D., Xie, L., Wu, J., Dong, M., Li, H.: Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. In: Proceedings of the 17th INTERSPEECH 2016, pp. 2453–2457. ISCA (2016)
39. An, S., Ling, Z., Dai, L.: Emotional statistical parametric speech synthesis using LSTM-RNNS. In: Asia-Pacific Signal and Information Processing Association Annual Samit and Conference (APSIPA ASC), pp. 1613–1616, IEEE (2017)
40. Skerry-Ryan, R., et al.: Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In: Proceedings of the 34th International Conference on Machine Learning, pp. 4693–4702. PMLR (2018)
41. Wu, P., Ling, Z., Liu, L., Jiang, Y., Wu, H., Dai, L.: End-to-end emotional speech synthesis using style tokens and semisupervised training. In: Asia-Pacific Signal and Information Processing Association Annual Samit and Conf. (APSIPA ASC), pp. 623–627. IEEE (2019)
42. Zhou, K., Sisman, B., Rana, R., Schuller, B.W., Li, H.: Speech synthesis with mixed emotions. IEEE Trans. Affect. Comput. **14**(4), 3120–3134 (2022)
43. Van den Oord, A., Dieleman, S., Zen, H., et al.: WaveNet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 12 (2016)
44. Van den Oord, A., et al.: Parallel WaveNet: fast high- fidelity speech synthesis. In: Proceedings of the 35th International Conference on Machine Learning, pp. 3915–3923. Stockholm, Sweden (2018)
45. Arik, S.O., et al.: Deep voice: real-time neural text-to-speech. In: Proceedings of the 34th International Conference on Machine Learning, pp. 195–204. PMLR, Sydney, Australia (2017)
46. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis. In: Proceedings of the 18th INTERSPEECH 2017, pp. 4006–4010. ISCA, Stockholm, Sweden (2017)
47. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In: Proceedings of ICASSP, pp. 4779–4783. Calgary, Canada (2018)
48. Ping, W., Peng, K., Gibiansky, A., et al.: Deep voice 3: scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
49. Arik, S.Ö, Chen, J., Peng, K., Ping, W., Zhou, Y.: Neural voice cloning with a few samples. In: Advances in Neural Information Processing Systems 31, 32nd Conference on Neural Information Processing Systems, pp. 10040–10050, Montreal, Canada (2018)
50. Nachmani, E., Polyak, A., Taigman, Y., Wolf, L.: Fitting new speakers based on a short untranscribed sample. In: Proceedings of the 35th International Conference on Machine Learning, pp. 3680–3688. Stockholm, Sweden (2018)
51. Akuzawa, K., Iwasawa, Y., Matsuo, Y.: Expressive speech synthesis via modeling expressions with variational autoencoder. In: Proceedings of the 19th INTERSPEECH, pp. 3067–3071. ISCA, Hyderabad, India (2018)
52. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. Adv. Neural Inf. Process. systems **32** (2019)
53. Ren, Y., et al.: Fastspeech 2: Fast and high-quality end-to-end text to speech. Preprint arXiv:2006.04558 (2020)
54. Nosek, T., Suzić, S., Sečujski, M., Stanojev, V., Pekar, D., Delić, V.: End-to-end speech synthesis for the Serbian language based on Tacotron. In: Karpov, A. Delić, V., (eds.) SPECOM 2024, LNAI Part I - 15299, Springer, Heidelberg, Belgrade, Serbia (2024)
55. Wang, C., et al.: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111 (2023)
56. Zhang, Z., et al.: Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. arXiv preprint arXiv:2303.03926 (2023)

57. Han, B., et al.: VALL-E R: Robust and Efficient Zero-Shot Text-to-Speech Synthesis via Monotonic Alignment. arXiv preprint arXiv:2406.07855 (2024)
58. Meng, L., et al.: Autoregressive Speech Synthesis without Vector Quantization. arXiv preprint arXiv:2407.08551 (2024)
59. Casanova, E., Weber, J., Shulby, C., Candido Junior, A., Gölge, E., Antonelli Ponti, M.: YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. arXiv preprint arXiv:2112.02418 (2024)
60. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint arXiv:2010.05646 (2020)
61. Prenger, R., Valle, R., Catanzaro, B.: WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv preprint arXiv:1811.00002 (2018)
62. Casanova, E., et al.: XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv preprint arXiv:2406.04904 (2024)
63. Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., Delić, V.: AlfaNum system for speech synthesis in Serbian language. In: Proceedings of the 5th International Conference Text, Speech and Dialogue (TSD 2002), pp. 237–244. Brno, Czech Republic (2002)
