PRINCIPLES OF DIGITAL SPEECH SIGNAL PROCESSING AND ANALYSIS OF THE ACOUSTIC CHARACTERISTICS OF HUMAN AND SYNTHETIC SPEECH

S.U.Nasirov
sultan250593@gamil.com
Tashkent University of Information Technologies named after Muhammad al-Khorezmi

Abstract: This article examines the growing importance of digital speech signal processing in fields such as cybersecurity, biometrics, and human-computer interaction, particularly in response to synthetic speech and deepfake voice threats. It traces the historical development of speech synthesis from early mechanical models (Kratzenstein, von Kempelen) to modern systems, while highlighting a critical gap: the lack of comprehensive TTS systems for the Uzbek language. The article discusses the technical requirements for building a Uzbek-language speech synthesizer - including phonetic modeling, prosody assignment, and acoustic feature extraction (e.g., MFCCs) - and emphasizes its potential to improve digital accessibility for visually impaired users and enable voice-based interaction in underrepresented languages.

Keywords: Speech signal processing, Text-to-speech synthesis, Uzbek language speech synthesis, Acoustic features, Assistive technology for visually impaired, Synthetic speech detection

Today, digital speech signal processing has become one of the most important areas in artificial intelligence, biometric identification, voice assistants, automatic translation, human-computer interaction, and cybersecurity systems. In particular, the rapid proliferation of synthetic speech, deepfake voices, and automated voice-based attacks has transformed the analysis of speech signals, the discrimination between human and synthetic speech, and the improvement of speech recognition accuracy into significant scientific and practical challenges. Consequently, the development of advanced speech signal processing algorithms, the identification of informative acoustic features, and their adaptation to modern intelligent systems have emerged as important directions of contemporary research.

At present, these technologies are widely applied in the following domains:

- Cybersecurity - for protecting voice authentication systems against spoofing attacks and unauthorized access.
- Banking and Finance - for detecting fraudulent activities and identity impersonation attempts conducted through telephone communications.
- Media and Journalism - for verifying the authenticity of audio recordings and identifying manipulated or artificially generated speech content.
- Forensic Science - for examining audio evidence and determining the authenticity and integrity of recorded speech signals.

These applications highlight the growing importance of speech signal analysis and human-synthetic speech discrimination in ensuring security, reliability, and trustworthiness across various sectors.

In the era of rapidly advancing information technologies, blind and visually impaired pupils, students, and young people are increasingly able to access printed and electronic resources in foreign languages independently through modern computer technologies. Most of them rely on auditory perception to obtain information and therefore use assistive technologies developed abroad, including specialized devices that scan and read printed texts aloud, as well as screen-reader software such as JAWS and NVDA. However, accessing textual information in the Uzbek language through such

software and hardware solutions remains challenging. This limitation arises because the speech synthesizers integrated into these systems are designed for specific languages and generate speech according to the phonetic rules and pronunciation standards of those languages. Consequently, Uzbek-language texts are often rendered inaccurately or unnaturally. To date, comprehensive Uzbek-language speech synthesizers and dedicated software tools that enable visually impaired individuals to effectively utilize computer and Internet technologies remain limited and insufficiently developed.

Voice-based computer technologies are becoming increasingly widespread in modern society. The challenge of enabling computers to communicate with humans through natural spoken language requires extensive research efforts involving specialists and leading scientists in the field of information and communication technologies. Modern computer systems have already achieved significant progress in understanding human commands and converting textual information into spoken output. However, such interactions are generally limited to languages that have been explicitly incorporated into the system during development. At present, comprehensive methods, speech synthesizers, and software tools that enable effective human-computer interaction in the Uzbek language remain insufficiently developed.

A speech synthesizer is a system capable of converting textual or visual information into spoken language through software and hardware technologies. Speech synthesis is required in a wide range of applications where information is intended to be delivered through auditory channels. The quality of a speech synthesizer is primarily evaluated based on two criteria: its similarity to natural human speech and the intelligibility of the generated output. The simplest form of synthesized speech can be produced by concatenating pre-recorded speech segments that are stored in a database and combined to form complete utterances. This approach served as the foundation for many early speech synthesis systems and contributed significantly to the development of modern text-to-speech technologies.

