Check for updates

Hindawi
Computational Intelligence and Neuroscience
Volume 2019, Article ID 4368036, 19 pages
https://doi.org/10.1155/2019/4368036

Review Article

Speech Technology Progress Based on New Machine Learning Paradigm

Vlado Delić,[1] Zoran Perić,[2] Milan Sečujski,[1] Nikša Jakovljević,[1] Jelena Nikolić,[2] Dragiša Mišković,[1] Nikola Simić,[2] Siniša Suzić,[1] and Tijana Delić[1]

[1]*University of Novi Sad, Faculty of Technical Sciences, 21000 Novi Sad, Serbia*
[2]*University of Niš, Faculty of Electronic Engineering, 18000 Niš, Serbia*

Correspondence should be addressed to Nikša Jakovljević; jakovnik@uns.ac.rs

Received 22 February 2019; Revised 29 April 2019; Accepted 9 May 2019; Published 25 June 2019

Academic Editor: Paolo Gastaldo

Copyright © 2019 Vlado Delić et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Speech technologies have been developed for decades as a typical signal processing area, while the last decade has brought a huge progress based on new machine learning paradigms. Owing not only to their intrinsic complexity but also to their relation with cognitive sciences, speech technologies are now viewed as a prime example of interdisciplinary knowledge area. This review article on speech signal analysis and processing, corresponding machine learning algorithms, and applied computational intelligence aims to give an insight into several fields, covering speech production and auditory perception, cognitive aspects of speech communication and language understanding, both speech recognition and text-to-speech synthesis in more details, and consequently the main directions in development of spoken dialogue systems. Additionally, the article discusses the concepts and recent advances in speech signal compression, coding, and transmission, including cognitive speech coding. To conclude, the main intention of this article is to highlight recent achievements and challenges based on new machine learning paradigms that, over the last decade, had an immense impact in the field of speech signal processing.

1. Introduction

According to Kuhn's theory of scientific revolutions [1], the science makes progress through the revolutionary changes of prevailing scientific paradigms, where a paradigm represents a set of beliefs and values and technical and methodological procedures common to a scientific community. Paradigms define frames and models for solving scientific challenges. New solutions come with new generations who are ready to accept new truths and interdisciplinary approaches. New paradigms appear suddenly and provide new lights to a scientific problem, based on synergy of particular and specialized knowledge consolidated into a functional and coherent unity. Speech technology community investigates spoken language processing as an interdisciplinary research area (Figure 1), [2]. After a short retrospective of the main scientific paradigms based on the knowledge of speech production and auditory perception, this article presents new achievements and perspectives based on the new machine learning paradigm related to neuroscience and advanced signal processing.

The roots of speech signal processing research were closely related to the needs of speech signal digitization. The pioneering solutions were deployed during the World War II due to a need of secure communication between the Allies. The system was named SIGSALY, and it utilized pulse-code modulation (PCM) to enable the first transmission of voice using digital equipment [3]. In the next decades, the focus of researchers was on standardizing rules of digital telephony in order to provide high quality of reconstructed speech signal in the wide range of speech signal variances [4–7]. The compression paradigms regarding these systems have not changed significantly for decades. Particularly, the focus of research has slightly been moved toward improving the signal quality at the receiving end or toward reducing the required bit rate [8–13]. However, the significant

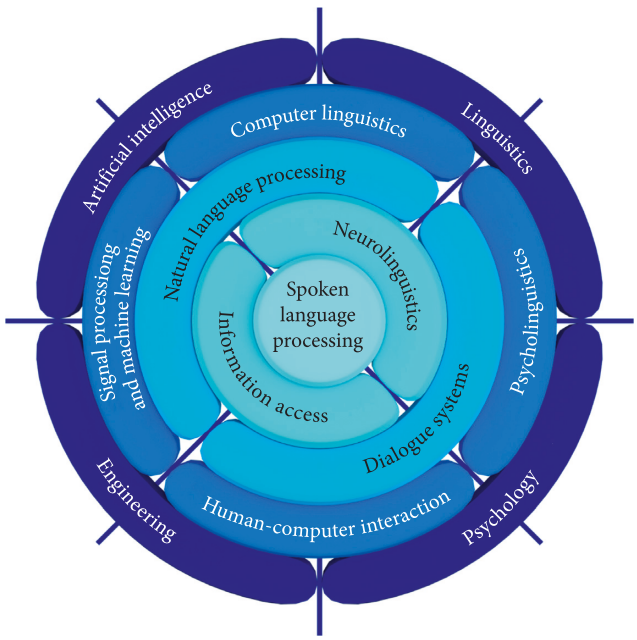

Figure 1: Interdisciplinary nature of speech technologies, i.e., spoken language processing (adopted from [2]).

development of computer technology in the last decade has enabled research into new approaches to advanced speech signal processing including adaptive machine learning methods [14]. Recent trends include cognitive speech coding so that there is a paradigm shift from perceptual (auditory) toward cognitive (auditory plus cortical) speech signal processing [15].

Modern speech technology systems rely on interdisciplinary research in the areas of multimodal signal processing and artificial intelligence, and a number of methods and algorithms have been developed with the aim of solving various problems: dialogue systems based on speech recognition and synthesis, including emotional speech, speaker identification and verification, as well as speech signal coding and transmission, denoising and detection of signals in the presence of noise, quality enhancement, and medical diagnostics based on the analysis of human voice. Recent progress in most of these speech technology topics will be discussed in more details in the following sections.

Spoken language processing (SLP) is an interdisciplinary research area that has attributes of computational intelligence. SLP lies in the intersection of linguistics, psychology, engineering, and artificial intelligence (AI) [2]. Advanced signal processing and machine learning methods are positioned in the adopted view to the interdisciplinary character of SLP, and both interconnections and intersections of different disciplines are shown and presented in a novel point of view (Figure 1). Instead of using the original term "pattern processing" in Figure 1, we have opted for the more common term "signal processing and machine learning (SP&ML)," which represents the overlap between the community of engineering and AI disciplines. With linguistic aspects included, they compose the natural language processing (NLP) field. Human-computer interaction (HCI) draws experience and methodology from the fields of engineering and psychology, and with the knowledge from linguistics included, they form a basis for the study and development of dialogue systems.

The interconnection of psycholinguistics and AI is the foundation of cognitive science or neurolinguistics. Neurolinguistics has been treated here as the neuroscience of speech. Neurolinguistics is presented in Figure 1 as dominantly linguistics discipline but connected to AI through computer linguistics which is on the intersection between AI and linguistics from one side and also connected to psychology through psycholinguistics, which is on the intersection between psychology and linguistics from other side. Neurolinguistics is on the opposite side from the engineering point of view. The neuroscience of speech can also be considered as an area of cognitive science, and cognition is inherent part of both speech perception (in the phase of understanding, the meaning of the message conveyed by spoken language) and speech production (in the phase of composing, a message intended to convey a certain meaning). Finally, SLP combines knowledge from the interdisciplinary areas of SP&ML, HCI, psycholinguistics, and computer linguistics, or more precisely NLP, cognitive sciences, dialogue systems, and information access.

Speech technologies are based on speech signal processing that spans a wide range of topics, while the focus in this review article is on three areas where the authors have the most expertise:

(i) Fundamental topics (speech analysis and synthesis, sound waves and speech features, speech production, auditory perception, and cognition including the linguistic aspect)

(ii) Dialogue systems based on speech recognition and text-to-speech synthesis (emotional speech recognition and text-to-speech synthesis including voice and style conversion)

(iii) Speech coding, compression, and transmission

Speech technology fields within the scope of the paper are presented in Figure 2 as a unified framework that encompasses covered topics, showing their complementarity, ranges and borders, interconnections, and intersections in the interdisciplinary area of SLP.

A brief retrospective and some perspectives of the speech technology fields shown in Figure 2 are presented in the following sections. Fundamental topics are shown in the middle of Figure 2 and presented in Section 2, covering speech production and perception analysis, including cognitive and linguistic point of views. More details related to the progress in speech recognition and speech synthesis, as well as their contributions to a new generation of human-machine speech dialogue systems, are presented in Section 3. Finally, the progress in speech signal compression, coding, and transmission is presented in Section 4, including contributions of the authors to the area. Most of these advancements are based on the new deep learning paradigm and our better understanding of neuroscience

PGR2026-00033
Krisp Technologies EX1010 Page 2

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

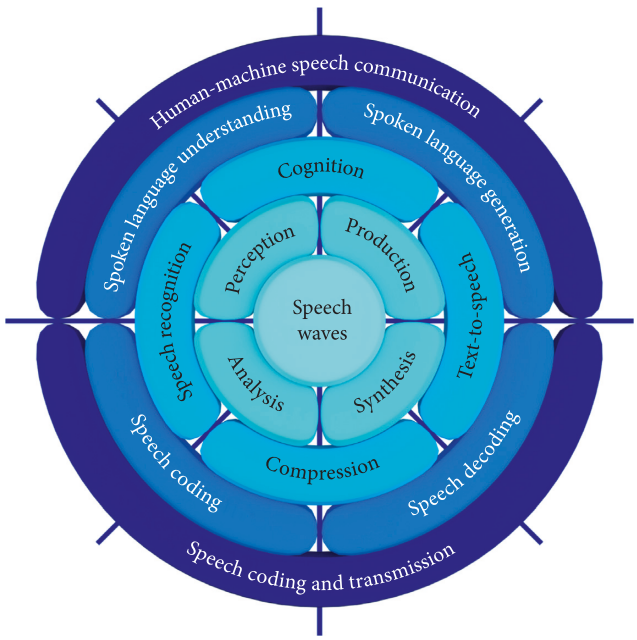

Figure 2: Unified framework that encompasses speech signal processing fields in the scope of the article.

and modelling of cognitive aspects of spoken language communication.

2. Progress in Speech Analysis and Knowledge of Spoken Language Nature

Knowledge related to the nature of spoken language is essential for efficient coding and transmission as well as satisfactory real-time human-machine speech interaction. Speech models based on either speech production or auditory perception were inherent parts of most successful algorithms. Most recent neuro-inspired computational models are based on knowledge of cognitive speech processing models [16]. After a brief review of sound pressure waves and speech signal features, speech production and auditory perception including cognitive and linguistic points of view will be elaborated in more detail in the following subsections.

2.1. Sound Pressure Waves and Speech Signal Features. Sound propagates as a continuum of acoustic waves (sound pressure), and, once received, it can be recorded, digitized, coded, transmitted, processed, and reproduced. In case of speech sounds, frequencies relevant for recognizing what was said and who has said it are located mostly below 4 kHz and hardly ever above 7-8 kHz, which is just a portion of the entire frequency range of the human auditory sense [17]. This fact was the basis of the design of analogue telephone communication systems, including the choice of microphones used. For that reason, a speech signal is sampled at 8 kHz (for a basic level of quality) or 16 kHz (if a higher level of quality is desired). It is also well known that the dynamic range from the softest to the loudest sounds in average human speech is approximately 40 dB. Even if whisper and elevated voice are included, this dynamic range is rarely above 50 dB [14]. For these reasons, the requirements for a microphone needed to record voice are typically less strict than in case of recording, e.g., music. As to quantization, it is known that each bit contributes to the signal-to-noise ratio (SNR) by 6 dB, which means that the quantization noise is practically inaudible if 8 bits are used for coding every sound sample [4]. Thus, the typical case of using the sampling rate of 8 kHz and 8 bits per sample produces the bit rate of 64 kbits/s. A lot of effort has been invested to reduce this rate without significant loss of quality at the receiving side, and Section 4 is devoted to this subject.

