IRSTI 16.21.37 https://doi.org/10.26577/EJPh202520047

G.T. Kussepova[1], R.Zh. Kondybaeva[2]*, K.A. Chingissova[2]

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: kondybaeva.raushan85@gmail.com

COMPARATIVE ANALYSIS OF PROSODIC CHARACTERISTICS OF SPONTANEOUS AND SYNTHESIZED SPEECH (Based on Kazakh and English Ted Talks Video Materials)

The study aims to conduct an instrumental-comparative analysis of the prosodic characteristics of spontaneous (based on TED Talks materials) and synthesized speech in Kazakh and English. The paper examines existing prosody research approaches and an acoustic analysis of key prosodic parameters (pitch frequency, intensity, and tempo) for spontaneous and synthesized speech types. For the comparative analysis, a corpus was developed, containing 10 speech excerpts drawn from TED Talks each in Kazakh and English, which were then transcribed and converted into audio files using modern speech synthesis systems. The acoustic analysis was conducted using PRAAT software and own proprietary software, Pro-AG-2025 (protected document No. 58731, dated May 27, 2025). This article formulates a hypothesis that spontaneous speech is characterized by greater variability in prosodic features, while synthesized speech differs from natural speech in acoustic and prosodic features. The instrumental analysis results confirm that synthesized speech, despite its structural conformity, retains a set of parameters that allow it to be reliably differentiated from spontaneous speech in increased amplitude uniformity and frequency contours, the absence of stochastic variations, and a simplified rhythmic-pause pattern. The obtained data are of practical significance for the further improvement of speech synthesis algorithms, increasing the degree of naturalness, and optimizing the communicative effectiveness of media applications.

Keywords: spontaneous speech, synthesized speech, prosody, acoustic parameters, tonality, pitch frequency.

Г.Т. Кусепова[1], Р.Ж. Кондыбаева[2]*, К.А. Чингисова[2]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
*e-mail: kondybaeva.raushan85@gmail.com

Спонтанды және синтезделген сөйлеістің просодикалық сипаттамаларын салыстырмалы талдау (қазақ және ағылшын тілдеріндегі Ted Talks бейне материалдары негізінде)

Зерттеудің мақсаты – қазақ және ағылшын тілдеріндегі спонтанды (аудиожазба материалдарына негізделген) және синтезделген сөйлеістің просодикалық сипаттамаларына инструменталды және салыстырмалы талдау жүргізу. Мақалада просодика мәселесін зерттеуде кеңінен қолданылатын әдіс-тәсілдер қарастырылған. Сөйлеу материалының негізгі просодикалық параметрлеріне (тон жиілігі, қарқындылығы және темпі) акустикалық талдау жүргізілген. Салыстырмалы талдау үшін ағылшын және қазақ тілдерінде әрқайсысы 10 TED Talks бейнебаяндамасы бар корпус жинақталды, олар кейін транскрипцияланып, заманауи сөйлеу синтезі жүйелерін пайдалана отырып, аудиофайлдарға түрлендірілді. Акустикалық талдау Praat бағдарламасы және біз әзірлеген ProAG-2025 бағдарламасы (қорғалған құжат № 58731, 2025 жылғы 27 мамыр) арқылы жүргізілді. Бұл мақалада синтезделген сөйлеіске қарағанда спонтанды, яғни табиғи сөйлеістің просодикалық сипаттамалары біршама өзгеше деп, ал синтезделген сөйлеіс табиғи сөйлеістен статистикалық тұрғыдан маңызды акустика-просодикалық ерекше белгілерге ие деген болжам айтылады. Инструменталды талдау нәтижелері синтезделген сөйлеудің құрылымдық сәйкестігіне қарамастан, оны табиғи сөйлеуден сенімді түрде ажыратуға мүмкіндік беретін параметрлер жиынтығын сақтайтынын растайды: амплитудалық және жиілік контурларының біркелкілігінің артуы, стохастикалық вариациялардың болмауы және жеңілдетілген ырғақты-үзіліс үлгісінің болу және т.б. Алынған деректер сөйлеуді синтездеу алгоритмдерін одан әрі жетілдіру, табиғилық дәрежесін арттыру және медиа қолданбаларының коммуникативтік тиімділігін оңтайландыру үшін практикалық маңызға ие.

Түйін сөздер: спонтанды сөйленіс, синтезделген сөйленіс, просодика, акустикалық параметрлер, тоналдылық, дыбыс жиілігі.

Г.Т. Кусепова[1], Р.Ж. Кондыбаева[2*], К.А. Чингисова[2]
[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
*e-mail: kondybaeva.raushan85@gmail.com

Сравнительный анализ просодических характеристик спонтанной и синтезированной речи (на видеоматериалах TED TALKS на казахском и английском языках)

Целью данного исследования заключается в осуществлении инструментально-сравнительного анализа просодических характеристик спонтанной (на материалах подкастов) и синтезированной речи на казахском и английском языках. В работе рассмотрены существующие исследовательские подходы к изучению просодии, а также выполнен акустический анализ основных просодических параметров (частоты основного тона, интенсивности и темпа) для указанных типов речевого материала. Для сравнительного анализа сформулирован корпус, включающий по 10 выступлений TED Talks на английском и казахском языках, которые далее транскрибированы и преобразованы в аудиофайлы с применением современных систем синтеза речи. Акустический анализ проводился программой Praat и нами разразботанной программой ProAG-2025 (с охраняемым документом № 58731 от «27» мая 2025 года). В данной статье формулируется гипотеза, согласно которой спонтанная речь характеризуется большей вариативностью просодических показателей, тогда как синтезированной речь отличается от естественной по статистически значимым акустико-просодическим признакам. Результаты инструментального анализа подтверждают, что синтезированная речь, несмотря на структурную нормативность, сохраняет комплекс параметров, позволяющих надёжно дифференцировать её от естественной: повышенную равномерность амплитудного и частотного контуров, отсутствие стохастических вариаций, а также упрощённый ритмико-паузовый рисунок. Полученные данные представляют практическую значимость для дальнейшего совершенствования алгоритмов синтеза речи, повышения степени её естественности и оптимизации коммуникативной эффективности медийных приложений.

Ключевые слова: спонтанная речь, синтезированная речь, просодика, акустические параметры, тональность, частота основного тона.

Introduction

Speech is a complex, multi-level communication system, utilizing both traditional and new formats of oral content, including prepared and spontaneous speech or even synthetic speech generated by automatic synthesis systems (Cooper E. et al., 2024) in social media and news aggregators. In these settings, prosodic characteristics are a fundamental parameter, determining the perception of a message, its credibility, expressiveness, and emotional impact on the audience (Galdino J. C. et al., 2025). A comparative analysis of the spontaneous and synthesized speech prosodic features allows to identify the linguistic parameters that distinguish natural speech from its artificial counterparts (Gabler P. et al., 2023), revealing differences in intonation contours, pause distribution, temporal variability, and loudness dynamics. Furthermore, it assesses the degree to which modern speech synthesis systems correspond to natural prosodic organization, determining how fully synthesized speech reproduces the characteristic melodic-rhythmic, accent-stress, and expressive features of living human speech.

