

A.B. ABEN¹, D. GABDULLAULY²

Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan

¹ E-mail: arypzhan.aben@ayu.edu.kz*

² E-mail: gabdulladimas@gmail.com

DEVELOPING AN OPEN-SOURCE DATASET FOR SPEECH SOUND DISORDERS IN KAZAKH CHILDREN

Abstract. Speech sound disorders (SSD) in childhood affect intelligibility, literacy, social participation and emotional well-being, yet no annotated corpus of child speech exists for Kazakh, a low-resource language, which limits the development of diagnostic and therapeutic tools. This study describes the construction of a Kazakh child SSD dataset from 100 children aged 3–10 years (50 with SSD diagnosed by speech-language therapists and 50 typically developing controls). Each child completed a fixed elicitation protocol (30 single words, 10 phrases and 5 sentences) recorded at 44.1 kHz/16 bit under controlled clinical conditions; after segmentation and quality screening, the corpus contains 10,000 utterance-level audio files with IPA transcriptions, error annotations and participant metadata. Compared with controls, the SSD group showed higher mean F0, F1 and F2 (Cohen's $d = 0.44$ – 0.53), longer articulation duration ($d = 0.88$) and a markedly higher error rate (35% vs. 5%), dominated by substitutions and omissions of velar and uvular consonants. The dataset, preprocessing scripts and documentation are released at <https://doi.org/10.5281/zenodo.14589203> under the CC BY 4.0 (metadata and code) and CC BY-NC 4.0 (audio) licences, providing a basis for Kazakh-specific research on child speech development and for AI-based screening tools.

Keywords: speech sound disorders, Kazakh language, child speech corpus, phonological disorders, articulation disorders, low-resource languages, acoustic analysis.

A.B. АБЕН¹, Д. ГАБДУЛЛАҰЛЫ²

Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ.,
Қазақстан

¹ E-mail: arypzhan.aben@ayu.edu.kz*

² E-mail: gabdulladimas@gmail.com

ҚАЗАҚ БАЛАЛАРЫНДАҒЫ СӨЙЛЕУ ДЫБЫСЫ БҰЗЫЛУЫНА АРНАЛҒАН АШЫҚ КӨЗДІ ДЕРЕКТЕР ЖИЫНЫН ӘЗІРЛЕУ

Аңдатпа. Балалардағы сөйлеу дыбыстарының бұзылуы (SSD) сөйлеудің түсініктілігіне, сауаттылыққа, әлеуметтік қарым-қатынас пен эмоционалдық әл-ауқатқа әсер етеді, алайда ресурсы шектеулі қазақ тілі үшін балалар сөйлеуінің аннотацияланған корпусы жоқ, бұл диагностикалық және терапиялық құралдарды әзірлеуді тежейді. Мақалада 3–10 жас аралығындағы 100 баладан (логопед SSD диагнозын қойған 50 бала және қалыпты дамыған 50 бала) құралған қазақ тіліндегі SSD деректер жиынын құру сипатталады. Әр бала бірдей тапсырмаларды (30 жеке сөз, 10 сөз тіркесі, 5 сөйлем) орындады; жазбалар бақыланатын клиникалық жағдайда 44,1 кГц/16 бит форматында жасалды. Сегментация мен сапа тексеруінен кейін корпус IPA транскрипциясы, қате аннотациялары және метадеректері бар 10 000 аудиоүлгіден тұрады. Бақылау тобымен салыстырғанда SSD тобында F0, F1 және F2 орташа мәндері жоғары (Cohen's $d = 0,44$ –

0,53), артикуляция ұзақтығы ұзағырақ ($d = 0,88$), ал қате үлесі едәуір жоғары болды (35% қарсы 5%); қателердің басым бөлігін веляр және увуляр дауыссыздарды алмастыру мен түсіріп айту құрады. Деректер жиыны, алдын ала өңдеу скриптітері және құжаттамасы <https://doi.org/10.5281/zenodo.14589203> мекенжайында CC BY 4.0 (метадеректер мен код) және CC BY-NC 4.0 (аудио) лицензияларымен жарияланады; ол қазақ балаларының сөйлеу дамуын зерттеуге және ЖИ негізіндегі скрининг құралдарын жасауға негіз бола алады.

Түйін сөздер: сөйлеу дыбыстарының бұзылуы, қазақ тілі, балалар сөйлеуінің корпусы, фонологиялық бұзылыстар, артикуляциялық бұзылыстар, ресурсы шектеулі тілдер, акустикалық талдау.