Digitalization including quantization is the basis of all digital speech processing techniques. If the aim is to represent speech compactly and robustly, as is the case in automatic speech recognition or most types of speech coding for transmission, one of the basic questions is the selection of relevant features that will enable fast, accurate, and robust recognition of speech (or the speaker, language, or even emotion), and/or fast and efficient speech coding for transmission without significant loss of quality. Linear predictive coding (LPC) and LPC analysis have fundamental significance in speech signal modelling and speech feature estimation [18]. Many speech coding schemes are based on LPC including Low Delay-Code Excited Linear Prediction (LD-CELP) coding scheme defined by G.728 standard, Conjugate Structure Algebraic Code-Excited Linear Prediction (CS-ACELP) coding scheme defined by G.729A standard, Algebraic Code-Excited Linear Prediction coding scheme defined by G.723.1 standard, and Adaptive Multi-Rate Wide-Band (AMR-WB) coding scheme defined by G.722.2, standards which are used in today mobile voice communication and VoIP [5, 7].

One of speech production models is also based on LPC analysis and provides speech feature sets describing speech spectrum, which is most important for speech recognition [19]. The main scope of speech signal and data processing in real time (or limited time) is to reduce the amount of data (speech features), while providing high quality of representation of such a reduced signal, that is, data source. The realization of this goal is supported by statistical signal and data processing as well as methods and algorithms which deal with signal and data reduction [20]. The most efficient methods and algorithms incorporate adaptation, and these topics will be elaborated in more details in the next sections.

2.2. Speech Production and Auditory Perception. Figure 3 shows a block diagram of both speech production and perception. Text-to-speech synthesis (TTS) and automatic speech recognition (ASR) are shown in parallel as corresponding processes performed by machines. Speech and language are learned, while the sense of hearing is innate. There are a lot of differences among human and machine speech production and perception, but the increase in the ability of machine learning paradigms to simulate human speech production mechanism, as well as auditory perception and cognition abilities, will inevitably bring about an increase in the accuracy of ASR and naturalness of TTS.

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

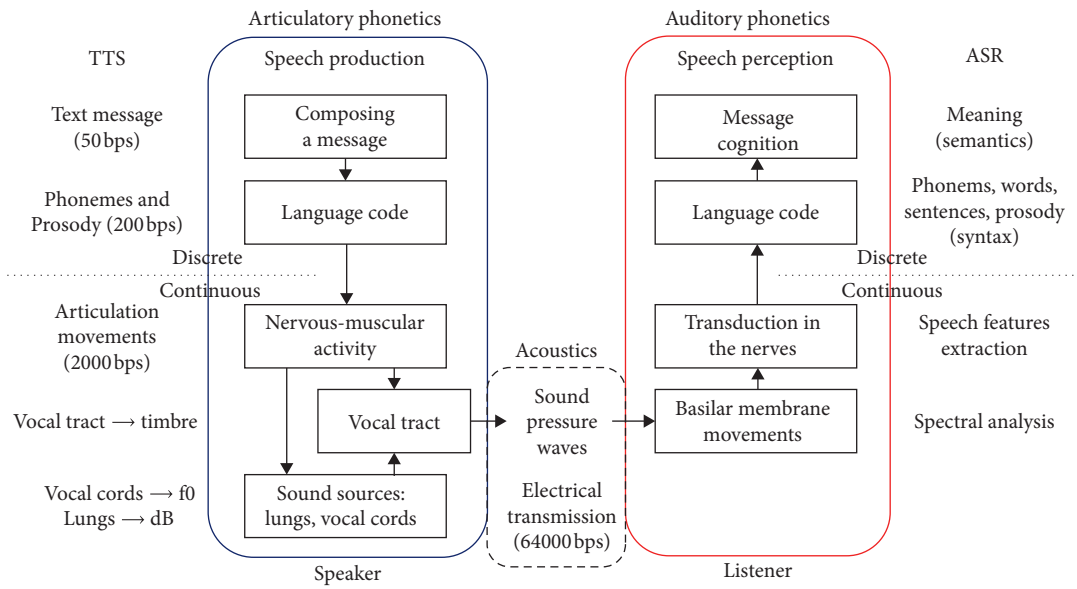

Figure 3: Block diagram of speech production and speech perception and corresponding processes performed by machines carrying out text-to-speech synthesis (TTS) and automatic speech recognition (ASR).

Speech communication between humans begins and ends at the cognitive level of message composition and interpretation. Taking into account the average speech rate of 10–12 phones per seconds and the number of phones in a language, which typically corresponds to 5 or 6 bits needed to encode them, a speech message conveyed as text could be considered to correspond to a bit rate of 50–60 bits per second. The speaker plans not only what to say but also how to say it—(s)he controls the volume, speech rate, and intonation (prosody)—any of which can carry linguistic, and also paralinguistic and extralinguistic information [21]. With that information added, the bit rate can be considered to increase to several hundreds of bits per second.

Once the speaker decides what to say and how to say it, an appropriate sound wave is produced through nervous and muscular activity [22, 23]. In that, phones are not pronounced in isolation, but the articulatory targets required for corresponding phonemes are rarely reached, leading to the coarticulation effect, which aggravates the task of ASR. Most often, the entire speech apparatus is considered through the source-filter model, where the activity of vocal folds defines the excitation and the remainder of vocal tract acts as a filter and shapes the sound spectrum [19, 21]. Besides being dependent on the phone, the acoustic features of the speech signal at a particular moment also carry information relevant to the speaker and thus represent a biometrical feature which can reveal the speaker's identity [24] and possibly other factors related to the speaker or to the message. Including the influence of speaker variability, the bit rate at this level increases to several thousand bits per second. This segment of speech communication is studied by articulatory and acoustical phonetics, and its machine counterpart is TTS, namely, the module charged with the production of the artificial speech signal itself.

Distribution of speech sample amplitudes is nonuniform, and this knowledge is used in nonuniform speech signal coding defined by μ-law and A-law [25], while some new research results provide better solutions based on adaptive algorithms. The speech production mechanism articulates a series of phonemes nonuniformly, according to an empirical statistical law formulated by George Kingsley Zipf, a linguist [26], referring to the principle of the least effort from evolutionary biology field: interlocutors try to understand each other using phonemes and words that are easier for production and perception in a particular context. The knowledge of phoneme and word statistics has been introduced into ASR algorithms long ago, and stochastic speech models like Hidden Markov model (HMM) [27] were the prevailing scientific paradigm and represented the state of the art in speech recognition and synthesis community for decades.

On the other side, the continuum of acoustic waves reaches the ear of the listener and certain frequencies excite the eardrum, and over the malleus, incus, and stapes, they excite the cochlea, where spectral analysis is performed, based on the movement of the basilar membrane, whose length is about 35 mm [17, 22, 23, 25, 28]. The hair cells in the cochlea respond to different sounds based on their frequency so that high-pitched sounds stimulate the hair cells in the lower part of the cochlea, while low-pitched sounds stimulate the upper part of the cochlea [28]. Thus formed neural impulses are sent to the central auditory system in the brain [22], and based on spectral differences, the brain recognizes relevant acoustic differences and attempts to recover the string of phones that the original message was composed of, taking into account its language model (at the level of morphology, syntax, semantics, and pragmatics). It can thus be considered that the task of ASR is to reduce the bit rate of, e.g., 64 kbits/s (digitized speech) to a bit rate of 50–60 bits/s (plain text), which would correspond to the textual contents of the message without speech prosody.

However, speech perception, which principally relies on the sense of hearing, is a nonlinear process. As is the case

PGR2026-00033
Krisp Technologies EX1010 Page 4

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with other human senses (vision, taste, touch, and smell), auditory perception of both sound pressure level (SPL) and fundamental frequency (f0, pitch) follows the Weber–Fechner law [28] from psychophysics: a change perceived as linear corresponds to an exponential change in the physical stimulus. Apart from SPL and pitch, perception of sound is affected by the distribution of sound energy across frequencies, i.e., the spectrum of the sound, which usually represents a mixture of a sequence of discrete frequency components (timbre), as in the case of periodic sounds, and a continuous mix of nonharmonic or random frequency components, as in the case of various types of noise [22, 28]. This is why common speech features like cepstral coefficients are considered to be located at frequencies rescaled from Hz to mel-scale–MFCC; they are estimated by cepstral analysis from speech frames of 20–30 ms together with their first and second derivatives calculated from several successive frames [29].

Auditory scene analysis is the process by which the auditory system separates individual sounds in natural-world situations [30, 31]. Regardless of whether sound is received by a human ear or a microphone, the incident sound pressure wave represents a sum of pressure waves coming from different individual sources, which can be either human voices or any other sound sources. These sounds usually overlap in both time and frequency. Nevertheless, the human auditory system is usually able to concentrate on an individual sound source at a time [23, 31]. While listening and separating one source, the listener constructs a separate mental description for that source. Although he/she cannot actively listen to two sound sources simultaneously, he/she can switch immediately his/her attention from one to the other [30]. For example, if a student listens to the teacher, he ignores the noise from LCD projector and a colleague who may be speaking to him; if he switches the focus to his colleague, he cannot actively listen to the teacher anymore. Furthermore, if a human listener follows the context, he/she is able to reconstruct some phonemes or entire words that he/she may not be able to hear for some reason. Humans are as successful in sound separation as they are more experienced in real-word situations and they always analyse the incoming signal using heuristic processes. As the ultimate step of the hearing process, human auditory cortex constructs a cognitive representation of the received sound wave. Without the cognition step, sound waves coming to the ears are not perceived. Heuristic analysis is based on (ir)regularities in the sum of underlying sounds.

Individual sounds differentiate from each other in at least one of the following dimensions: time, space, and frequency spectrum [28, 31]. Temporal and spatial sensations in the human auditory system are presented in more details in [32]. In the time dimension, two sounds can have some onset/offset asynchrony. In a specific environment, binaural hearing enables the localization of sound sources, which is easier, but also often more important, in the horizontal plane where human ears are positioned than in the vertical plane. The spectrum of frequency components can determine the perceived pitch, timbre, loudness, and the difference in the spectra of sounds received by both ears enables the localization of sound sources [23, 31, 32]. Pitch is related to the fundamental frequency f0 in periodic sound waves such as musical tones or vowels in speech; their spectrum consists of f0 and its harmonics. Temporal variation of f0 results in melody in music and intonation in speech. Timbre represents a specific distribution in the intensities of f0s and its harmonics in the spectrum. Two renditions of the same tone from two different musical instruments, having the same f0, will have different timbres due to the difference in the relative intensities of particular harmonics (the spectral envelope), and as a result, they will sound different [22]. If a sound spectrum does not contain just harmonic tones (f0s and their harmonics), the spectrum is not discrete; sound spectrum is rich with frequency components in parts or in the entire frequency range of the human auditory sense. Such sounds, with a spectrum that is more or less continuous, are much more frequent in nature (e.g., noise of a car or a machine or any transient noise). Magnitudes of spectral components contribute to the loudness; sound pressure level is defined in dB relative to the threshold of hearing at 1 kHz (20 μPa) and has range 0–120 dB to the threshold of pain [17, 22]. To conclude, two sounds can be separated from each other in an auditory scene analysis according to the differences in loudness, pitch (f0, if present), and timbre or spectrum as a whole, as well as in their temporal and/or spatial variations that can create a variety of sound impressions.