The study aims to conduct an instrumental comparative analysis of the spontaneous and synthesized speech prosodic characteristic.

The study objectives are as follows:

- to analyze existing instrumental approaches and methods to the spontaneous and synthesized speech prosody analysis;
- to use specialized software applications (PRAAT, ProAG-2025) to conduct a prosodic acoustic analysis (pitch frequency, intensity, tempo) for the spontaneous and sythesized speech;
- carry out statistical processing of the obtained data to identify differences and correlations between prosodic characteristics of each speech type.

Hypothesis 1. Spontaneous speech demonstrates the greatest variability of prosodic parameters (wide range of pitch frequencies) compared to synthesized speech.

Hypothesis 2. Synthetic speech, despite its high level of naturalness, statistically significantly differs from natural speech (spontaneous) in key prosodic parameters, especially in the area of intonation contours.

Materials and methods

To conduct a comparative analysis, a speech data corpus was compiled, comprising 20 excerpts of spontaneous speech from TED Talks: 10 excerpts each for Kazakh and English. The sample was randomly selected; the topic of the talks was not controlled, as the primary goal was to obtain spontaneous speech samples.

The collected materials were transcribed using our proprietary ProAG-2025 program (patented document no. 58731, May 27, 2025). The application is written in Python. Its functionality includes audio file extraction, time-slice extraction, and automatic transcription using the Whisper model, which provides highly accurate segmentation and speech recognition. After transcription, the text transcriptions of natural fragments were converted to audio using two modern text-to-speech (TTS) systems:

- System 1 (for Kazakh): https://freereadtext.com/ru/text-to-speech/kazakh-kazakhstan;
- System 2 (for English): https://surl.li/nyjdyk (presumably another commercial or open-source service).

Before analysis, all speech fragments were pre-normalized for loudness and converted to a unified audio signal format. To ensure data comparability, the recording quality was checked, and noise artifacts and non-speech segments were removed.

Instrumental analysis of all speech fragments was conducted using Praat. The following acoustic-prosodic parameters were measured for each fragment: pitch frequency (F0), intensity, temporal characteristics, and segmentation. All fragments were segmented into sentences and words to measure prosodic contours. Additional acoustic analysis was conducted using ProAG -2025 , which integrates the librosa , soundfile, and matplotlib libraries .

Literature review

Kane et al. J. (2024) added realistic prosody using machine learning to enhance the authenticity and naturalness of a synthesized voice. Analysis of key prosodic elements at the syllable level was performed using the PRAAT program, extracting parameters of fundamental frequency, amplitude and intensity, duration and position of the syllable in a word or phrase, as well as the duration of the pause before and after the syllable. Training was carried out with a multiple- input, single-output LSTM. The mean squared error (MSE) was used as the loss function. The trained model was used to transform monotonous speech key elements into dynamic prosodic features, enhancing the naturalness and expressiveness of the synthesis. The average MOS score was 2.98, and the median was 3 on a scale from 1 (robotic speech) to 5 (natural, authentic speech). These results demonstrate the high effectiveness of the approach and its potential to generate more natural-sounding and authentic speech, even with a limited amount of data and LSTM models. During the transformation process, an increase in the overall loudness of the synthesized voice was observed: the median value increased from 51 dB to 65 dB, reflecting more expressive and dynamic speech production.

Correctly conveying the prosodic characteristics of utterances remains the most challenging issue in speech synthesis. In a study, O'Mahony J. et al. (2022) conducted a comparative analysis of the synthesized speech prosodic characteristics of two models: a baseline model trained exclusively on monologue speech and a model trained on a mixed corpus including both monologue and spontaneous dialogue data from podcasts. The results of the perceptual experiment showed that when synthesizing interrogative sentences, the DataMix model was perceived by listeners as significantly more conversational compared to the baseline model (significant intercept, $p < 0.01$). In a second preference test, in which subjects were asked to choose a more natural option, the DataMix model also demonstrated a statistically significant advantage in generating questions ($\beta = 0.44$, $p < 0.01$). When synthesizing response utterances, no significant differences were found between the models. The effects estimated on the basis of listener preferences did not reach statistical significance ($\beta = -0.17$, $p = 0.09$; $\beta = -0.21$, $p < 0.06$). Analysis of mean quality ratings (Mean Opinion Score (MOS) did not reveal a significant main effect of either model or sentence type.

Thorson J.C. and Morgan J.L. (2021) conducted a systematic analysis of prosodic structure in spontaneous English speech to identify prosodic vocabulary, semantics (functions and relations), and syntax (combination rules), as well as a comparison of spontaneous speech with scripted speech. The data included two spontaneous speech corpora (CallHome, Santa Barbara Corpus) and three audiobooks (scripted speech). Intonation units were automatically extracted using tempo modulation as a boundary marker. To construct a prosodic vocabulary, unsupervised clustering of pitch and intensity contours in latent space was used, which

allowed the identification of approximately 200 typical patterns. Sequences of intonation unit pairs were analyzed using Markov logic, identifying stable pattern combinations that occur more frequently than random ($P < 0.0001$). Manual cluster analysis revealed context-dependent features and speaker attitudes: all clusters exhibited 2–5 recurring features, and 90% of clusters expressed a specific attitude in ≥50% of the corresponding intonation units. Comparison with scripted speech revealed that in professionally voiced audiobooks, there was no statistical correlation between clusters, confirming that rehearsal and following the text simplify prosodic structure.

Amanbaeva A.Zh. and Zhumabaeva Zh.T. (2017) substantiate a segmental-prosodic model necessary for creating a natural synthesis of Kazakh speech. The key element is the syntagma, which determines semantic segmentation and ensures the intelligibility of synthesized speech. The authors identify eight intonemes that form an intonation contour and reflect the type of utterance. For prosody modeling, the following key parameters are defined: fundamental tone frequency (F0), duration, pause, and amplitude. Natural synthesis requires precise differentiation between orthography and orthoepy, consideration of individual phonetic characteristics, and the inclusion of intonemic models in the algorithm. The authors emphasize the lack of automated synthesis for the Kazakh language and the need to develop a model based on modern theories of intonation.