The earliest significant research on speech synthesis can be traced back to the late eighteenth century and the work of the Danish scientist Christian Kratzenstein, a member of the Russian Academy of Sciences. He developed a mechanical model of the human voice capable of producing five vowel sounds (a, e, u, o, and y). The system consisted of a set of acoustic resonators of various shapes that generated vowel sounds through vibrating reeds excited by airflow. In 1778, the Austrian scientist Wolfgang von Kempelen extended Kratzenstein's model by incorporating representations of the tongue and lips, resulting in an acoustic-mechanical speech machine capable of reproducing certain speech sounds and their combinations. Later, in 1837, Charles Wheatstone introduced an improved version of the machine that could reproduce vowels as well as most consonant sounds. In 1846, Joseph Faber demonstrated his speech synthesis device known as Euphonia, which attempted not only to synthesize speech but also to produce singing voice output. Today, the most advanced speech synthesizers are employed for widely spoken languages and are extensively used across the world. However, each speech synthesizer is typically designed according to the linguistic, phonetic, and grammatical rules of a specific language, making it a language-dependent system. Consequently, a speech synthesizer developed for one language cannot be directly applied to another language without substantial modifications. Therefore, the development of a dedicated speech synthesizer for each language is considered the most effective approach. Achieving this objective requires the completion of several development and adaptation stages, including linguistic analysis, phonetic modeling, acoustic data collection, and speech generation model training.

Speech synthesis is one of the fundamental speech processing techniques concerned with converting electronic text stored in a computer system into spoken language. Several approaches to speech synthesis have been developed, among which one of the most common is the generation of words through the concatenation of phonemes and allophones. After determining the sequence of

phonemes, appropriate pitch and intonation parameters are assigned, and the resulting sequence is converted into speech. Although this approach produces intelligible speech, listeners can often recognize that the output has been generated by a machine rather than a human speaker.

Achieving high-quality speech synthesis requires substantial computational resources, including high data-processing speeds and considerable memory capacity for storing linguistic and acoustic information. In most speech synthesis systems, natural-sounding speech is generated through the automatic assignment of stress patterns and prosodic features while taking into account the linguistic characteristics of the target language. Another important aspect of the proposed project is that the planned Uzbek-language speech synthesizer and voice-based human-computer interaction system are intended to be applied not only in conventional computer environments but also in computer telephony applications. The project aims to develop an Uzbek-language speech synthesis system and software platform that enables users of all ages, particularly individuals with visual impairments, to access computer and Internet technologies through the auditory presentation of on-screen information.

A speech signal is a complex acoustic signal that contains distinctive biometric characteristics of an individual speaker. Although speech is inherently time-varying, it exhibits quasi-stationary behavior over short intervals. Typically, the spectral composition of a speech signal remains relatively stable within time windows ranging from approximately 5 to 100 milliseconds. For this reason, speech signal analysis relies heavily on short-term spectral parameters, including amplitude, energy, formant frequencies, Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency, and other acoustic features. Temporal variations in these characteristics correspond to different phonemes and speech sounds, thereby conveying linguistic information.

The primary information contained in a speech signal is represented through its short-term spectral amplitudes and waveform characteristics. During transmission, storage, and processing, speech signals are often subjected to various technical distortions. Noise, channel impairments, compression algorithms, and transmission losses can significantly degrade signal quality. Therefore, the main objective of any speech processing system is to transform the signal into a form suitable for processing, transmission, or storage while preserving the useful information it contains. In this process, maintaining the semantic content, naturalness, and individual characteristics of speech is of critical importance. Speech processing technologies are widely utilized in telecommunications, banking services, public administration, security systems, healthcare, robotics, and intelligent devices. In particular, the use of voice-based biometric authentication for user identification, interaction with virtual assistants, automatic translation systems, and voice-enabled customer support services in call centers has significantly increased the practical importance of speech technologies.

The process of speech production is closely associated with the physiological mechanisms of the human vocal apparatus. Speech sounds are generated through the interaction of airflow from the lungs, vibration of the vocal folds, and the articulatory movements of the tongue, lips, jaw, and vocal tract. Speech sounds can be categorized into several types. Voiced sounds are produced by the vibration of the vocal folds. Plosive sounds result from the sudden release of accumulated air pressure. Fricative sounds are generated by turbulent airflow passing through a narrow constriction in the vocal tract. Affricates are produced as a combination of plosive and fricative articulations. The analysis of these physiological and acoustic properties plays a crucial role in the development of speech synthesis and speech recognition systems.