Acoustic signals are received by a listener and transformed into linguistic and nonlinguistic categories, but it is not known exactly how. There is ongoing research on neurophysiology of speech communication using the latest advances in invasive and noninvasive human recording techniques, with the aim to uncover fundamental characteristics of cortical speech processing [16]. The research team in question has studied phonetic feature encoding and mechanisms of noise robust representation of speech in auditory cortex based on the evidence that humans and animals can reliably perceive behaviourally relevant sounds in noisy and reverberant environments.

Neuro-inspired computational models try to provide progress in artificial deep neural network (DNN) performance, based on better understanding of the representation and transformation performed by these models. A case study in ASR given in [33] attempts to identify the mechanisms that normalize the natural variability of speech and compares these mechanisms with findings of speech representation in the human auditory cortex. The aim is to compare DNNs with their biological counterparts, identify their limitations, and reduce the performance gap between biological systems and artificial computing. For example, a human is able to concentrate on one speaker voice and ignore other sounds and voices [23, 31], based on their differences in spatial positions, pitch, and timbre, coherence of changes in level and/or frequency, and time characteristics (onset/offset asynchrony) [30]. An algorithm aimed at focusing on one speaker in a group of many speakers based on deep attractor network is proposed in [34], based on similar principles. It has been shown that switching attention

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to a new speaker instantly changes the neural representation of sound in the brain. An adaptive system should change the sensory representation in real time to implement novel, task-driven computations that facilitate the extraction of relevant acoustic parameters.

Human listeners have a remarkable ability to understand quickly and efficiently the world around them based on behaviour of known sound sources. Moreover, they are able to pay attention and focus on the meaning of speech of a particular speaker. Attentional focus can be integrated into HCI dialogue strategy [35], while data related to human cognitive effort can be used in postprocessing and improvement of the performance of ASR systems [36]. Humans are able not only to separate one speaker or concentrate only to one sound source but also to group more sound sources and hear, e.g., the entire orchestra as one musical sound based on harmonicity and synchrony of particular sound sources. Concurrent and sequential grouping processes are described in more details in [37].

The role of the nonlinearities in DNN in categorization of phonemes by their nonuniform and nonlinear warping of the acoustic space are studied in [38], as well as the way perceptual invariant categories are created. Biological neurons are able to dynamically change the synaptic efficacy in response to variable input conditions. It is called synaptic depression and when it is added to the hidden layers of a DNN trained for phoneme classification, ASR system becomes more robust to noisy conditions without explicitly being trained for them. The results from [39] suggest that more complete neuron models may further reduce the gap between the biological performance and artificial computing, resulting in networks that better generalize to novel signal conditions.

2.3. Engineering vs. Linguistic Point of View to NLP as a Typical AI Topic. The mechanism of speech production and the physical component of sound perception are relatively well-studied topics [22, 31], while cognitive aspects of speech communication still represent a widely open research area. All aspects of human-machine speech communication that are related to linguistics, such as natural language processing (NLP), cognitive sciences—neurolinguistics, and dialogue management (see Figure 1), represent great challenges to the scientific community. In the recent past, the development of speech technology and spoken dialogue systems has gained most momentum from the engineering disciplines, through the possibility of automatic learning from vast quantities of data, in terms of development of computational facilities, complex learning algorithms, and sophisticated neural model architectures addressing particular phenomena and problems of cognitive linguistics. At the same time, cognitive speech sciences mostly remain outside of the scope of the immediate interest of engineering disciplines relevant to speech technology development. Nevertheless, the knowledge in these areas overlaps in the concept and scope with machine learning, which, inspired by neurosciences, has brought about progress not only in human-computer interaction and computational linguistics but also in the area of spoken language processing, which lies in their intersection. This is indicated in Figure 1, which also shows a relatively wide gap between cognitive sciences (neuroscience) and psycholinguistics on one side and predominantly engineering disciplines on the other.

As regards the role of machine learning in the development of speech technology, it has offered a powerful alternative to models dependent on linguistic resources and modules performing particular linguistically motivated subtasks. Linguistic resources such as dictionaries and speech databases are typically quite expensive and time-consuming to collect and annotate, while the development of modules that compose a speech technology system requires deep domain knowledge and expert effort. In the last two decades, some of the tasks performed by rule-based systems or simpler machine learning methods have, one by one, been overtaken by neural networks. Namely, in the case of acoustic speech recognition, neural networks have been shown to outperform hidden Markov models (HMMs) in acoustic modelling [40] but have also outperformed classical *N*-gram language models in terms of generalization, using either architectures based on long short-term memory (LSTM) neurons [41] or recurrent neural networks (RNN) [42]. Solutions based on neural networks have been shown to reach human parity in tasks as complex as casual conversational speech recognition [43]. In combination with a range of data-synthesis techniques for obtaining large quantities of varied data for training, it is now possible to obtain an end-to-end ASR capable of outperforming state-of-the-art pipelines in recognizing clear conversational speech as well as noisy one [44, 45]. They have also been used in multimodal speech recognition, i.e., recognition of speech from audio and video [46]. The task of speech synthesis is a more language-dependent one, and in that it is more challenging since it aims to reintroduce the redundancy which is lost when speech is converted into text, and to do it in such a way that, among a multitude of prosodic renditions of a particular utterance, it produces one that the listener will consider acceptable in a given context. Here again, neural networks have shown to overperform classical models working on parameterized speech such as HMMs [47, 48] in acoustic modelling, and they have also been employed for prosody modelling [49] as well as modelling of acoustic trajectories [50]. Neural networks have also addressed the problem of a somewhat muffled character of synthesized speech due to the use of a vocoder, by performing synthesis of raw speech waveforms instead [51]. Finally, to overcome the need for sophisticated speech and language resources that require deep domain expertise, a range of end-to-end architectures were proposed, with the ultimate end that the system should be trained on pairs of text and audio, exploiting the capability of neural networks to automatically develop higher-level abstractions [52]. The flexibility of such a powerful data-driven approach in comparison with classical speech concatenation synthesizers has also brought significant progress in the area of multispeaker TTS and speaker adaptation [53–55] as well the ability to conform to a particular speech style or emotion [56]. This is particularly relevant as it coincides with the emergence of applications

PGR2026-00033
Krisp Technologies EX1010 Page 6

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

such as smart environments, virtual assistants, and intelligent robots, demanding high-quality speech synthesis in different voices and different styles and conveying different emotional states of the perceived speaker [57]. Other language technology tasks have also been successfully addressed by neural networks, such as question answering [58], text classification [59, 60], machine translation [61, 62], and sentiment analysis [63]. Neural networks have also been used as a powerful linguistic tool, for modelling sentence syntax [64] or exploring particular linguistic phenomena such as establishing word representations in vector spaces [65]. However, rather than providing a decomposition of the problem and a clear analytical insight into it, neural networks provide an alternative, data-driven point of view, and thus cannot be considered a classical tool of theoretical linguistics. On the other hand, their performance in solving these problems justly makes neural networks state of the art in the development of speech technology.

3. Progress in Speech Recognition and Synthesis, as well as Dialogue Systems

Apart from automatic speech recognition (ASR) and text-to-speech synthesis (TTS), a human-machine speech dialogue system also includes a dialogue management module with corresponding dialogue strategies and language technologies for spoken language understanding (SLU) and spoken language generation (SLG), as illustrated in Figure 4.

This section presents some achievements in the field of speech technologies such as ASR and TTS. They have been developed with an effort to combine interdisciplinary knowledge from different areas such as linguistics, acoustics, computer science, and mathematics. Signal processing engineers usually have integrating roles among linguists from one side and mathematicians from the other side.

3.1. Progress of Automatic Speech Recognition Systems. Research and development of ASR systems began in the 1950s in Bell Labs, with simple digit recognition systems, and since then the recognition tasks have become more complex—from the recognition of isolated digits, then isolated words, then continuously spoken words in a silent environment, up to the recognition of spontaneous speech in a noisy environment. Consequently, the complexity of the algorithms used also increased drastically. A brief review of historical development of ASR can be found in [66]. There were three important moments in the development of ASR systems: introduction of mel-frequency cepstral coefficients [67], introduction of statistical methods (hidden Markov models (HMM) with Gaussian mixture models (GMM)) [68], and introduction of deep neural networks (DNN) [69]. This development was also supported by the technological development in the computer industry as well as the increase in the amount of data available for training these systems.

The domination of DNNs in ASR started with [40], which showed that feedforward DNN outperforms GMM in the task of estimation of context-dependent HMM state emitting probabilities. For a small database, such as English

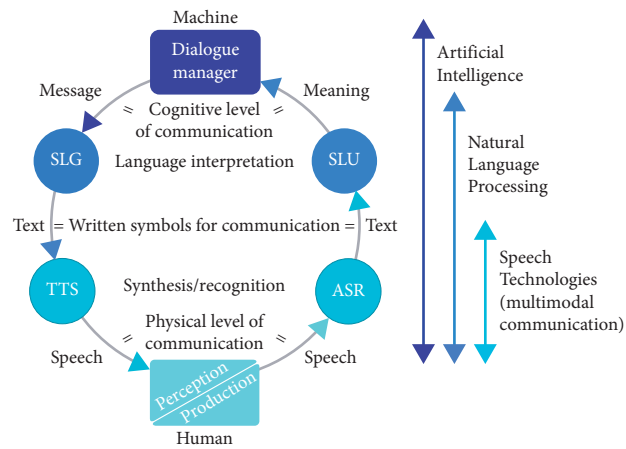

Figure 4: Components of a human-machine speech dialogue system.