64. Pakoci, E., Mak, R.: HMM-based speech synthesis for the Serbian language. In: Proceedings of the 56th ETRAN, vol. TE4, pp. 1–4. Zlatibor, Serbia (2012)
65. Delić, T., Sečujski, M., Suzić, S.: A review of serbian parametric speech synthesis based on deep neural networks. TELFOR J. **9**(1), 32–37 (2017)
66. Sečujski, M., Pekar, D., Suzić, S., Smirnov, A., Nosek, T.: Speaker/style-dependent neural network speech synthesis based on speaker/style embedding. J. Univ. Comput. Sci. **26**(4), 434–453 (2020)
67. Suzić, S., Sečujski, M., Nosek, T., Delić, V., Pekar, D.: HiFi-GAN based text-to-speech synthesis in Serbian. In: Proceedings of 30th EUSIPCO, pp. 2231–2235, Belgrade, Serbia (2022)
68. Sakai, T., Doshita, S.: Phonetic Typewriter. J. Acoust. Soc. Am. **33**, 1664 (1961)
69. Davis, K.H., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. Am. **24**, 637–642 (1952)
70. Vintsyuk, T.K.: Speech discrimination by dynamic programming. Cybern. Syst. Anal. **4**, 52–57 (1972)
71. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**, 43–49 (1978)
72. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. **50**, 637–655 (1971)
73. Jelinek, F., Bahl, L., Mercer, R.: Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. Inf. Theory **21**, 250–256 (1975)
74. Klatt, D.H.: Review of the ARPA speech understanding project. J. Acoust. Soc. Am. **62**, 1345–1366 (1977)
75. Jelinek, F.: Continuous speech recognition by statistical methods. Proc. IEEE **64**, 532–556 (1976)
76. Levinson, S.E., Rabiner, L.R., Sondhi, M.M.: An Introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. Bell Syst. Tech. J. **62**, 1035–1074 (1983)
77. Juang, B.-H.: Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains. AT&T Tech. J. **64**, 1235–1249 (1985)
78. Juang, B.-H., Levinson, S., Sondhi, M.: Maximum likelihood estimation for multivariate mixture observations of Markov chains. IEEE Trans. on Inform. Theory **32**, 307–309 (1986)
79. Lee, K.-F.: Context-independent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Trans. Acoust. Speech Signal Process. **38**, 599–609 (1990)

80. Young, S.J., Woodland, P.C.: State clustering in hidden Markov model-based continuous speech recognition. Comput. Speech Lang. **8**, 369–383 (1994)
81. Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. Pattern Recogn. Artif. Intell. 374–388 (1976)
82. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**, 1738–1752 (1990)
83. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun. **25**, 133–147 (1998)
84. Prasad, N.V., Umesh, S.: Improved cepstral mean and variance normalization using Bayesian framework. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 156–161. IEEE, Olomouc, Czech Republic (2013)
85. Rehr, R., Gerkmann, T.: Cepstral noise subtraction for robust automatic speech recognition. In: Proceedings of ICASSP, pp. 375–378. IEEE, South Brisbane, Queensland, Australia (2015)
86. Hermansky, H., Morgan, N.: RASTA processing of speech. IEEE Trans. on Speech Audio Processing **2**, 578–589 (1994)
87. Bahl, L., Brown, P., De Souza, P., Mercer, R.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proceedings of ICASSP, pp. 49–52. IEEE, Tokyo, Japan (1986)
88. Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J.: MMIE training of large vocabulary recognition systems. Speech Commun. **22**, 303–314 (1997)
89. Juang, B.-H., Hou, W., Lee, C.-H.: Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. **5**, 257–265 (1997)
90. Povey, D., Woodland, P.C.: Minimum phone error and i-smoothing for improved discriminative training. In: Proceedings of ICASSP, pp. I-105-I–108. IEEE, Orlando, FL, USA (2002)
91. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, pp. 841–848. MIT Press, Cambridge, MA, USA (2001)
92. Macherey, W.: Discriminative training and acoustic modeling for automatic speech recognition. Ph.D. Thesis, Aachen Techn. Hochsch (2010)
93. Baker, J.: The DRAGON system–An overview. IEEE Trans. Acoust. Speech Signal Process. **23**, 24–29 (1975)
94. Bahl, L.R., Jelinek, F., Mercer, R.L.: A maximum likelihood approach to continuous speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 179–190 (1983)
95. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. **13**, 359–393 (1999)
96. Goodman, J.T.: A bit of progress in language modeling. Comput. Speech Lang. **15**, 403–434 (2001)
97. Lippmann, R.P.: Review of neural networks for speech recognition. Neural Comput. **1**, 1–38 (1989)
98. Bourlard, H.A., Morgan, N.: Connectionist Speech Recognition: a Hybrid Approach. Springer, US, Boston, MA (1994)
99. Mohamed, A., Dahl, G.E., Hinton, G.E.: Deep belief networks for phone recognition. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, pp. 1–9. Vancouver, BC, Canada (2009)
100. Dahl, G.E., Dong Yu, Li Deng, Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **20**, 30–42 (2012)

101. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 369–376. ACM Press, Pittsburgh, Pennsylvania (2006)
102. Maas, A., Xie, Z., Jurafsky, D., Ng, A.: Lexicon-free conversational speech recognition with neural networks. In: Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 345–354. Denver, Colorado (2015)
103. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: Proceedings of ICASSP, pp. 4945–4949. Shanghai (2016)
104. Karita, S., et al.: A comparative study on transformer vs RNN in speech applications. In: Automatic speech recognition and understanding workshop (ASRU), pp. 449–456. IEEE, SG, Singapore (2019)
105. Zhu, H., Wang, L., Cheng, G., Wang, J., Zhang, P., Yan, Y.: Wav2vec-S: semi-supervised pre-training for low-resource ASR. In: Proceedings of the 23th INTERSPEECH, pp. 4870–4874. ISCA (2022)
106. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
107. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the International Conference on Machine Learning, pp. 28492–28518 (2023)
108. Suzić, S., Ostrogonac, S., Pakoci, E., Bojanić, M.: Building a speech repository for a Serbian LVCSR system. Telfor J. **6**(2), 109–114 (2014)
109. Nosek, T., Suzić, S., Delić, V., Sečujski, M.: Cross-lingual text-to-speech with prosody embedding. In: Proceedings of IWSSIP, 5 pages (2023)
110. Pakoci, E.T., Popović, B.Z.: Recurrent neural networks and morphological features in language modeling for Serbian. In: 29th Telecommunication Forum (TELFOR), 8 pages. IEEE (2021)
111. Delić, V., Sečujski, M., Sedlar, N.V., Mišković, D., Mak, R., Bojanić, M.: How speech technologies can help people with disabilities. In: Ronzhin, A., Potapova, R., Delić, V. (eds.) 16th SPECOM 2014, LNAI, vol. 8773, pp. 243–250. Springer. Novi Sad, Serbia (2014)
112. Delić, V., et al.: Central audio-library of the university of Novi Sad. In: Proceedings of the Intelligent Distributed Computing XIII, pp. 467–476. Springer International Publishing (2020)
113. Pakoci, E., Pekar, D., Popović, B., Sečujski, M., Delić, V.: Overcoming data sparsity in automatic transcription of dictated medical findings. In: Proceedings of the 30th EUSIPCO, pp. 454–458. IEEE (2022)
114. Popović, B., Pakoci, E., Jakovljević, N., Kočiš, G., Pekar, D.: Voice assistant application for the Serbian language. In: 23rd Telecommunication Forum (TELFOR), pp. 858–861. IEEE (2015)
115. Reitmaier, T., et al: Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In Proceedings of the CHI Conference on Human Factors in Computing Systems, p. 17 (2022)
116. Mu, Z., Yang, X., Dong, Y.: Review of end-to-end speech synthesis technology based on deep learning. arXiv preprint arXiv:2104.09995 (2021)
117. Ogayo, P., Neubig, G., Black, A.W.: Building TTS systems for low resource languages under resource constraints. In: Proceedings Speech for Social Good Workshop, p. 5 (2022)
118. Jimerson, R., Liu, Z., Prud'Hommeaux, E.: An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In: Proceedings of the 61st Annual Meeting of the Association for Comp. Linguistics (Vol. 2 Short Papers), pp. 1008–1016 (2023)

119. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
120. Popović, B.Z., Pakoci, E.T., Pekar, D.J.: Transfer learning for domain and environment adaptation in Serbian ASR. Telfor Journal **12**(2), 110–115 (2020)
121. Delić, V.D., Pekar, D.J., Sečujski, M.S., Popović, B.Z., Pakoci, E.T., Suzić, S.B.: Development of speech technology for Serbian and its applications. In: Proceedings of the First Serbian International Conference on Applied Artificial Intelligence, p. 7. Kragujevac, Serbia (2022)