A study by Bazarbayeva Z.M. (2022), examining the prosodic means of spoken English discourse, demonstrates that prosody and spoken language function as independent grammatical levels. The informational structure of discourse is considered a component of semantics, and prosodic parameters are elements integrated into the phonological system. Prosody performs three key functions: the melodic contour marks the communicative type and speech act; pauses structure the flow of speech, breaking it into information units; accentual emphasis sets the focus of the utterance and organizes its internal structure. The phonological approach, which assumes an abstract representation of intonation in the form of tonal structures (in particular, in the autosegmental-metric model), is recognized as the most productive. This allows for the description of the systemic connections between intonation, grammar, and discourse. Paralinguistic parameters, including the general tone level, convey the speaker's emotional and pragmatic attitude and are characterized by gradation, reflecting the degree of expression of states such as respect or excitement.

Defining the segmental and prosodic parameters of Kazakh speech is a key requirement for developing an intonationally adequate synthesis (Bazarbayeva, 2025). Segmental and prosodic elements ensure the semantic unity of an utterance and act as differentiating features. Syntagma is defined as the basic unit of synthesis, requiring mandatory identification during text markup. For automatic intonation modeling, it is necessary to formalize eight Kazakh intonemes (completeness, incompleteness, general and specific questions, categorical and polite imperatives, exclamations, and insertions) that correlate with punctuation marks and determine the communicative type of utterance. The naturalness of synthesized speech is ensured by taking into account orthoepy and the individual prosodic characteristics of the speaker.

The research by Berdalieva R.Sh. (2022) explores the paralinguistic characteristics of Kazakh speech (intonation, tempo, rhythm, and timbre) as a means of expressing the individual, social, and ethnocultural characteristics of the speaker. The Kazakh language is distinguished by specific phonetic features and an accentuated final syllable, while consonant realization is subordinated to vowel harmony. Melody (F_0) plays a key role in distinguishing communicative types of utterances and organizing the logical structure of speech. Voice parameters (volume, pitch) perform a social function: quiet speech expresses respect, while loud speech is associated with authority. It is also noted that Kazakhs, on average, speak louder than other Turkic and Asian peoples, which is associated with the historical conditions of their nomadic lifestyle. Some types of utterances have fixed vocal characteristics (e.g., "Oybay!" – a high voice, commands – a firm voice, parting words – a soft voice), which emphasizes the cultural determinacy of paralinguistic means.

Another study conducted by Taldibayeva M. (2016) is aimed at a systematic description and theoretical substantiation of labial vowel harmony as a key phonological feature of the Kazakh language. Kazakh speech is characterized by labial vowel segmentation. The vocalism of the language is formed by three vowel syngemes: a high vowel (4 allosingemes), a low vowel (2 allosingemes), and a diphthong-type vowel (3 allosingemes), which form a trivocalic pattern. It is shown that synharmonically soft mid vowels and hard back vowels are articulated near the central row. Historical analysis revealed

that the most accurate data on labial vowel harmony were obtained in the mid-19th century thanks to perceptual observations not limited by orthographic norms.

Taken together, the reviewed studies highlight that accurate modeling of segmental, suprasegmental, and paralinguistic features to achieve intelligibility, naturalness, and communicative adequacy across different languages, including Kazakh and English.

Results and discussion

An acoustic-prosodic comparison of two speech samples (spontaneous and synthesized) demonstrates consistent differences between spontaneous speech and a synthesized version formally imitating conversational mannerisms. Figure 1 shows an example of speech analysis in English. During speech synthesis, the speaker's style was chosen to be similar to spontaneous speech.

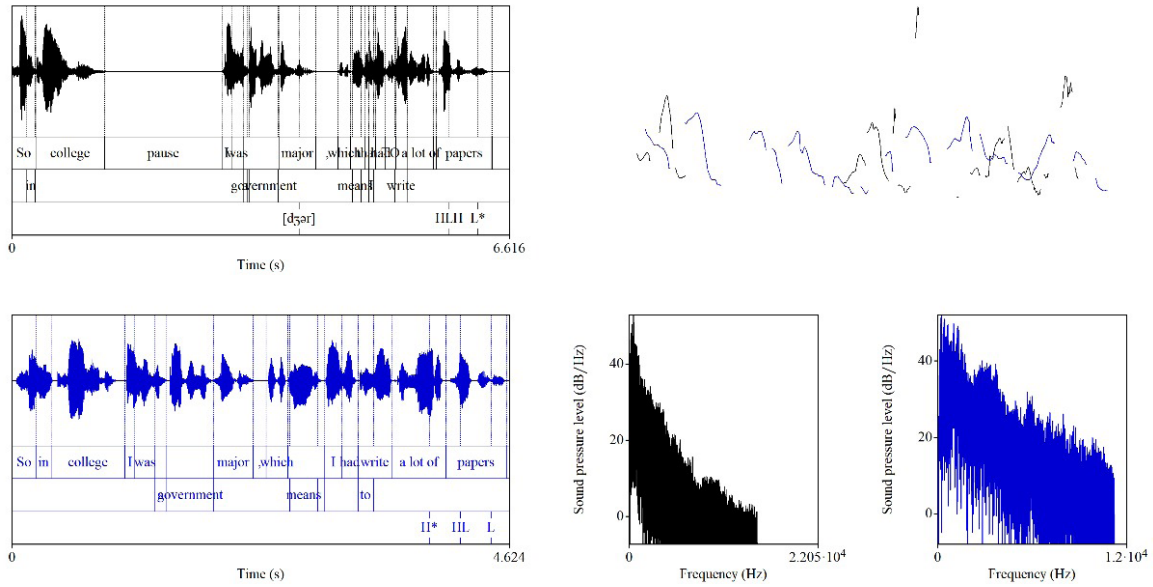

Figure 1 – Analysis of natural and synthesized speech in English

Natural speech (black) is characterized by a pronounced irregularity of the amplitude contour. Intensity fluctuations within phrases vary widely, reflecting the natural mechanisms of utterance planning, emotional modulation, and the physiological characteristics of breathing. The waveform displays an alternation of short micropauses and longer semantic stops, which is typical of unprepared speech. Intonation exhibits local and global fluctuations in the fundamental frequency, including microvariations (jitter), rapid transitions between intonation levels, and characteristic F0 lowerings at the ends of phrases, sometimes accompanied by creative phonation. The harmonic structure in real speech is irregular: the HNR constantly changes depending on the type of segment, and the high-frequency noise components of sibilants and fricatives are clearly and unpredictably present.

The synthesized fragment (blue) exhibits a noticeably smoother acoustic organization. Amplitude peaks are more evenly distributed, transitions between phrases have a fixed duration, and pauses lack the stochastic nature inherent in spontaneous speech. The fundamental frequency contour is characterized by less variability and order: intonation movements seem programmed, smooth, and devoid of micro-fluctuations. This creates the effect of a grammatically correct, but prosodically simplified utterance. Phonation in the synthesis most often remains within a stable modal register; breath sounds, irregular harmonic bursts, and signs of vocal tension are virtually absent (Teixeira, 2013). In the spectral domain, the synthesis exhibits a more pronounced and uniform power drop with increasing frequency, indicating filtered smoothing and insufficient realization of high-frequency components, especially important for the natural perception of noisy consonants. Compared to a real recording, the synthesized signal has a higher predictability of spectral slope (Sisman B. et al., 2020) and less amplitude differentiation of fricative consonants.