Broadcast News (about 30 h of training data), the difference in word error rates (WER) was not significant, but for the Switchboard database, which is bigger (about 300 h of training data), the difference became substantial. Further improvement of DNN was based on better optimization, new activation functions, new network architectures, new speech preprocessing methods, and leveraging multiple languages and dialects [70]. One of the important findings was that layer-by-layer pretraining using restricted Boltzmann machines (RBM) is not obligatory and that back-propagation algorithm is sufficient for training in case of a large quantity of available training data as well as a large number of units in the hidden layers. Additionally, LeCun et al. showed in [71] that in case of sufficiently wide DNN (large number of units in a layer), there is no problem with the local minima and that the values of local minima are very close. The next big step was a complete elimination of HMM from the model. Graves and Jaitly in [72] reported a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. The system is based on a combination of the deep bidirectional long-short term memory (LSTM) recurrent neural network architecture and the connectionist temporal classification (CTC) objective function. Such a direct mapping of an audio signal into a grapheme sequence allows easy application of the system on new languages such as Serbian [73]. Inspired by CTC, Povey at al. in [74] developed lattice-free maximum mutual information using phone *n*-gram language model starting from randomly initialized neural networks. This method was also successfully applied to Serbian [75]; i.e., the relative reduction of WER was about 25% with respect to the best previous system.

3.2. Progress of Speech Emotion Recognition. Since humans are not always rational and logical beings—emotions play very important aspects in acceptance of new products and technologies [76]. The earliest attempts to recognize speaker emotional state on the basis of voice characteristics date back to the 1980s [77]. The initial motive for this research direction was the adaptation of an ASR system to emotionally

PGR2026-00033
Krisp Technologies EX1010 Page 7

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

stressed speech [78], but another motive appeared with the development of spoken dialogue systems, where it was useful to modify the dialogue strategy based on, e.g., user annoyance [79]. There are a number of emotions that can be easily represented in the activation-evaluation space [80], but classification of such a large number of emotions is difficult. Hence, classification space has been reduced to neutral and 6 archetypal emotions: anger, disgust, fear, joy, sadness, and surprise, which are the most obvious and distinct emotions [80]. It should be noted that archetypal emotions are not primary emotions in so-called "pallet theory," where each emotion can be represented as a combination of the primary ones.

One of the important steps in the design of a speech emotion recognition system is the extraction of features that efficiently discriminate between emotions independently of lexical content, speaker, and acoustic environment. It is well known that prosodic features are correlated with emotions [80], which is why standard features used in emotion recognition systems include pitch, energy, and phone duration [81]. These features are also related to the voice quality that is related to the emotions [82]. Emotions affect speech energy distribution across a wide range of frequencies, thus spectral features such as MFCCs, linear prediction cepstral coefficients, log frequency power coefficients, and formants were further proposed [83, 84]. Feature extraction procedure starts with the segmentation of the input signals into 20–30 ms long frames shifted by 10 ms, since speech is a nonstationary signal. After that, the features extracted from a chosen segment (corresponding to a particular phoneme, syllable, word, or sentence) are mapped into a single vector using functionals such as mean, second moment, contour slope, and range. Hereafter, features "condensed" in such a way represent the input of standard classification algorithms such as linear Bayes [85], k nearest neighbours [85, 86], support vector machines [87], GMM [86], and artificial neural networks [88]. On the other hand, such frame-based features can also be classified as a sequence using HMM [84] and RNN [89]. Besides low-level acoustic features, individual words or sequences of words obtained by ASR can also be used to perform emotion classification [90]. After a huge success of convolutional neural networks (CNN) in image classification, where lower layers of the network perform feature extraction, some research groups have tried to implement CNN in the same manner to obtain features [91, 92]. Since speech emotion recognition is a scarce data problem, one of the future trends will be the application of semisupervised learning [93]. More details about features, classification algorithms, and databases can be found in [94–97].

3.3. Progress in the Development of Text-to-Speech Synthesis. The very first "speech machines" were mechanical devices capable of producing single phonemes, and some of their combinations were introduced by Christian Kratzenstein and Wolfgang von Kempelen at the end of eighteenth century [98]. The VODER, presented in 1939 by Homer Dudley, can be considered as the first synthesizer which could generate whole sentences [99]. The first full TTS system for English was introduced in 1968 by Teranishi and Umeda [100]. It was an articulatory-based system which could perform text analysis and determine pauses in text using a sophisticated parser [101].

However, it was not until concatenative synthesizers were invented, that TTS gained widespread usage. The idea of concatenative TTS is to concatenate appropriate parts of a prerecorded database [102]. If the goal is domain-specific synthesis or a very large speech database is available, this approach can produce high-quality speech. However, there are audible glitches at the concatenation points if the appropriate units cannot be found in the database. The method is also extremely inflexible in terms of changing the speaking style or the voice of the speaker; it requires a whole new database to be recorded and annotated.

As applications of TTS became more popular and more widely used, the necessity of algorithms that could produce different voices and speaking styles from smaller databases has grown. From around 2000, statistical parametric speech synthesis, where the spectrum, fundamental frequency, and duration of speech were modelled by multispace probability distribution HMMs and multidimensional Gaussian distributions [103], became popular. The HMM synthesis enables transformation of speaker-independent system toward a target speaker using a very small amount of speech data [104], creating expressive voices [105], as well as multilingual voices [106]. However, this method never achieved the naturalness of concatenative TTS. One of the main problems is the signal smoothness caused by modelling similar contexts with the same Gaussian mixtures. Another big problem introduced with parametric methods is the usage of vocoder, a system that produces speech waveforms from predicted acoustic features. Vocoders, although significantly improved over the time, introduce some artefacts which affect the overall quality of generated speech. A detailed review of HMM-based speech synthesis can be found in [107].

The first attempts to use neural networks in speech synthesis can be found in [108]. However, the recent development of hardware, especially graphical processing units (GPUs), has popularized this approach and established its dominant status in the TTS research society. Deep neural networks (DNN) replaced decision trees and Gaussian mixture models in mapping input linguistic features to output acoustic features, enabling nonlinear mappings [109]. Although simple feedforward NN with several hidden layers and sigmoid or tangent hyperbolic activations are sufficient for the production of intelligible and natural sounding synthetic speech, introduction of LSTM (long short-term memory) units has brought further improvement into the quality of synthesized speech [110]. Some improvements were also reported by introducing generative adversarial networks [111] and stacked bottleneck features [112].

DNNs have not only just enabled generating synthetic speech of high quality but also introduced many possibilities for production of speech in different voices and speaking styles. A majority of methods for creating new DNN voice using limited amount of training data is based on usage of multispeaker models. In multispeaker modelling, a large

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

database consisting of multiple speakers is required. Each speaker is usually represented with less data than in case of single-speaker modelling. Due to a variety of contextual information and better network generalization, the quality of speech produced with multispeaker models is similar or even better than speech obtained with single speaker models. Speaker identity in multispeaker systems can be represented in several ways. One group of approaches is based on the use of a unique vector for each speaker. This vector can be represented as *i* vector [113] or just one-hot vector [54] and is used in training as extension of standard input or additional input to any of the hidden layers. Another group of methods for representing speaker identity is based on splitting network to speaker-specific and shared parts. In [53], separate output layers for each speaker have been introduced. In [114], even language-dependent parts of the network have been added, but this approach requires data from the same speaker in multiple languages. Creating a new voice, whose samples have not been seen in the training phase, in a multispeaker framework is based on adapting only the speaker-dependent part of the network [53], estimating the speaker-specific vector for the new speaker [55] or adjusting the parameters of neurons in starting models [113]. As opposed to the usage of multispeaker models as starting models for adaptation, in [115], adaptation starting from a single speaker model is investigated. It has been shown that only ~10 min of target speaker voice is required in order to produce synthetic speech in target speaker's voice reaching the quality of conventional methods built on several hours of speech database. The hypothesis was that the models of speakers A and B are more similar than a randomly initialized model and the model of speaker B, consequently requiring less data to train the model of speaker B starting from the model of speaker A than starting from a randomly initialized model.

Synthetic speech should convey not only just information but also paralinguistic information such as emotional state. There is also a need to support some task-specific speech styles such as news, commercials, storytelling, and warnings [116]. It has been shown that emotion, mood, and sentiment affect attention, memory, performance, judgement, and decision-making in humans [117], which supports the necessity of using different speaking styles in synthetic speech for many applications. Three different methods for style modelling are compared in [118]. The presented methods are based on ideas introduced in multispeaker modelling using input codes, network adaptation, and separated output layers. It has been shown that only ~5 min of speech per style is sufficient in order to produce speech of acceptable quality in a specific style. Using input codes for representing different styles is also presented in [119, 120]. There have also been attempts at style transplantation, i.e., producing speech in the voice of speaker A in style X without having any sentence from speaker A in style X in the training data, in which case the network is forced to learn the style X from other speakers in the training database [121, 122].

Although DNNs have shown to be extremely powerful and flexible, for a long time, one of their main disadvantages in speech synthesis has been their dependence on the usage of a vocoder. For the first time in 2016, raw audio samples were directly predicted by DNN using WaveNet architecture [51]. This model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones. When conditioned on linguistic features derived from text and speaker identity, it can be used as TTS and it significantly outperforms all other TTS systems. The main drawbacks of this system are its need for extremely large databases and extreme computational power, although the synthesis has since been accelerated by the introduction of approaches such as Parallel WaveNet [123]. A similar model called DeepVoice was introduced in 2017 [124]. In DeepVoice, every part of TTS pipeline is replaced by a corresponding DNN. Its main drawback is the fact that all components of TTS system are trained independently, and it leads to a cumulative error in synthesized speech in the end.

As opposed to WaveNet and DeepVoice systems, which use lexical features as inputs, there are systems which use raw orthographic text as input, such as Tacotron [52], Tacotron 2 [125], and Deep Voice 3 [126]. Tacotron outputs spectrograms that are transformed to speech samples using Griffin–Lim algorithm, which also introduces artefacts in generated speech. On the other hand, the Tacotron 2 system-generated spectrograms are used for conditioning standard WaveNet architecture, which generates speech samples. DeepVoice 3 architecture 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 [127, 128] as well as synthesis in different styles [129, 130].

The main advantage of an integrated end-to-end TTS system is that requires minimal human effort since there is no need to label input data. Since in end-to-end systems, direct sample values are often predicted [29], the usage of 16 bit samples would make the prediction complicated and some type of quantization is performed. For this reason, improved coding and compression algorithms are important for TTS.

3.4. Dialogue Systems. Automatic speech recognition and speech synthesis are technologies with a long history. During the last five decades, a wide spectrum of algorithms shaped our knowledge within the speech technology field. With the recent advances in the world of deep learning and artificial neural networks, we are able to imitate to some extent the human auditory system sensitivity, recognition accuracy, human voice intelligibility and naturalness, emotions in synthesized speech, etc. As the result, machines are able to identify particular speakers, recognize human words in a noisy environment or to perform large-vocabulary continuous speech recognition with high accuracy. Furthermore, with a small amount of speech data from target speakers, they synthesize high quality speech, good enough to become a threat to automatic speaker verification systems.

One question that naturally arises is whether this is enough to achieve intelligent-like, natural and long-term human-machine speech interaction. Unfortunately, it is not.