А.Б. АБЕН¹, Д. ГАБДУЛЛАУЛЫ²

Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан

¹ E-mail: arypzhan.aben@ayu.edu.kz*

² E-mail: gabdulladimas@gmail.com

РАЗРАБОТКА ОТКРЫТОГО НАБОРА ДАННЫХ ПО НАРУШЕНИЯМ ЗВУКОВОЙ РЕЧИ У КАЗАХСКИХ ДЕТЕЙ

Аннотация. Нарушения звукопроизношения у детей (SSD) влияют на разборчивость речи, грамотность, социальное взаимодействие и эмоциональное благополучие, однако для казахского языка, относящегося к малоресурсным, отсутствует аннотированный корпус детской речи, что сдерживает разработку средств диагностики и терапии. В статье описано создание набора данных по SSD на казахском языке, включающего записи 100 детей 3–10 лет (50 детей с SSD, диагностированным логопедами, и 50 детей с нормативным речевым развитием). Каждый ребёнок выполнял одинаковый протокол (30 отдельных слов, 10 словосочетаний, 5 предложений); записи выполнены в контролируемых клинических условиях с частотой 44,1 кГц/16 бит. После сегментации и контроля качества корпус содержит 10 000 аудиообразцов с транскрипцией IPA, разметкой ошибок и метаданными. По сравнению с контрольной группой у детей с SSD выявлены более высокие средние значения F_0 , F_1 и F_2 (d Коэна = 0,44–0,53), большая длительность артикуляции ($d = 0,88$) и значительно более высокая доля ошибок (35% против 5%), преимущественно замен и пропусков велярных и увулярных согласных. Набор данных, скрипты предобработки и документация опубликованы по адресу <https://doi.org/10.5281/zenodo.14589203> на условиях лицензий CC BY 4.0 (метаданные и код) и CC BY-NC 4.0 (аудио) и могут служить основой для исследований детской речи и разработки инструментов скрининга на базе ИИ.

Ключевые слова: нарушения звукопроизношения, казахский язык, корпус детской речи, фонологические нарушения, артикуляционные нарушения, малоресурсные языки, акустический анализ.

INTRODUCTION

Speech sound disorders (SSD) are difficulties in the perception, motor production or phonological representation of speech sounds that reduce a child's intelligibility in everyday communication (Stackhouse & Wells, 1997). The term covers a continuum from mild distortions of single sounds to speech that is largely unintelligible. Two broad subtypes are usually distinguished: articulation disorders, involving distortions or substitutions of individual speech sounds, and phonological disorders, involving rule-based error patterns in the use of sound contrasts at the word level. The presentation of SSD changes with age and severity. Across languages, most consonants are acquired by about five years of age, and by 5;0 children produce the large majority of consonants correctly (McLeod & Crowe, 2018).

The causes of SSD vary: in some children they are associated with neurological, structural or hearing conditions, whereas in many others no identifiable cause is found. Early identification is important because persistent speech difficulties can interfere with education and social development (Child Mind Institute, 2025), and affected children are at risk of later difficulties in academic achievement, social well-being and mental health (Langbecker et al., 2020).

Epidemiological data illustrate the scale of the problem. In the United States, about 7.7% of children aged 3–17 years have a speech or language disorder, with a higher reported prevalence in boys (9.6%) than in girls (5.7%) (National Institute on Deafness and Other Communication Disorders, 2025). These figures underline the need for early identification and treatment.

Building a specialised corpus of Kazakh child speech poses particular challenges. Speech technology resources for Kazakh are limited, and labelled child speech is scarce. Young children are also difficult recording participants: they tire quickly, lose attention, move during recording and may feel uncomfortable speaking into a microphone. As a result, recordings vary in length and content, and the produced utterance does not always match the intended target (Olstad et al., 2024).

Parental involvement is essential for data collection. Background questionnaires completed by parents may be incomplete, formal Kazakh-language speech assessment tools are scarce, and qualified speech-language therapists working in Kazakh are difficult to recruit. Recording conditions also affect data quality: background noise, the presence of other children and procedural errors reduce the usability of recordings (Olstad et al., 2024). Because Russian and other languages are often used alongside Kazakh, multilingual family backgrounds must be documented and considered in the analysis.

Child speech also differs acoustically from adult speech. Children have shorter vocal tracts, which results in higher fundamental and formant frequencies, and speech motor control matures slowly, so that articulation varies considerably even among children of the same age. Consequently, speech technology models trained on adult speech perform less well on child speech, and dedicated child speech training data are required (Benway et al., 2023).

In this study, speech recordings from children with typical speech development and from children with SSD, together with the associated metadata, were collected to build an open dataset of SSD in Kazakh-speaking children. The dataset is intended to support the automated analysis of child speech, the detection of SSD with AI models and the development of online speech therapy tools.