Articulatory features also differ. In natural speech, vowel reduction in unstressed positions, asymmetrical formant transitions, and increased coarculation variability are observed. Consonants weaken or strengthen depending on the tempo and communicative task, which manifests itself in the instability of their spectral characteristics. In the synthesized version, vowels are more often realized as complete articulatory targets regardless of their position in the word, and coarculation transitions acquire an "idealized" character: overly clear, symmetrical, and rhythmically precise. This lends the speech a certain technical clarity not characteristic of untrained, live speech. The temporal structure of the synthesized fragment exhibits a more uniform tempo, a low frequency of disfluencies, and the absence of restarts, further enhancing the sense of syntheticity.

A spectral comparison confirms the general difference: the spectrum of natural speech exhibits localized power fluctuations associated with articulatory features and noise components, while the spectrum of the synthesized signal is smoothed, with a predictable decrease in energy in the high-frequency range. This distribution indicates insufficient implementation of the noise component and a lack of microarticulatory variations in the synthesis model.

Taken together, the obtained data demonstrate that synthesized speech, despite its structural correctness, retains a number of acoustic and prosodic features that distinguish it from real speech.Uniformity of the amplitude contour and F0, the absence of stochastic variations and creative elements, insufficient articulatory reduction, a smoothed spectrum, and a simplified rhythmic-pause pattern are the main characteristics of synthesized speech. These features can be used as diagnostic markers of synthesized speech and simultaneously serve as benchmarks for improving speech generation algorithms.

Acoustic-prosodic analysis of the Kazakh speech materials shows systematic differences between natural spontaneous speech and the synthesized version as well. The specificity of Kazakh phonetics makes these differences especially clear.

In Figure 2, we presented an example of spontaneous (black) and synthesized (blue) speech examples. In a natural fragments, the amplitude structure is irregular. The speaker's speech intensity fluctuates both within and between syntagmas. This reflects the free rhythm organization of the pronunciation. The presence of meaningful pauses and micro-stops associated with utterance planning show that speaker intentionally pauses longer in some places. Variability in vowel duration is clearly visible, especially in positions before sonorants and in strong syntactic positions. Reduction processes show vowels in weak positions, slightly shortened or even weakened, which is evident in the oscillatory structure of the formants with the irregularity of the amplitude. Consonant transitions have a pronounced coarculation character by softened segments and transitions such as -ды, -нда demonstrate asymmetrical formant trajectories.

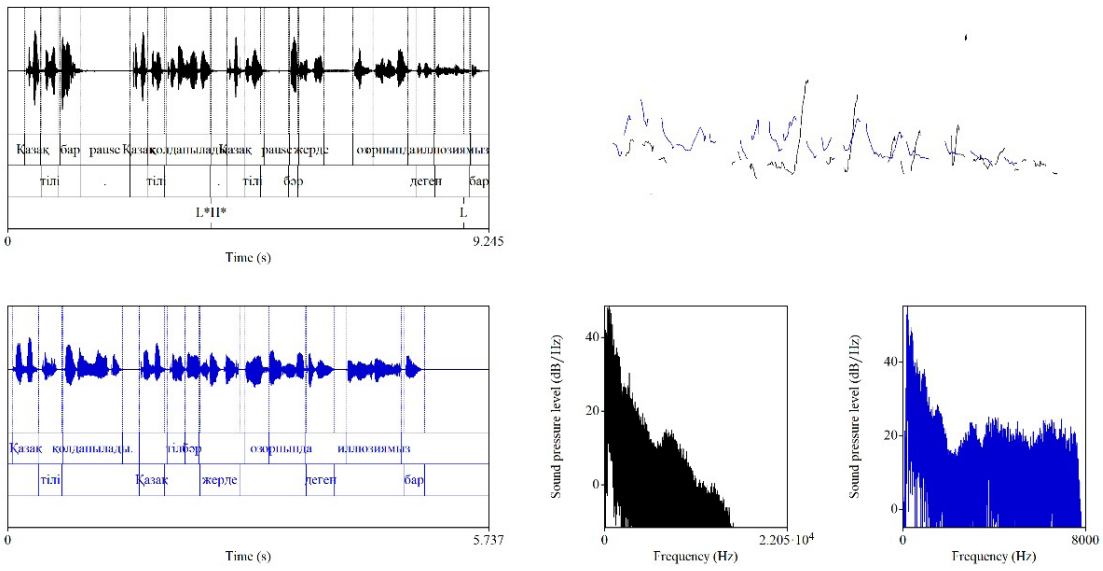

Figure 2 – Analysis of natural and synthesized speech in the Kazakh language

The intonation contour of real speech in almost all examples show a wide range of F0 fluctuations. Rising and falling tones follow an uneven pattern that reflect the communicative structure of the utterance demonstrating the speaker's subjective attitude. For example, in presented Figure 2, the contour is characterized by sharp jumps in F0 and intonation dominants, as seen in sections with *the phrases "қолданылады", "өз орнында", and "иллюзиямыз"*. Micro-oscillations in the fundamental frequency (jitter) are observed, as well as natural variations in intensity (shimmer), which are markers of live phonation (Teixeira, 2013). The tempo is irregular: there are slowings before key words and speedings up in less significant segments.

Spectral analysis of natural speech reveals a rich high-frequency content through the energy of noise consonants (к, з, ж) is unevenly distributed, and the spectrum clearly displays peaks associated with articulatory difficulties. These high-frequency components indicate natural turbulent noise at the points where fricatives are formed, which cannot be fully reproduced by synthesizers.

In the synthesized fragment, the speech structure is significantly more organized, as presented in synthesized speech in English. The amplitude of the signals is evened out, the intensity remains almost constant, even where prosodic accents are observed in natural speech. The tempo of the synthesis is rhythmically regular, pauses are evenly spaced based on the full stops, and do not reflect the actual semantic division of the utterance (Yu, 2025). Reduction processes are virtually absent: vowels retain their full duration regardless of position, making the speech formally clear but prosodically unnatural, where reduction is minimal but temporal variability is significant.

The intonation contour of synthesized speech is smoothed. F0 moves primarily along predetermined trajectories, without micro-variations and without the natural "dips" or rapid rises characteristic of human speech, which appeared in "қолданылады". Intonation movements are extended and "mathematically smooth", leading to a feeling of monotony. The tone realization in the words appears simplified.

The spectral structure of the synthesized version shows typical high-frequency smoothing: noise consonants lack their natural turbulent spectrum, making the synthesized speech sound softer than natural speech. High-frequency energy is reduced more uniformly, indicating filtering and insufficient articulatory variability in the model.