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Speech recognition and synthesis are only two of the six modules in a typical architecture of a speech dialogue system, depicted in Figure 4, and each of the six modules corresponds to certain cognitive aspects underlying the human language processing system. As a result, dialogue management becomes a complex structure that involves cooperation of several, quite different functional units. Leaving aside the division of dialogue systems into task-oriented and non-task-oriented, we must emphasize that the biggest challenge scientists face during the research in this area is to emulate human ability to understand meaning and conduct a conversation that is forward-looking, informative, and coherent. Regarding the dialogue initiatives, machines are successful in handling conversations that are system initiative (or single initiative). In such cases, the system completely controls the conversation and maintains the processes of speech recognition, meaning extraction, answer generation, and speech production. However, as it is known, natural dialogue is not deterministic and real improvements in human-machine speech interaction can be achieved only if adaptive behaviour with respect to the intention, the current context, and history of interaction are provided.

As a traditional paradigm shift, recent works in this area have addressed a series of data-driven, end-to-end trainable, non-goal-driven systems based on generative probabilistic models [131]. As such, these models can be viewed as artificial cognitive systems, aimed at grouping and carrying out traditional dialogue management tasks: language understanding, reasoning, decision-making, and natural language generation. They are corpus-based, data-driven dialogue systems, based on machine learning algorithms using corpora created from real word data. The statistics observed in dialogue corpora is the main knowledge for the optimization of parameters and variables.

It is worth pointing out that, besides the importance of domain knowledge, linguistic context has the crucial impact for active and engaging conversation. However, one of the main drawbacks of these approaches is related to sparsity issues that can be expected during integration of contextual information into statistical models. In the work of Sordoni et al. [132], the neural network architecture is used to address this problem, allowing the system to take into account the previous dialogue utterances. While modelling contextual information, the authors identify models for three linguistic entities in a conversation: the context (c), the message (m), and response (r). On that basis, they suggest three context-based generation models to estimate a response $r = r_1, \ldots, r_{\mathrm{T}}$ as follows:

$$p(r \mid c, m) = \prod_{t=1}^{\mathrm{T}} p(r_t \mid r_1, \ldots, r_{t-1}, c, m). \quad (1)$$

This work extends the recurrent neural network language models (RNNLM) as a generative model of sentences [133]. As the basic principle in this neural network model, input vector, representing the current word at time instant t, is concatenated to the output from neurons in network context layer at time $t-1$. In order to capture long-span dependencies together with semantic and syntactic similarities, the authors select word embedding as a continuous representation of words and phrases. Similar approaches already advanced classical language modelling, based on traditional n-gram language models [134, 135].

In line with this, many researches are trying to take advantage of combining neural network and end-to-end training possibilities with the large amount of freely available text or audio material from social media, movie scripts, etc. [136]. Serban et al. [137], for example, demonstrated introduction of latent variables to hierarchical recurrent encoder-decoder architecture. The research presented in [138] extended the hierarchical structure with the attention mechanism (word level attention and utterance level attention), taking into account that words and utterances in the context are differentially important.

Although end-to-end, statistical models have drawn most of the recent research on dialogue systems, many problems remain unresolved [131]. Neural network-based models are capable of handling large amount of data, but still it is hard to design an intelligent system based on imitating responses (especially if we take into account that the dialogue data for a specific domain are quite limited). Hence, to reduce these limitations, Mišković et al. [36] proposed a different, representational approach. This work extends the focus tree model, a cognitively inspired computational model of working memory that allows for adaptive dialogue management in human-machine interaction. The research not only is focused on improvements of speech recognition module but also points to possible new architectural aspect of dialogue systems. Following the assumption of the hierarchical and associative nature of human memory system and facts that the processing of the user's dialogue acts in human-machine interaction is always context-dependent, this model enables, to some extent, understanding of language and real word data.

4. Progress in Speech Signal Compression, Coding, and Transmission

In general, speech coders can be classified into three major categories depending on the applied coding technique: waveform coders, parametric coders, and hybrid coders. The primary idea behind a waveform coder design is to preserve the shape of a speech signal waveform, thus encoding information about the original time-domain waveform [4–6, 14, 139]. Such coders are broadly used in embedded applications due to several reasons: low cost of manufacturing, low computational resource usage, and high speech quality [4–7, 14, 139]. The simplest and most well-known type of waveform coders is pulse code modulation (PCM) coder, which is considered as a standard in digital telephony. One of the key advantages of PCM coders is that they are instantaneous, implying a coding delay of no more than one sample period [4]. Unlike waveform coders, which tend to reconstruct the original shape of the speech signal in time-domain, parametric coders reconstruct the speech signal from certain parameters that model the source signal, making no attempt to preserve the original shape of the waveform [4–7, 14, 139]. Due to this limitation, parametric coders are

PGR2026-00033
Krisp Technologies EX1010 Page 10

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

more signal dependent and less versatile. Additionally, compared to waveform coders, they provide a lower quality of speech signals. In parametric coders, human speech production mechanism is modelled with a time-varying filter, having coefficients commonly determined by the linear prediction analysis procedure. In the end, hybrid coders represent a class of coders, which combine features of both previously described classes of coders, namely, hybrid coders tend to preserve the shape of the signal in time domain and also exploit perceptive characteristics, that is, parametric approach [4–6, 14, 140]. Performance comparison of these three classes of coders is presented in Figure 5, where mean opinion score (MOS) is used as one of the standard subjective measures of reconstructed speech signal quality [4].

From Figure 5, one can conclude that waveform coders provide excellent quality of reconstructed speech signal and that they represent the best choice at bit rates higher than 16 kbits/s, whereas parametric coders cannot provide high quality regardless of the bit rate. On the other hand, parametric coders provide much better quality than waveform coders at low bit rates. Finally, hybrid coders are most suitable at medium bit rates. As for the purposes of speech synthesis and automatic speech recognition, the highest possible quality of reconstructed signal is desirable and waveform coders are usually considered as an adequate choice. Thus, what follows is focused on PCM and adaptive PCM (ADPCM) coding techniques.

4.1. Adaptive PCM. Speech signal can be considered as a nonstationary process, whose average power significantly fluctuates in time domain, resulting in a wide dynamic range [4]. However, speech signal can be considered as almost stationary in a short period of time (up to 30 ms). This means that speech signal has a highly predictable characteristics during short periods of time, which is suitable for utilizing adaptive quantization [4, 8–10, 141–147]. Commonly, adaptive quantization is frame-based, where frames are formed by dividing an input speech signal into sets of samples.

There are two fundamentally distinct categories of adaptive quantization techniques: forward and backward adaptive quantization techniques [141]. Forward adaptive techniques require transmission of additional information about the estimated gain, which is used for adaptation. Moreover, forward adaptive techniques require a longer processing delay than backward adaptive techniques as samples within a frame have to be stored in a buffer, in order to estimate predictable characteristics of every frame. When the gain is estimated and the quantizer is adapted, samples can be quantized and further transmitted to the decoder along with the quantized gain.

A general forward adaptive PCM model is presented in Figure 6 [8, 10]. The encoder is formed of two parts: a fixed (nonadaptive) part, consisting of a fixed quantizer Q_f, and an adaptive part, consisting of a buffer, a gain estimator, one divider, and a fixed gain quantizer Q_g. If Q_f is a piecewise linear μ-law quantizer designed for 8 bit/sample and $\mu = 255$,

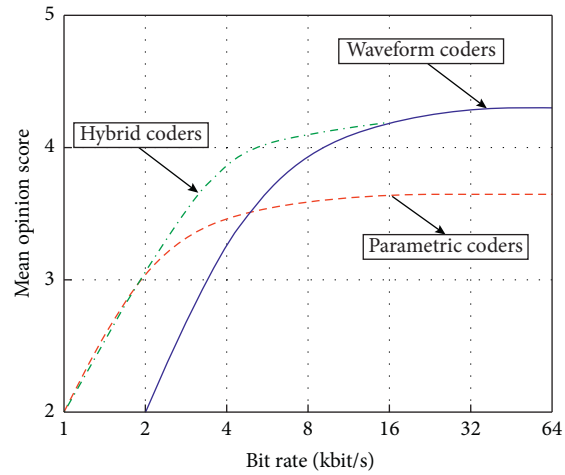

Figure 5: Speech signal quality according to MOS versus bit rate for various speech signal coding techniques.

the general forward adaptation model becomes a forward adaptive PCM model defined by G.711 standard [148].

Unlike forward adaptation, backward adaptation does not estimate characteristics of samples in a frame while encoding, which means that there is not additional information that has to be transmitted [149]. In fact, gain estimation is performed at the receiver after decoding, considering previously quantized samples. The simplest backward adaptive quantization model is based on uniform quantization with one codeword memory exploited for gain estimation and it is commonly referred as Jayant's model [4].

Advanced backward adaptive models commonly incorporate more sophisticated gain estimation methods, or variance ($\hat{\sigma}_y^2(n)$) estimation methods, which, for quantization of a current sample $x(n)$, typically exploit a larger number of previously decoded samples $y(n-i)$ [4]:

$$\hat{\sigma}_y^2(n) = \frac{1-\alpha}{1+\alpha} \sum_{i=1}^{+\infty} \alpha^{i-1} y^2(n-i), \tag{2}$$

where α is a weighting parameter, which can take values $0 < \alpha < 1$. Parameter α defines a learning period, that is, a time required for variance estimation [4]:

$$L = \frac{1+\alpha}{1-\alpha}. \tag{3}$$

Equation (2) can be written in the following recursive form:

$$\hat{\sigma}_y^2(n) = \frac{1-\alpha}{1+\alpha} y^2(n-1) + \alpha \hat{\sigma}_y^2(n-1), \tag{4}$$

which is straightforwardly used in the simplest mathematical model of Jayant's backward quantizer with one codeword memory. One of the realizations of backward adaptive PCM with one codeword memory that incorporates a widely used companding quantization model is shown in Figure 7, where $M(n-1)$ denotes a step size multiplier, used for adaptation, and $c(x)$ and $c^{-1}(x)$ are a compressor function and an expandor function, respectively.

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

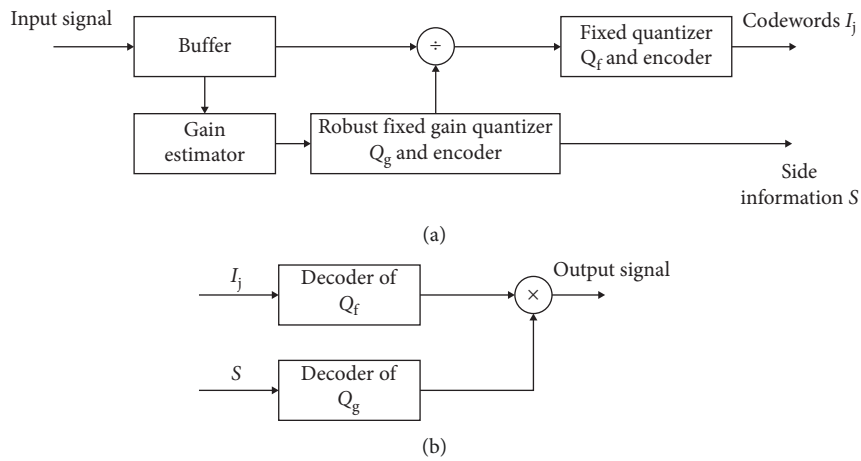

Figure 6: Forward adaptive PCM: (a) encoder; (b) decoder.