The shape of the vocal tract determines the frequency characteristics of a speech signal. The resonance frequencies of the speech signal are known as formants, which represent one of the most important acoustic features of human speech. Formant frequencies can be utilized to distinguish

vowel sounds, identify speaker-specific characteristics, and detect differences between synthetic and natural speech. Consequently, formant analysis plays a crucial role in speech recognition, speaker identification, and synthetic speech detection systems. The primary function of speech is the transmission of information. According to Claude Shannon's information theory, every message conveys a certain amount of information, and the volume of transmitted information is measured in bits. In speech communication, this information is conveyed through analog acoustic waveforms. Using a microphone, the acoustic signal is converted into an electrical signal, which subsequently undergoes analog and digital processing stages. As a result, the signal can be transmitted, recorded, stored, or reproduced on various devices. These technologies have formed the foundation for the development of telephone communication systems, mobile devices, Internet-based communications, and multimedia applications. Today, automatic speech recognition and synthetic speech detection systems are becoming increasingly sophisticated. Advances in deep neural networks, transformer-based architectures, and generative artificial intelligence technologies have enabled the creation of highly realistic synthetic speech. As a consequence, distinguishing between genuine human speech and artificially generated speech has become significantly more challenging. Therefore, contemporary research focuses on the extraction of informative acoustic features, the development of noise-robust algorithms, and the design of intelligent real-time speech processing and recognition systems. Speech analysis and recognition systems commonly employ syntactic, semantic, and acoustic constraints. Human listeners reconstruct speech not only through auditory perception but also by relying on prior knowledge and contextual information. In contrast, machine-based recognition systems perform this process using mathematical models and statistical algorithms. For this reason, the development of speech processing systems extensively relies on acoustic models, language models, and probabilistic decision-making algorithms to achieve accurate and reliable recognition performance. The continued advancement of speech technologies is making human-machine interaction increasingly natural, intuitive, and efficient. At the same time, issues related to voice security, biometric authentication, and synthetic speech detection are gaining growing importance in both research and practical applications. Studies focused on digital speech signal processing, the extraction of informative acoustic features, and the analysis of speech using intelligent algorithms play a crucial role in enhancing the performance, reliability, and security of future intelligent systems. As speech-based technologies become more deeply integrated into everyday life, the development of robust and accurate speech processing methods will remain a key direction for scientific and technological innovation.

Table 1.1.

Digital Speech Signal Processing Systems and Their Primary Functions

System	References	Description
DSP Methods	1	Analysis of speech signals by dividing them into short-term quasi-stationary segments (5-100 ms) and examining their spectral amplitudes and phonetic characteristics.
Speech Synthesis and Production	2	Generation of speech waveforms based on vocal tract models, including the synthesis of voiced, plosive, and fricative sounds.
Biometric Recognition	3	Identification and verification of individuals using speech signals for biometric authentication and security applications.

Speech Coding and Transmission	4	Encoding speech while preserving its informational content and representing it as a sequence of bits for storage or transmission through communication channels.
Speech Enhancement and Modification	5	Reduction of noise in speech signals, modification of voice characteristics, and improvement of speech quality without altering the underlying message.
Automatic Speech Recognition (ASR)	6	Decoding acoustic waveforms into textual representations using machine-based algorithms and syntactic constraints.

Speech is a complex acoustic signal that conveys not only linguistic information but also additional characteristics such as a speaker's age, gender, emotional state, and pronunciation patterns. At the same time, speech signals may be affected by background noise, stress variations, accent differences, and environmental conditions, which can degrade signal quality and complicate analysis. Therefore, one of the key challenges in speech processing is the identification and extraction of informative acoustic features that enable the accurate discrimination and recognition of different phonemes under varying conditions.

References

1. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1980. – Vol. 28(4). – P. 357–366.

2. Rabiner L., Juang B. H. Fundamentals of Speech Recognition. – New Jersey: Prentice Hall, 1993.

3. Oppenheim A. V., Schafer R. W. Discrete-Time Signal Processing. – 3rd ed. – Pearson, 2010.

4. Huang X., Acero A., Hon H. W. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. – New Jersey: Prentice Hall, 2001.

5. Kinnunen T., Li H. An overview of text-independent speaker recognition: From features to supervectors // Speech Communication. – 2010. – Vol. 52(1). – P. 12–40.

6. Reynolds D. A. An overview of automatic speaker recognition technology // IEEE International Conference on Acoustics, Speech, and Signal Processing. – 2002. – P. 4072–4075.

7. Todisco M., Delgado H., Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification // Computer Speech & Language. – 2017. – Vol. 45. – P. 516–535.