Overall, the comparison shows that synthesized Kazakh speech does not reproduce key parameters of natural speech flow: irregularity of intensity, vowel duration variability, coarculation asymmetry, a rich spectrum of fricatives, F0 microdynamics, and natural rhythmic and intonational patterns. These differences can serve as important criteria for assessing the naturalness of synthesized Kazakh speech and as benchmarks for further improving the quality of synthesis models.

Text-to-speech conversion was carried out on the following resources: (www.freereadtext.com; www.mureka.ai).

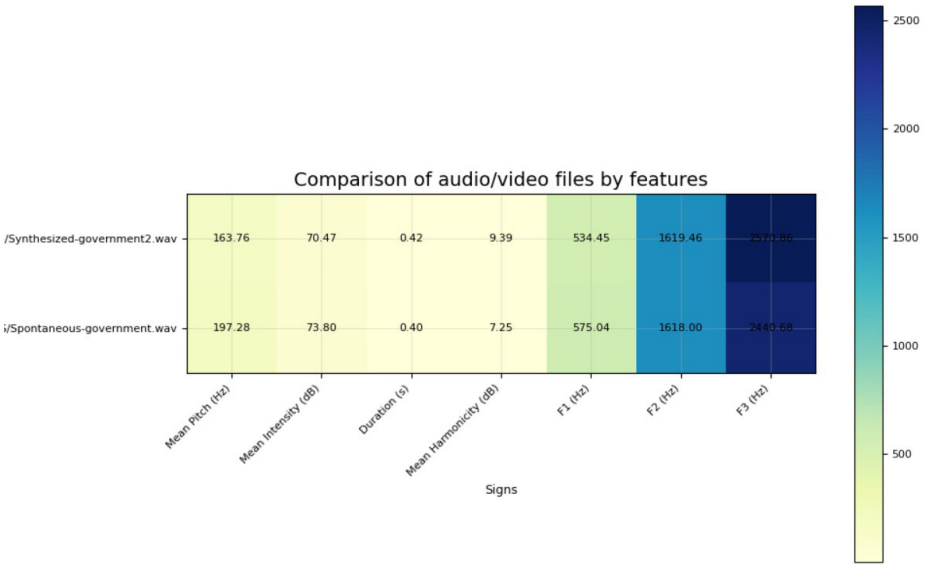

Figure 3 – Heat map of the comparison of the acoustic characteristics of synthesized and spontaneous speech " government " based on the main features

Analysis of the presented data reveals significant differences between the spontaneous and synthesized speech. After comparing the prosodics of sentences, we decided to investigate if the words contain any similarities of differences according to the reproduction type. For example, the phrase "government" demonstrates different acoustic and prosodic parameters. Average F0 values indicate that synthesized speech is characterized by a higher F0 fundamental frequency, while spontaneous speech, in contrast, exhibits a lower one. The intensity of spontaneous speech is also lower as its F0, while the sythesized option show slightly greater mean than expected. It might be the result of the spectral richness of the signal. In most of the words, synthesized speech exhibits more stable amplitude due to normalization, resulting in a moderate but perceptually distinguishable difference. All words are pronounced according to the algorythm provided to the application. Thus, a slight increase in the duration of the synthesized audio probably indicates the tendency of TTS models to generate slightly slower and smoother articulatory trajectories. The harmonicity index is significantly higher in synthesized speech, reflecting the dominance of the periodic component characteristic of modern vocoders such as WaveNet/HiFi-GAN. In spontaneous speech, most of the examples differ greatly, which ends with the decrease in harmonicity. We suggest it due to phonetic irregularities, respiratory noises, and natural vocal fold vibration. Formant analysis reveals the accuracy of the synthesized speech's reproduction of F2 values and duration, but also reveals deviations in F1 and, especially, F3. Taken together, these features confirm that synthesized speech approaches natural speech in key parameters (Wester, 2016) while retaining spectral and intonational markers of artificial origin.

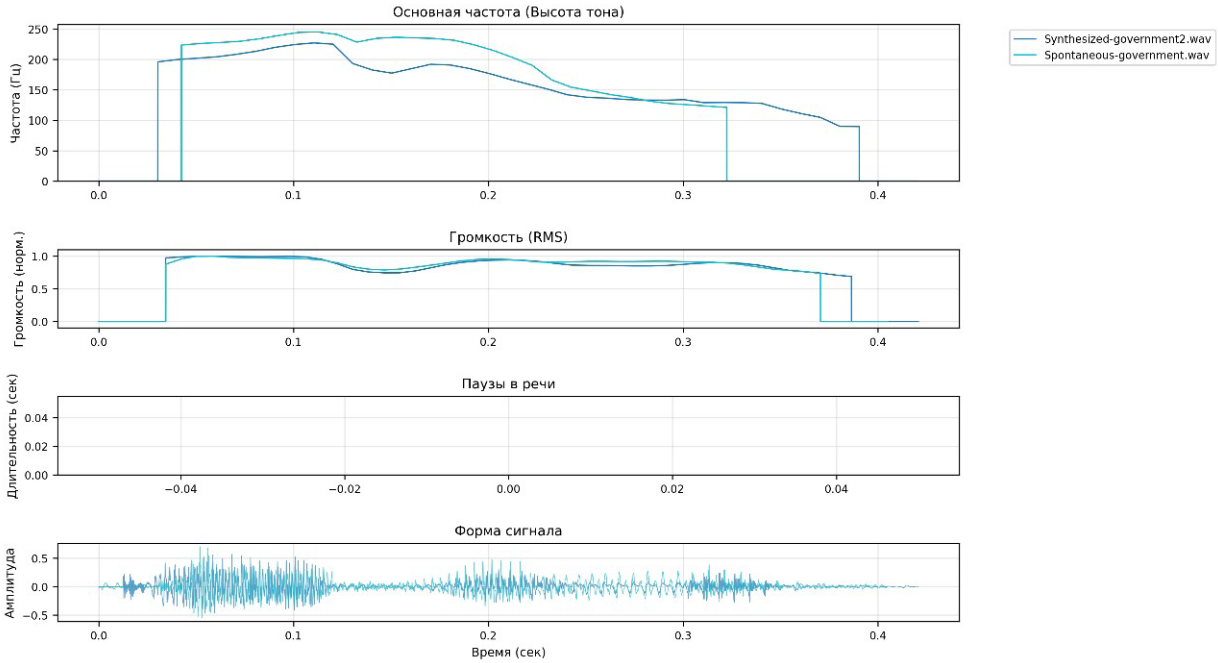

Figure 4 – Dynamics of changes in pitch, volume and signal shape of synthesized and spontaneous speech "government"