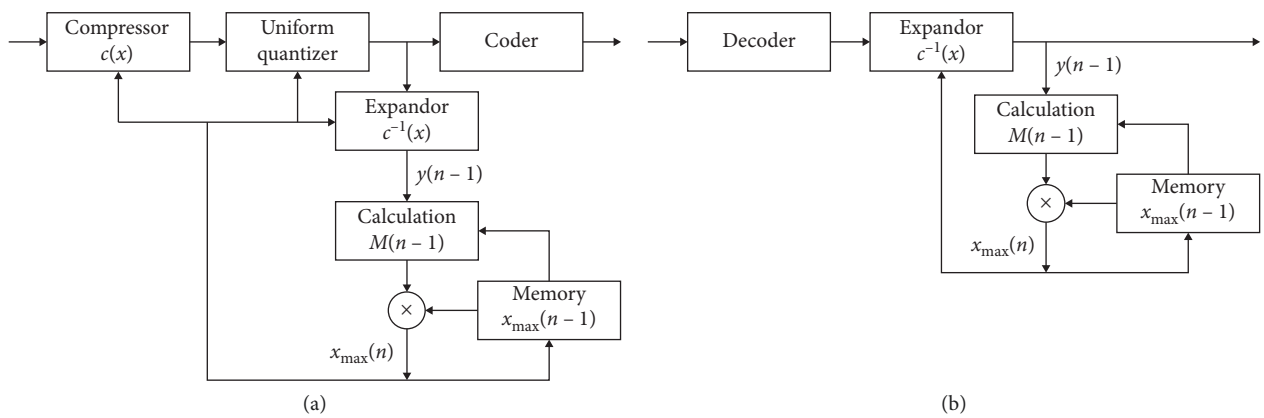

Figure 7: One of the realizations of backward adaptive PCM with one codeword memory: (a) encoder; (b) decoder.

4.2. Dual-Mode Quantization. Dual-mode and adaptive dual-mode quantizers belong to a relatively new class of quantizers whose design is based on multiparameter adaptation, such as variance and maximum amplitude [8, 9, 11]. Depending on their purpose and application, they can perform quantizer adaptation according to the frame variance and to the frame maximum amplitude x_{max} and also according to the subframe maximum amplitude. By utilizing two quantizers, which compose the dual-mode system, and by applying switched technique, it is possible to achieve a better quality of the quantized signal, or a higher compression, compared to the common single-mode quantizers. In Figure 8, a dual-mode quantization scheme is shown, where Encoder 1 and Decoder 1 are related to the quantizer applied for processing signals having restricted amplitude range, whereas Encoder 2 and Decoder 2 are related to the quantizer applied for processing the signals having unrestricted amplitude range [8, 9, 11]. The switched process is frame-based, and it is performed so that the restricted quantizer is used in the case if all samples within a frame belong to the restricted quantizer's support region, while the unrestricted quantizer is used otherwise [8, 9, 11]. The main idea behind such quantization model is to enable a more preferable selection of the restricted quantizer, with a narrower support region, than the unrestricted one, since, in such a manner, an increase of the signal to quantization noise ratio can be provided.

Considering that speech signal can be described using Gaussian probability density function (PDF) or Laplacian PDF, which is heavy-tailed, it is expected that only small percentage of speech frames will have some samples of large values. However, this also depends on the size of a frame. Consequently, the support region threshold values for both quantizers should be chosen so that the restricted quantizer usage should be dominant, but taking also into account the frame size and the whole input signal dynamics in order to achieve a minimum of the total distortion introduced in the quantization process [8, 9, 11].

4.3. Differential Pulse Code Modulation. Differential pulse code modulation (DPCM) represents a simple but high-quality speech signal coding scheme for middle bit rates. It initially exploited uniform quantization and the first-order prediction [150, 151]. As it was already discussed, speech signal has highly predictable characteristics within a frame, which is exploited to reduce the dynamic range of amplitudes

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

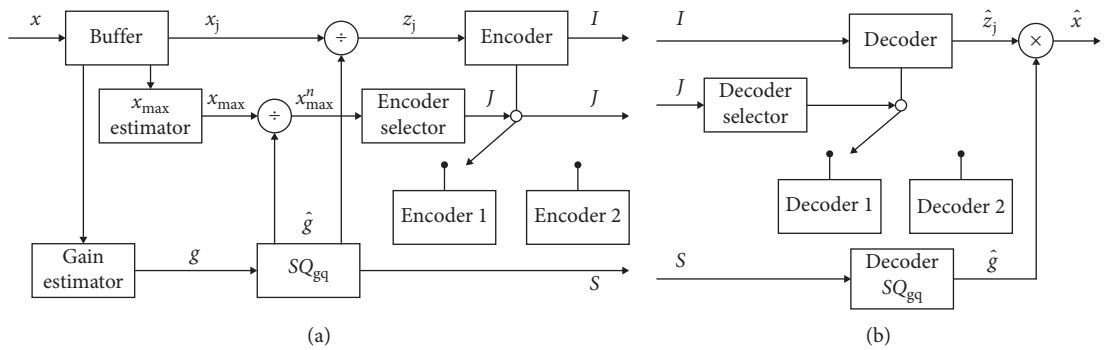

Figure 8: Dual mode quantization scheme: (a) encoder; (b) decoder.

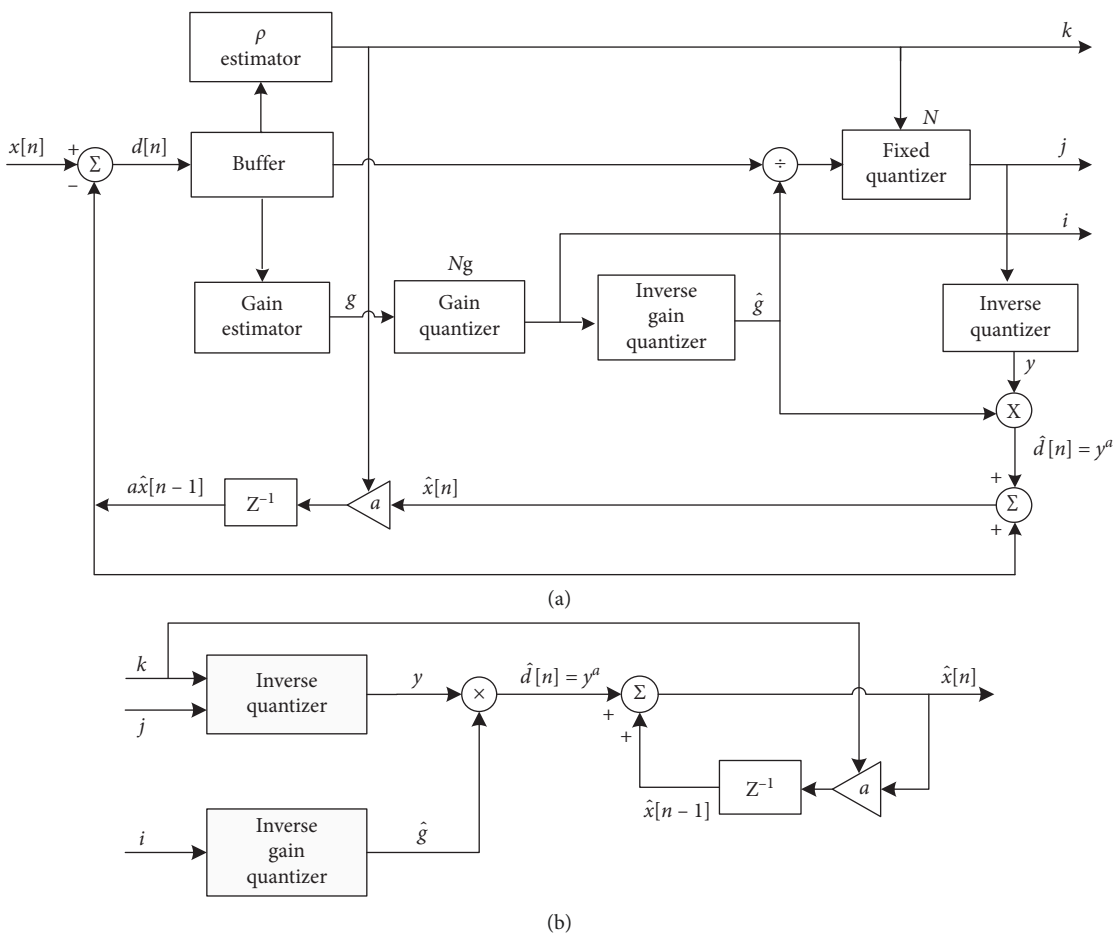

Figure 9: DPCM: (a) encoder; (b) decoder.

for quantization in the DPCM scheme [4, 12, 13, 152, 153]. In particular, DPCM predicts the next sample amplitude value and encodes the difference between the predicted value and the value of the current input signal amplitude. Due to the high correlation, these differences have much smaller values compared to ones of the input signal samples, so that the dynamic range of amplitudes is significantly reduced before quantization. Accordingly, with a suitable design of a DPCM system, a certain distortion may be provided at lower bit rates compared to the PCM system. In other words, a worthy compression may be achieved with the DPCM system compared to the PCM.

More sophisticated solutions may incorporate prediction of a higher order or other kinds of quantization models, such as a gain-adaptive quantization model [154]. Figure 9 shows a DPCM scheme with incorporated simple first-order predictor and forward gain-adaptive quantizer based on optimal companding model [151].

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In the DPCM system given in Figure 9, the reconstructed speech signal $\hat{x}$ is determined by

$$\hat{x}[n] = a \cdot \hat{x}[n-1] + y^a[n], \tag{5}$$

where y^a denotes the output of the adaptive quantizer, whereas n denotes the n-th sample of the signal. It can be noted that the value of parameter a depends on the nature of the input signal. If an input signal is highly correlated, it is preferred to use values close to 1 (e.g., $a = 0.8$), whereas values close to zero are preferred for lowly correlated signals (e.g., $a = 0.3$). However, the choice of parameter a is not an easy task even if adaptation is applied. The determination of linear predictor coefficients can be done using methods that are based on statistical learning such as least mean squares (LMS) estimation method [155]. LMS search algorithm reduces distortion by adapting coefficients for each input sample, and its main features, which attract researches, are low computational complexity, proof of convergence in stationary environment, unbiased convergence in the mean to the Wiener solution, and stable behaviour when implemented with finite-precision arithmetic [156]. Moreover, coefficients of linear predictor as well as determination of other important parameters for quantizer design may be determined by exploiting artificial neural networks or regression methods.

5. Conclusions

This review article has provided an overview in the recent development of speech technologies and other scientific areas related to them, mostly due to the development of the new machine learning paradigm, which has had a tremendous impact in this domain. Apart from natural speech production and speech perception, understanding of cognitive aspects of speech communication is very important for future HCI systems including both spoken language understanding and generation as language technologies. The machine learning paradigm has had a great impact on automatic speech recognition (ASR) and text-to-speech synthesis (TTS) as basic speech technologies. It is expected that ASR systems based on deep learning and adaptive algorithms in the near future will be able to recognize spontaneous speech in complex acoustic environments, with the accuracy that will surpass the corresponding ability of humans. Synthetic speech has already reached such quality that is hard or impossible to differentiate from human speech. With flexibility of changing speaker and style, HCI is becoming as pleasant and natural as human-human interaction. Unsupervised and reinforcement-based machine learning algorithms will also develop further, which will, in turn, bring about progress in areas where large data sets are not available, as is the case in speech analysis for speech recognition and synthesis for under-resourced languages. A short overview of speech coding techniques and of current progress in adaptive scalar quantization has been presented as the quality of digitized and compressed speech signal is important for accurate automatic speech signal detection and synthesis. Although these techniques can be designed to be robust in a wide dynamic range of speech signal variations, or to be frame-adaptive, one can anticipate that machine learning tools of increasing popularity will lead to novel solutions, which will improve performances of various systems by adapting predictive coefficients. To conclude, we are witnessing an increasingly fast progress in the field of speech signal processing due to machine learning paradigms, and it appears very hard to predict what they will bring about next and how soon that can be expected.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work described in this paper was supported in part by the Ministry of Education, Science and Technological Development of the Republic of Serbia, through the project "Development of Dialogue Systems for Serbian and Other South Slavic Languages" (TR32035).

References

[1] T. S. Kuhn, *The Structure of Scientific Revolutions-50th Anniversary Edition*, vol. 3, 4th edition, 2012.

[2] R. K. Moore, "Cognitive informatics: the future of spoken language processing?," in *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)*, Patras, Greece, October 2005.

[3] J. D. Paul, "Re-creating the sigsaly quantizer: this 1943 analog-to-digital converter gave the allies an unbreakable scrambler-(resources)," *IEEE Spectrum*, vol. 56, no. 2, pp. 16-17, 2019.

[4] N. S. Jayant and P. Noll, "Digital coding of waveforms. Principles and applications to speech and video," *Signal Processing*, vol. 9, no. 2, pp. 139-140, 1985.

[5] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, John Wiley & Sons, Hoboken, NJ, USA, 2003.

[6] L. Hanzo, F. C. A. Somerville, and J. P. Woodard, *Voice and Audio Compression for Wireless Communications*, John Wiley & Sons Ltd., Chichester, England, 2nd edition, 2007.

[7] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, John Wiley & Sons Ltd., Chichester, UK, 2nd edition, 2004.

[8] Z. H. Peric and J. Nikolic, "An adaptive waveform coding algorithm and its application in speech coding," *Digital Signal Processing*, vol. 22, no. 1, pp. 199–209, 2012.

[9] J. Nikolic, Z. H. Peric, and A. Z. Jovanovic, "Two forward adaptive dual-mode companding scalar quantizers for Gaussian source," *Signal Processing*, vol. 120, pp. 129–140, 2016.

[10] J. Nikolic and Z. H. Peric, "Lloyd-Max's algorithm implementation in speech coding algorithm based on forward adaptive technique," *Informatica (Lithuanian Academy of Sciences)*, vol. 19, no. 2, pp. 255–270, 2008.

[11] Z. Peric and J. Nikolic, "High-quality Laplacian source quantisation using a combination of restricted and unrestricted logarithmic quantisers," *IET Signal Processing*, vol. 6, no. 7, pp. 633–640, 2012.

[12] O. Ordentlich and U. Erez, "Performance analysis and optimal filter design for sigma-delta modulation via duality

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with DPCM," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 1153–1164, 2019.

[13] R. C. Farias and J. M. Brossier, "Adaptive quantizers for estimation," *Signal Processing*, vol. 93, no. 11, pp. 3076–3087, 2013.

[14] I. V. McLoughlin, *Speech and Audio Processing: a MATLAB-Based Approach*, Cambridge University Press, Cambridge, UK, 2016.

[15] M. Cernak, A. Asaei, and A. Hyafil, "Cognitive speech coding: examining the impact of cognitive speech processing on speech compression," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 97–109, 2018.

[16] Neurophysiology of Speech Communication, http://naplab.ee.columbia.edu/speech-neurophysiology.html.

[17] F. A. Everest and K. C. Pohlmann, *Master Handbook of Acoustics*, McGraw-Hill, New York, NY, USA, 5th edition, 2009, http://www.roletech.net/books/HandbookAcoustics.pdf.

[18] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[19] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1-2, pp. 1–194, 2007.

[20] X. Huang, A. Acero, H. W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[21] S. Jovičić, *Speech Communication—Physiology of Psychoacoustics and Perception*, Izdavačko Preduzeće Nauka, Belgrade, Serbia, 1999, in Serbian.

[22] T. D. Rossing, *Springer Handbook of Acoustics*, Springer Nature, New York, NY, USA, 2007.

[23] J. Blauert, *Communication Acoustics*, Springer, Berlin, Germany, 2005.

[24] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[25] C. Brokish and M. Lewis, *A-Law and Mu-Law Companding Implementations Using the TMS320C54x. Application Note SPRA163A*, Texas Instruments, Dallas, TX, USA, 1997, http://www.ti.com/lit/an/spra163a/spra163a.pdf.

[26] D. M. W. Powers, "Applications and explanations of Zipf's law," in *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning, NeMLaP/CoNLL*, pp. 151–160, Sydney, Australia, January 1998.

[27] X. D. Huang, M. Jack, and M. Ariki, *Hidden Markov Models for Speech Recognition*, Columbia University Press, New York, NY, USA, 1990.

[28] D. M. Howard and J. A. S. Angus, *Acoustics and Psychoacoustics*, Elsevier Ltd., Oxford, UK, 4th edition, 2009, https://www.sciencedirect.com/book/9780240521756/acoustics-and-psychoacoustics.

[29] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[30] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA, USA, 1994.

[31] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, Berlin, Germany, 2008.

[32] Y. Ando, *Auditory and Visual Sensations*, Springer, New York, NY, USA, 2010, https://link.springer.com/book/10.1007/b13253.

[33] T. Nagamine and N. Mesgarani, "Understanding the representation and computation of multilayer perceptrons: a case study in speech recognition," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 2564–2573, Sydney, Australia, August 2017.

[34] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[35] M. Gnjatović and V. Delić, "Cognitively-inspired representational approach to meaning in machine dialogue," *Knowledge Based Systems*, vol. 71, no. 1, pp. 25–33, 2014.

[36] D. Mišković, M. Gnjatović, P. Štrbac, B. Trenkić, N. Jakovljević, and V. Delić, "Hybrid methodological approach to context-dependent speech recognition," *International Journal of Advanced Robotic Systems*, vol. 14, no. 1, 2017.

[37] S. McAdams and C. Drake, "Auditory perception and cognition," in *Stevens' Handbook of Experimental Psychology*, H. Pashler, Ed., pp. 397–452, John Wiley & Sons, New York, NY, USA, 2002.

[38] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "On the role of nonlinear transformations in deep neural network acoustic models," in *Proceedings of Interspeech 2016 17th Annual Conference of the International Speech Communication Association*, pp. 803–807, San Francisco, CA, USA, September 2016.

[39] W. Zhang, H. Li, M. Yang, and N. Mesgarani, "Synaptic depression in deep neural networks for speech processing," in *Proceedings 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5865–5869, Shanghai, China, March 2016.

[40] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[41] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.

[42] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, and H. Ney, "LSTM, GRU, highway and a bit of attention: an empirical overview for language modeling in speech recognition," in *Proceedings of Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pp. 3519–3523, San Francisco, CA, USA, September 2016.

[43] W. Xiong, J. Droppo, X. Huang et al., "Achieving human parity in conversational speech recognition," https://arxiv.org/pdf/1610.05256.pdf.

[44] A. Hannun, C. Case, J. Casper et al., "Deep Speech: scaling up end-to-end speech recognition," http://arxiv.org/abs/1412.5567.

[45] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proceedings 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, Shanghai, China, March 2016.

[46] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[47] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7962–7966, Vancouver, BC, Canada, May 2013.

[48] H. Zen and A. W. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3844–3848, Florence, Italy, May 2014.

[49] Y. Fan, Y. Qian, F. L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proceedings of Interspeech 2014, 15th Annual Conference of the International Speech Communication Association*, pp. 1964–1968, Singapore, September 2014.

[50] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional deep recurrent neural networks," in *Proceedings of Interspeech 2014, 15th Annual Conference of the International Speech Communication Association*, pp. 2268–2272, Singapore, September 2014.

[51] A. van den Oord, S. Dieleman, H. Zen et al., "WaveNet: a generative model for raw audio," https://arxiv.org/abs/1609.03499.

[52] Y. Wang, R. J. Skerry-Ryan, D. Stanton et al., "Tacotron: towards end-to-end speech synthesis," in *Proceedings of the Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, pp. 4006–4010, Stockholm, Sweden, August 2017.

[53] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4475–4479, South Brisbane, Australia, April 2015.

[54] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proceedings of the Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pp. 2278–2282, San Francisco, CA, USA, September 2016.

[55] H. T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4905–4909, New Orleans, LA, USA, March 2017.

[56] Y. Wang, D. Stanton, Y. Zhang et al., "Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 5167–5176, Stockholm, Sweden, June 2018.

[57] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to expressive speech synthesis," in *Proceedings of the Fifth ISCA ITRW on Speech Synthesis*, pp. 79–84, Pittsburgh, PA, USA, June 2004.

[58] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1545–1554, San Diego, CA, USA, June 2016.

[59] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014.

[60] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pp. 649–657, Montreal, Canada, December 2015.

[61] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pp. 3104–3112, Montreal, Canada, December 2014.

[62] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[63] C. N. D. Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78, Dublin, Ireland, August 2014.

[64] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 655–665, Baltimore, MD, USA, June 2014.

[65] T. Mikolov, W. t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, GA, USA, June 2013.

[66] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011.

[67] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., vol. 116, pp. 374–388, Academic Press, New York, NY, USA, 1976.

[68] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[69] A. R. Mohamed, G. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in *Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, Whistler, Canada, December 2009.

[70] L. Deng, G. E. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8599–8603, Vancouver, Canada, May 2013.

[71] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[72] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning*, pp. 1764–1772, Beijing, China, June 2014.

[73] B. M. Popovic, E. Pakoci, and D. Pekar, "End-to-end large vocabulary speech recognition for the Serbian language," in *Proceedings of the 19th International Conference Speech and Computer (SPECOM)*, pp. 343–352, Hatfield, UK, September 2017.

[74] D. Povey, V. Peddinti, D. Galvez et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI,"

PGR2026-00033
Krisp Technologies EX1010 Page 16

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in *Proceedings of the Interspeech 2016 17th Annual Conference of the International Speech Communication Association*, pp. 2751–2755, San Francisco, CA, USA, September 2016.

[75] E. Pakoci, B. Popović, and D. J. Pekar, "Improvements in Serbian speech recognition using sequence-trained deep neural networks," *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 53–76, 2018.