In Figure 4, analysis of the fundamental frequency plot reveals consistent differences between synthesized and spontaneous speech. Initially, we expected that the synthesized signal exhibit intended smooth linear F0 in words, following application algorithm, implemented into the system. In the fundamental portion, both curves have a similar general contour; however, synthesized speech has a significantly smoother profile, while spontaneous speech exhibits microvariations associated with natural jitter, which is seen in 0.1-0.2 ms. The synthesized F0 terminates abruptly with smooth curve as seen in 0.2-0.3ms. Overall, synthesized speech exhibits a stable and predictable intonation pattern,

while spontaneous speech exhibits natural variability. The loudness dynamics (RMS) plot reveals significant differences: synthesized speech is characterized by a flat loudness level with minimal variability due to digital normalization and the absence of natural noise. Spontaneous speech, in contrast, exhibits wide amplitude dynamics associated with natural airflow fluctuations and phonation phases, which is seen in 0.2-0.4 ms. With comparable maximum RMS values, the synthesized signal exhibits abrupt transitions to zero at the beginning and end of the recording, a typical effect of digital clipping in TTS. Waveform analysis reveals key differences between synthesized and spontaneous speech. The synthesized signal exhibits a sharp initial increase in amplitude and a uniform, structurally ordered waveform in the central region, reflecting the absence of natural microvariations. Spontaneous speech, in contrast, contains initial noise components, previous word sound, as well as more pronounced amplitude fluctuations due to the biomechanics of phonation. The synthesized signal terminates abruptly, while natural speech gradually fades (0.3-0.4ms). These differences demonstrate a fundamental discrepancy between artificially generated voice and natural human phonation, conditioned by both the architecture of neural TTS models and the physiological characteristics of natural speech.

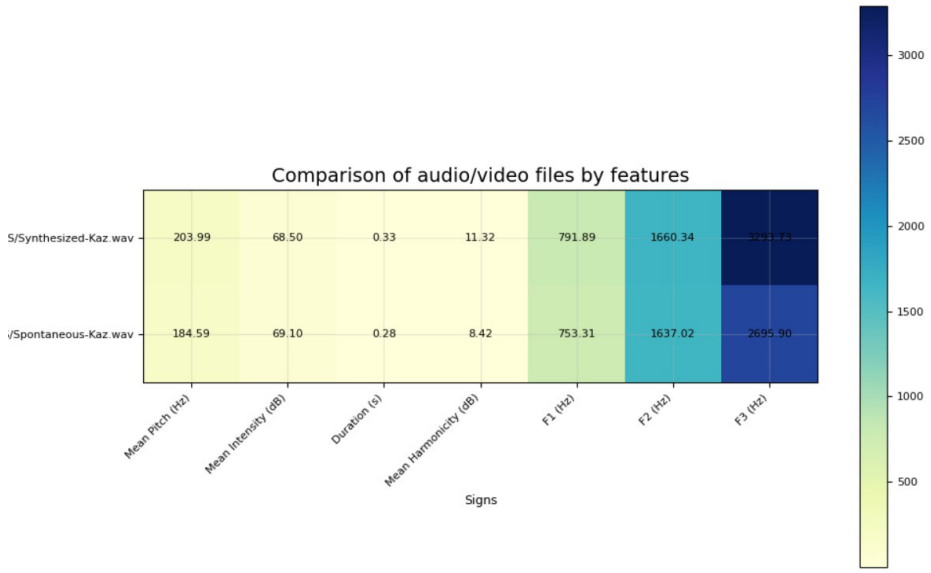

Figure 5 – Heat map of comparison of acoustic characteristics of synthesized and spontaneous speech "Kazakh" based on the main features

An acoustic-prosodic analysis of synthesized and spontaneous Kazakh speech revealed significant differences in a number of key parameters as well. Average pitch frequency values show that synthesized speech has a higher F0 (203.99 Hz) compared to spontaneous speech (184.59 Hz), which indicates the tendency of TTS systems to form a raised and more expressive intonation pattern. Natural spontaneous speech is characterized by lower F0 values, reflecting relaxed articulation and natural intonation fluctuations; a difference of almost 20 Hz indicates differences in prosodic dynamics. Intensity analysis reveals minimal differences. Spontaneous speech is slightly louder (69.10 dB) compared to synthesized speech (68.50 dB).

As we expected, synthesized speech also exhibits typical features of TTS models in terms of duration. It is longer (0.33 s versus 0.28 s), which is due to slower transient processes, smoother articulation, and the absence of natural reductions. We observed difference in the harmonicity index, where the high HNR value in synthesized speech (11.32 dB) reflects the almost complete absence of respiratory phonation irregularities, while the lower value for spontaneous speech (8.42 dB) corresponds to the natural variability of the voice signal. Moreover, formant analysis revealed characteristic spectral discrepancies. Synthesized speech exhibits an increased F1 (791.89 Hz), indicating a more open vowel articulation, and a significantly elevated F3 (3251.73 Hz), which indicates excessive timbre "purity" and smoothed resonance characteristics. Moreover, the F2 values (1660.34 Hz for synthesized and 1637.02 Hz for spontaneous speech) show a relatively ac-

curate transmission of the front-back position of the tongue by the TTS model.

Overall, as a result, we confirmed that synthesized Kazakh speech adequately reproduces the basic acoustic characteristics of the natural signal, but retains a number of typical differences: elevated F0, increased duration, greater harmonicity, elevated F3, and greater vowel openness.

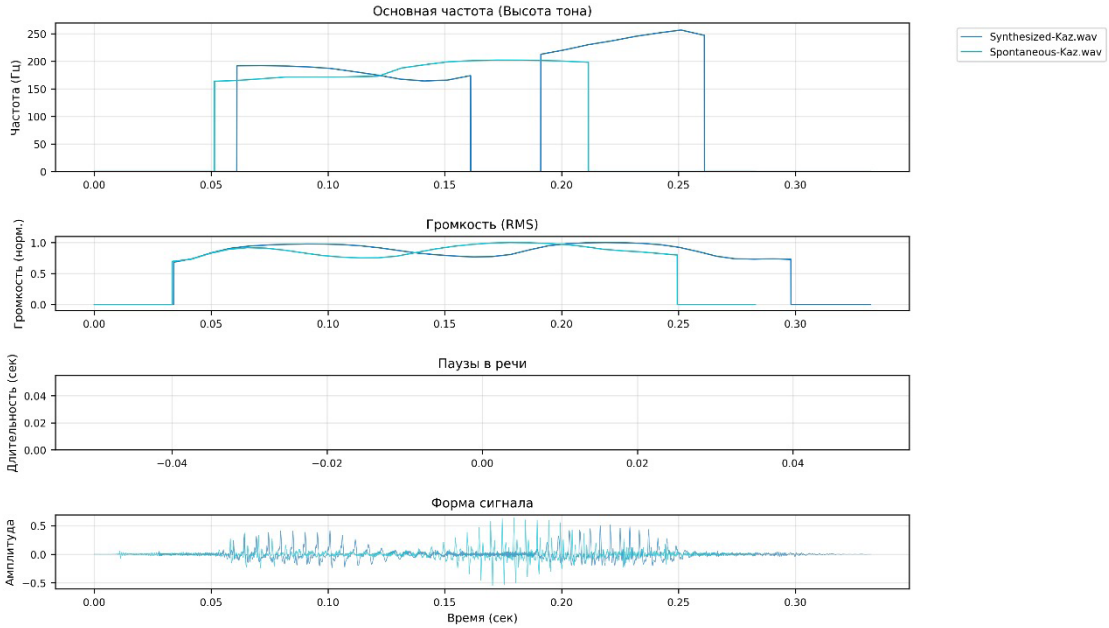

Figure 6 – Dynamics of changes in pitch, volume and signal shape of synthesized and spontaneous speech "Kazakh"

We decided to conduct similar word analysis towards Kazakh speech excerpts. As a result, F0 analysis shows that synthesized speech begins at a fixed frequency, has a smooth contour, and an abrupt end, whereas spontaneous speech develops gradually, contains natural micro-variations, and ends with a smooth frequency decline. RMS analysis shows that synthesized speech has a stable amplitude without noise or spikes, with an abrupt onset and end, whereas spontaneous speech exhibits natural loudness variations and smooth transitions. Waveform analysis shows that synthesized speech is free of initial noise (leaving out the sound [z]), has a regular, symmetrical waveform with minimal jitter and shimmer, and an abrupt end, whereas spontaneous speech (replaces [z] sound to similar, presenting a waveform in oscillogram) contains noise, amplitude fluctuations, and a smooth decay.

Table 1 – Comparative characteristics of the acoustic parameters of synthesized and spontaneous speech in English and Kazakh languages.

Parameter	Synthesized speech in English	Synthesized speech in Kazakh	Spontaneous speech in English	Spontaneous speech in Kazakh
Fundamental frequency	Smoothed contours, insufficient transmission of intonation jumps	Clear stepwise transitions, excessive morphemic regularity	High variability, sharp transitions	Smoother and more regular contours, less variation in amplitude
Volume	Strong smoothing, stress reduction is poorly conveyed	Even smoother, the volume is almost uniform	Pulsating dynamics, pronounced stress peaks	Smooth volume curve, low amplitude contrast
Signal form	The amplitude is more symmetrical, the natural rhythm is lost	Overly uniform, mechanically aligned	Strong amplitude variability	Relatively flat signal shape

Conclusion

A comprehensive instrumental analysis of the spontaneous and synthesized speech prosodic characteristics in English and Kazakh TED Talks revealed a number of significant differences, which generally confirms the previously proposed research hypotheses.

First, spontaneous speech exhibits a significantly higher degree of prosodic variability (in terms of F0, intensity, and tempo) compared to synthesized speech. This increased variability manifests itself in irregular amplitude and melodic contours, the presence of microfluctuations (jitter), elements of creative phonation, as well as the stochastic nature of pauses and temporal organization.

Secondly, synthesized speech, despite its stable segmental-temporal structure, exhibits statistically significant deviations from natural speech. Key markers of synthesized speech include a simplified rhythmic-pause pattern, a uniform distribution of amplitude peaks, a reduced level of F0 variability, as well as a smoothed spectral slope and insufficient realization of high-frequency noise components in both English and Kazakh language samples.

An analysis of Kazakh language data revealed that the presence of specific phonetic and prosodic characteristics, such as regular stress on final syllables and a developed system of intonemes, further worsens the differences between natural and synthesized speech. Current TTS systems for Kazakh do not fully account for the complex segmental and prosodic organization, resulting in simplified syntagmatic segmentation and reduced naturalness of intonation contours. These deviations are less noticeable when analyzing individual words, but are clearly evident at the sentence level and especially in the spectral characteristics of synthesized speech.

The practical significance of the obtained results lies in the fact that the identified acoustic-prosodic indicators of syntheticity can serve as benchmarks for optimizing speech generation algorithms. Improving the naturalness of synthesis requires the implementation of models capable of reproducing stochastic variations in F0 and intensity, accounting for articulatory vowel reduction in unstressed positions, more accurately modeling consonant noise components, and integrating ethnoculturally determined prosodic features, including the vowel harmony of the Kazakh language.

Prospects for further research include the development and implementation of a perceptual experiment aimed at quantitatively assessing the degree of naturalness of synthesized speech by listeners.

References

Cooper, E., et al. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology, Vol. 45*(4), P. 161–183. https://doi.org/10.1250/ast.e24.12

Galdino, J.C., et al. (2025). The evaluation of prosody in speech synthesis: A systematic review. *Journal of the Brazilian Computer Society, Vol. 31*(1), P. 466–487. https://doi.org/10.5753/jbcs.2025.5468

Gabler, P., Geiger, B. C., Schuppler, B., & Kern, R. (2023). Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition. *Information, Vol. 14*(2), Article 137. https://doi.org/10.3390/info14020137

Kane, J., Johnstone, M.N., Szewczyk, P. (2024). Voice synthesis improvement by machine learning of natural prosody. *Sensors, Vol. 24*(5), Article 1624. https://doi.org/10.3390/s24051624

O'Mahony, J., Lai, C., King, S. (2022). Combining conversational speech with read speech to improve prosody in text-to-speech synthesis. In *Proceedings of Interspeech 2022* (P. 3388–3392). ISCA. [Electronic resource]. URL: https://www.isca-archive.org/interspeech_2022/omahony22_interspeech.pdf (Date of use: 10.09.2025)

Thorson, J.C., & Morgan, J.L. (2021). Prosodic realizations of new, given, and corrective referents in the spontaneous speech of toddlers. *Journal of Child Language, Vol. 48*(3), P. 541–568. https://doi.org/10.1017/S0305000920000434

Аманбаева, А.Ж., Жұмабаева, Ж.Т. (2017). Создание синтеза речи с помощью просодических методов. *Universum: филология и искусствоведение,* Vol. 8(42). [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/sozdanie-sinteza-rechi-s-pomoschyu-prosodicheskih-metodov/viewer (Дата использования: 10.09.2025)

Базарбаева, З.М., Кожамсугирова, Б.О. (2022). Проблемы просодии и звучащего дискурса в английском языке [Issues of prosody and sounding discourse in the English language]. *Тілтаным,* Vol. 185(1), P. 3–14. https://doi.org/10.55491/2411-6076-2022-1-3-14

Базарбаева, З.М., Садык, Д., Аманбаева, А., Жұмабаева, З., Оспангазиева, Н. (2025). Segmental-prosodic foundations of Kazakh speech synthesis. *Eurasian Journal of Applied Linguistics, Vol. 11*(2), P. 69–80.