[76] P. Zhang and N. Li, "The importance of affective quality," *Communications of the ACM*, vol. 48, no. 9, pp. 105–108, 2005.

[77] R. Bezooijen, *Characteristics and Recognizability of Vocal Expressions of Emotion*, Foris Publications Holland, Dordrecht, Netherlands, 1984.

[78] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *Journal of Acoustical Society of America*, vol. 96, no. 6, pp. 3392–3400, 1994.

[79] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of the ICSLP 2002 7th International Conference on Spoken Language Processing*, pp. 2037–2040, Denver, CO, USA, September 2002.

[80] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[81] Y. Wang, S. Du, and Y. Zhan, "Adaptive and optimal classification of speech emotion recognition," in *Proceedings of the Fourth International Conference on Natural Computation ICNC*, vol. 5, pp. 407–411, Jinan, China, October 2008.

[82] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[83] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[84] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[85] V. Delic, M. Bojanic, M. Gnjatovic, M. Secujski, and S. T. Jovicic, "Discrimination capability of prosodic and spectral features for emotional speech recognition," *Electronics and Electrical Engineering*, vol. 18, no. 9, pp. 51–54, 2012.

[86] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques," *Procedia Computer Science*, vol. 49, pp. 50–57, 2015.

[87] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proceedings of the 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, vol. 2, pp. 621–625, Harbin, China, August 2011.

[88] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of the Interspeech 2014 15th Annual Conference of the International Speech Communication Association*, pp. 223–227, Singapore, September 2014.

[89] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proceedings of the Interspeech 2015, 16th Annual Conference of the International Speech Communication Association*, pp. 1537–1540, Dresden, Germany, September 2015.

[90] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, pp. 873–876, Denver, CO, USA, September 2002.

[91] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[92] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, no. 1, pp. 312–323, 2019.

[93] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.

[94] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[95] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[96] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[97] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: features and classification models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[98] M. R. Schroeder, "A brief history of synthetic speech," *Speech Communication*, vol. 13, no. 1-2, pp. 231–237, 1993.

[99] H. Dudley, "The carrier nature of speech," *Bell System Technical Journal*, vol. 19, no. 4, pp. 495–515, 1940.

[100] R. Teranishi and N. Umeda, "Use of pronouncing dictionary in speech synthesis experiments," *Reports of the Sixth International Congress on Acoustics*, vol. 2, pp. 155–158, 1968.

[101] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[102] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 373–376, Atlanta, GA, USA, May 1996.

[103] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of the Eurospeech 1999 6th European Conference on Speech Communication and Technology*, pp. 2347–2350, Budapest, Hungary, September 1999.

[104] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[105] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proceedings of the Eurospeech 2003 8th European Conference on Speech Communication and Technology*, pp. 2461–2464, Geneva, Switzerland, September 2003.

[106] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.

[107] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[108] T. Weijters and J. Thole, "Speech synthesis with artificial neural networks," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1764–1769, San Francisco, CA, USA, April 1993.

[109] Z. Wu, O. Watts, and S. King, "Merlin: an open source neural network speech synthesis system," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pp. 202–207, Sunnyvale, CA, USA, July 2016.

[110] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Proceedings of the Interspeech 2016 17th Annual Conference of the International Speech Communication Association*, pp. 2273–2277, San Francisco, CA, USA, September 2016.

[111] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[112] Z. Wu and S. King, "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[113] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proceedings of the Interspeech 2015 16th Annual Conference of the International Speech Communication Association*, pp. 879–883, Dresden, Germany, September 2015.

[114] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5540–5544, Shanghai, China, April 2016.

[115] T. Delic, S. Suzic, M. Secujski, and D. Pekar, "Rapid development of new TTS voices by neural network adaptation," in *Proceedings of the 2018 17th International Symposium Infoteh-Jahorina (Infoteh)*, pp. 1–6, Istočno Sarajevo, Bosnia and Herzegovina, March 2018.

[116] M. Abe, "Speaking styles: statistical analysis and synthesis by a text-to-speech system," in *Progress in Speech Synthesis*, J. P. H. Van Santen et al., Ed., pp. 495–510, Springer, New York, NY, USA, 1997, https://link.springer.com/chapter/10.1007/978-1-4612-1894-4_39.

[117] S. Brave and N. Clifford, "Emotion in human-computer interaction," in *Human-Computer Interaction Fundamentals*, A. Sears and J. A. Jacko, Eds., pp. 54–66, CRC Press, Boca Raton, FL, USA, 2009.

[118] S. Suzić, T. Delić, V. Jovanović et al., "A comparison of multi-style DNN-based TTS approaches using small datasets," in *Proceedings of the 13th International Scientific-Technical Conference on Electromechanics and Robotics Zavalishin's Readings*, vol. 161, pp. 1–6, St. Petersburg, Russia, May 2018.

[119] J. Lorenzo-Trueba, G. Eje Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[120] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1613–1616, Kuala Lumpur, Malaysia, December 2017.

[121] S. Suzić, T. Delić, D. Pekar, V. Delić, and M. Sečujski, "Style transplantation in neural network based speech synthesis," *Acta Polytechnica Hungarica*, 2019.

[122] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "An investigation to transplant emotional expressions in DNN-based TTS synthesis," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1253–1258, Kuala Lumpur, Malaysia, December 2017.

[123] A. van den Oord, Y. Li, I. Babuschkin et al., "Parallel WaveNet: fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3915–3923, Stockholm, Sweden, June 2018.

[124] S. O. Arik, M. Chrzanowski, A. Coates et al., "Deep voice: real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 195–204, Sydney, Australia, August 2017.

[125] J. Shen, R. Pang, R. J. Weiss et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, Calgary, Canada, September 2018.

[126] W. Ping, K. Peng, A. Gibiansky et al., "Deep voice 3: scaling text-to-speech with convolutional sequence learning," https://arxiv.org/abs/1710.07654.

[127] S. Ö. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pp. 10040–10050, Montreal, Canada, December 2018.

[128] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3680–3688, Stockholm, Sweden, June 2018.

[129] R. Skerry-Ryan, E. Battenberg, Y. Xiao et al., "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 4693–4702, Stockholm, Sweden, July 2018.

[130] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proceedings of the Interspeech 2018 19th Annual Conference of the International Speech Communication Association*, pp. 3067–3071, Hyderabad, India, September 2018.

[131] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[132] A. Sordoni, M. Galley, M. Auli et al., "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 2015 Conference of the North American Chapter of the of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pp. 196–205, Denver, CO, USA, May 2015.

[133] T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký, and S. Khudanpur, "Recurrent neural network based language

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

model," in *Proceedings of the Interspeech 2011 12th Annual Conference of the International Speech Communication Association*, pp. 1045–1048, Florence, Italy, August 2011.

[134] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1137–1155, 2003.

[135] J. Gao, X. He, W.-T. Yih, and L. Deng, "Learning continuous phrase representations for translation modeling," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 699–709, Baltimore, MD, USA, June 2014.

[136] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," https://arxiv.org/pdf/1512.05742.pdf.

[137] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3776–3783, Phoenix, AR, USA, February 2016.

[138] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5610–5617, New Orleans, LA, USA, February 2018.

[139] P. Sinha, *Speech Processing in Embedded Systems*, Springer, Berlin, Germany, 2009.

[140] M. Hasegawa-Johnson and A. Alwan, "Speech coding: fundamentals and applications," in *Wiley Encyclopedia of Telecommunications*, J. G. Proakis, Ed., John Wiley & Sons, Hoboken, NJ, USA, 2003.

[141] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Springer Nature, New York, NY, USA, 1991.

[142] Z. H. Perić, J. R. Nikolić, A. V. Mosić, and M. D. Petković, "Design of fixed and adaptive companding quantizer with variable-length codeword for memoryless Gaussian source," *Informatica (Lithuanian Academy of Sciences)*, vol. 24, no. 1, pp. 71–86, 2013.

[143] J. Nikolic, Z. Peric, A. Jovanovic, and D. Antic, "Design of forward adaptive piecewise uniform scalar quantizer with optimized reproduction level distribution per segments," *Electronics and Electrical Engineering*, vol. 119, no. 3, pp. 19–22, 2012.

[144] J. Nikolić, Z. Perić, D. Antić, A. Jovanović, and D. Denić, "Low complex forward adaptive loss compression algorithm and its application in speech coding," *Journal of Electrical Engineering-Elektrotechnicky Casopis*, vol. 62, no. 1, pp. 19–24, 2011.

[145] Z. Perić, A. Jocić, J. Nikolić, L. Velimirović, and D. Denić, "Analysis of differential pulse code modulation with forward adaptive Lloyd-Max's quantizer for low bit-rate speech coding," *Revue Roumaine des Sciences Techniques. Ser. Electrotechnique et Energetique*, vol. 58, no. 4, pp. 424–434, 2013.

[146] Z. H. Perić, J. R. Nikolić, and A. V. Mosić, "Design of forward adaptive hybrid quantiser with golomb-rice code for compression of Gaussian source," *IET Communications*, vol. 8, no. 3, pp. 372–377, 2014.

[147] S. Tomic, Z. Peric, M. Tancic, and J. Nikolic, "Backward adaptive and quasi-logarithmic quantizer for sub-band coding of audio," *Information Technology and Control*, vol. 47, no. 1, pp. 131–139, 2018.

[148] Telkom Institute of Technology, *Recommendation G.711, Pulse Code Modulation (PCM) of Voice Frequencies*, Telkom Institute of Technology, Bandung, Indonesia, 1972, https://www.itu.int/rec/T-REC-G.711.

[149] A. Ortega and M. Vetterli, "Adaptive scalar quantization without side information," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 665–676, 1997.

[150] P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," *Bell System Technical Journal*, vol. 52, no. 7, pp. 1105–1118, 1973.

[151] V. M. Despotovic, Z. H. Peric, L. Velimirovic, and V. D. Delic, "DPCM with forward gain-adaptive quantizer and simple switched predictor for high quality speech signals," *Advances in Electrical and Computer Engineering*, vol. 10, no. 4, pp. 95–98, 2010.

[152] S. S. Tomic, Z. H. Peric, and J. R. Nikolic, "An algorithm for simple differential speech coding based on backward adaptation technique," *Informatica*, vol. 29, no. 3, pp. 539–553, 2018.

[153] R. Pratheek and M. N. Suma, "Performance analysis of DPCM and ADPCM," in *Proceedings of the International Conference on Electronic Design and Signal Processing*, pp. 19–23, Manipal, India, December 2013.

[154] J. H. Chen and A. Gersho, "Gain-adaptive vector quantization with application to speech coding," *IEEE Transactions on Communications*, vol. 35, no. 9, pp. 918–930, 1987.

[155] S. O. Haykin, *Neural Networks and Learning Machines*, Prentice-Hall, New Jersey, NJ, USA, 3rd edition, 2009.

[156] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*, Springer, Boston, MA, USA, 4th edition, 2013, https://www.springer.com/gp/book/9781461441052.

8483, 2019, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/2019/4368036, Wiley Online Library on [11/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License