Бердалиева, Р.Ш. (2022). Характеристика звукового оформления казахской речи. *Тілтаным, Vol. 2,* P. 21–25.

Taldibayeva, M. (2016). Methodological analysis of the synharmonic composition of sounds in the Kazakh language. *Avrasya Uluslararası Araştırmalar Dergisi, Vol. 4*(9), P. 169–177.

Teixeira, J.P., Oliveira, C., Lopes, C. (2013). Vocal acoustic analysis – Jitter, shimmer and HNR parameters. *Procedia Technology, Vol. 9,* P. 1112–1122.

Sisman, B., et al. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29,* P. 132–157. https://doi.org/10.1109/TASLP.2020.3038524

Yu, J., Zihao, G., Li, C., Wang, Z., Yang, P., Chen, W., Yin, J. (2025). Eliciting implicit acoustic styles from open-domain instructions to facilitate fine-grained controllable generation of speech. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (P. 3679–3695).

Wester, M., Watts, O., Henter, G.E. (2016). Evaluating comprehension of natural and synthetic conversational speech. In *Speech Prosody 2016* (P. 766–770). [Electronic resource]. URL: https://www.research.ed.ac.uk/en/publications/evaluating-comprehension-of-natural-and-synthetic-conversational-/datasets/ (Date of use: 12.09.2025)

References

Amanbayeva, A.Zh., Zhumabayeva, Zh.T. (2017). Sozdanie sinteza rechi s pomoshiu prosodicheskih metodov [Speech synthesis using prosodic methods]. Universum: filologia i iskusstvovedenie [Universum: Philology and Arts]. Vol. 8 (42). [Electronic Resource]. URL: https://cyberleninka.ru/article/n/sozdanie-sinteza-rechi-s-pomoschyu-prosodicheskih-metodov/viewer (Date of use: 10.09.2025) (In Russian).

Bazarbayeva, Z.M., Kozhamsugirova, B.O. (2022). Problemy prosodii I zvuchashego diskursa v angliiskom yazyke [Issues of Prosody and Sounding Discourse in the English Language.]. Tiltanym. Vol. 185, Iss. 1, P. 3-14. https://doi.org/10.55491/2411-6076-2022-1-3-14 (In Russian).

Bazarbayeva, Z.M., Sadyk, D., Amanbayeva, A., Zhumabayeva, Z., Ospangaziyeva, N. (2025). Segmental-Prosodic Foundations of Kazakh Speech Synthesis. Eurasian Journal of Applied Linguistics. Vol. 11, Iss. 2, P. 69-80. (In Russian).

Berdalieva, R.Sh. (2022). Harakteristika zvukovogo oformlenia kazahskoi rechi. Tiltanym. Vol. 2, P. 21-25. (In Russian).

Cooper E. et al. (2024). A review on subjective and objective evaluation of synthetic speech. Acoustical Science and Technology. Vol. 45, Iss. 4, P. 161-183. https://doi.org/10.1250/ast.e24.12

Gabler, P., Geiger, B. C., Schuppler, B., Kern, R. (2023). Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition. Information.Vol. 14, Iss. 2, P. 137. https://doi.org/10.3390/info14020137

Galdino, J. C. et al. (2025). The evaluation of prosody in speech synthesis: a systematic review. Journal of the Brazilian Computer Society. Vol. 31, Iss. 1, P. 466-487. https://doi.org/10.5753/jbcs.2025.5468

Kane, J., Johnstone, M.N., Szewczyk, P. (2024). Voice synthesis improvement by machine learning of natural prosody. Sensors. Vol. 24, Iss. 5, P. 1624. https://doi.org/10.3390/s24051624

O'Mahony, J., Lai, C., King, S. (2022). Combining conversational speech with read speech to improve prosody in text-to-speech synthesis. Proceedings of Interspeech. ISCA. P. 3388-3392. [Electronic Resource]. URL: https://www.isca-archive.org/interspeech_2022/omahony22_interspeech.pdf (Date of use: 10.09.2025)

Sisman, B. et al. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 29, P. 132-157. https://doi.org/10.1109/TASLP.2020.3038524

Taldibayeva, M. (2016). Methodological analysis of the synharmonic composition of sounds in the Kazakh language. Avrasya Uluslararası Araştırmalar Dergisi. Vol. 4, Iss. 9, P . 169-177. (In Russian)

Thorson, J.C., Morgan, J.L. (2021). Prosodic realizations of new, given, and corrective referents in the spontaneous speech of toddlers. Journal of child language. Vol. 48, Iss. 3, P. 541-568. https://doi.org/10.1017/S0305000920000434

Teixeira, J.P., Oliveira, C., Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and HNR parameters. Procedia technology. Vol. 9, P.1112-1122.

Yu, J., Zihao, G., Li, C., Wang, Z., Yang, P., Chen, W., Yin, J. (2025). Eliciting Implicit Acoustic Styles from Open-domain Instructions to Facilitate Fine-grained Controllable Generation of Speech. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. P. 3679-3695.

Wester, M., Watts, O., Henter, G.E. (2016). Evaluating comprehension of natural and synthetic conversational speech. Speech Prosody. P. 766-770. [Electronic Resource]. URL: https://www.research.ed.ac.uk/en/publications/evaluating-comprehension-of-natural-and-synthetic-conversational-/datasets/ (Date of use: 12.09.2025)

Information about the authors:

Kussepova Gulzat Tungushbayevna – PhD, L.N. Gumilyov Eurasian National University (Kazakhstan, Astana, e-mail: kussepova_gt_2@enu.kz);

Kondybaeva Raushan Zhumakerimovna – PhD, Al-Farabi Kazakh National University (Kazakhstan, Almaty, e-mail: kondybaeva.raushan85@gmail.com);

Chingissova Kuralay Adilzhanovna – PhD student, Al-Farabi Kazakh National University (Kazakhstan, Almaty, e-mail: kuralay.cha@mail.ru).

Авторлар туралы мәлімет:

Кусепова Гульзат Тунгушбаевна – PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті (Қазақстан, Астана, e-mail: kussepova_gt_2@enu.kz);

Кондыбаева Раушан Жумакеримовна – PhD, Әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы, e-mail: kondybaeva.raushan85@gmail.com).

Чингисова Құралай Адилжановна – PhD докторанты, Әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы, e-mail: kuralay.cha@mail.ru).

Сведения об авторах:

Кусепова Гульзат Тунгушбаевна – PhD, Евразийский национальный университет имени Л.Н. Гумилева (Астана, Казахстан, e-mail: kussepova_gt_2@enu.kz).

Кондыбаева Раушан Жумакеримовна – PhD, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: kondybaeva.raushan85@gmail.com).

Чингисова Куралай Адилжановна – PhD докторант, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: kuralay.cha@mail.ru).

Date of receipt of the article: October 01, 2025.
Accepted: December 13, 2025.