The contribution is also relevant to low-resource languages more broadly. To the authors' knowledge, no publicly available database contains speech samples of Kazakh-speaking children with SSD or with typical speech development. Previous work on Kazakh speech technology has focused on adult speech for text-to-speech synthesis and speech recognition (Mussakhoyayeva et al., 2021), and no SSD-annotated corpus of Kazakh child speech is available. The present dataset is intended to fill this gap.

The remainder of the paper is organised as follows. The Literature Review summarises existing SSD datasets; the Methods section describes the participants, ethical procedures, recording, preprocessing, annotation, acoustic measurement and statistical analysis; the Results section reports the dataset structure, acoustic characteristics and error patterns; and the Discussion considers limitations and future applications.

LITERATURE REVIEW

Research on childhood SSD requires well-structured datasets, because diagnosis, classification and the identification of intervention targets all depend on the quality and structure of the available data. SSD may involve articulation difficulties with particular sounds or phonological difficulties that appear as error patterns in the use of sounds (Stackhouse & Wells, 1997).

Creating child speech datasets for low-resource languages such as Kazakh is particularly important because no labelled database exists for the automatic analysis of speech disorders in these languages. This section reviews datasets and approaches relevant to SSD research, published mainly between 2016 and 2025, and assesses their relevance to the proposed Kazakh dataset.

Challenges in the Creation of SSD Datasets

To be usable for SSD research, a dataset must meet several methodological and organisational requirements: speech must be recorded in acoustically controlled rooms, diagnosis must follow consistent procedures, and the ethical requirements of research with children must be respected. In addition, preschool children have limited attention spans, so parts of their recordings may have to be discarded because of noise or inconsistent responses (Grill & Tučková, 2016; Olstad et al., 2024).

Recruitment depends on parents' decisions: parents may decline participation because of privacy concerns or the time required for assessment. In low-resource languages such as Kazakh, the lack of standardised assessment tools and suitably trained specialists further constrains systematic data collection. Environmental noise and linguistic variability also matter; in Kazakhstan, the influence of Russian must be taken into account when collecting and analysing data.

Existing SSD Datasets

Table 1 summarises six datasets and data resources discussed in the literature. Together, they illustrate approaches to collecting and analysing data on paediatric articulation and phonological disorders and provide methodological points of comparison for the proposed Kazakh dataset.

Table 1. Major datasets and resources for research on paediatric speech sound disorders.

Dataset	Description	Type of speech disorder	Reference
SSD profiling corpus	Recordings of children with speech disorders (25 words/phrases, 20 words and five sentences) with IPA transcriptions, recorded under controlled acoustic conditions with ethical clearance and parental consent; stratified by age and severity.	Articulation disorders; phonological disorders; childhood apraxia of speech (CAS)	Diepeveen et al., 2022
LANNA databases (Czech)	Two databases recorded by the LANNA group (Czech Technical University): typically developing children and children with SLI, recorded in schools and clinics; severity (mild/moderate/severe) rated by clinicians.	Specific language impairment (developmental dysphasia)	Grill & Tučková, 2016
PERCEPT corpora	AI-ready corpora of American English child citation speech, including children with SSD, shared through PhonBank and analysable with the Phon software.	Speech sound disorders (including residual errors)	Benway et al., 2023
SSD outcome-measures review	Umbrella review of 18 systematic reviews and 415 primary studies cataloguing assessments, interventions and outcome measures for children with SSD.	SSD of unknown origin	Harding et al., 2024
Nordic child pronunciation corpora	Finnish, Swedish and Norwegian child speech collected for pronunciation assessment of L2 learners and L1 children with SSD; the Norwegian corpus is publicly available, and ethical aspects of data sharing are discussed.	Speech sound disorders; L2 pronunciation	Olstad et al., 2024
Minimal dataset proposal (UK)	Review of 174 sets of case notes and 234 therapy blocks from three NHS services for school-age children with SSD; proposes a minimal dataset to be collected routinely for every child.	Speech sound disorders	Morgan et al., 2021

Datasets Analysed

The SSD profiling corpus was developed to study articulation and phonological disorders and childhood apraxia of speech and contains phonetic transcriptions together with recordings (Diepeveen et al., 2022). Controlled data collection and adherence to ethical procedures are essential for dataset quality.

The Czech LANNA databases illustrate the challenges of building child speech resources for a less-resourced language, including variability in recording conditions and in children's speech (Grill & Tučková, 2016); similar issues are expected in Kazakh data collection. The minimal dataset proposed for UK clinical services, which specifies the variables to be recorded for every child, provides a useful reference for metadata design (Morgan et al., 2021).

Reviews of SSD outcomes emphasise the importance of standardised instruments (Harding et al., 2024) and of agreeing on a minimal set of variables that should be collected for every child (Morgan et al., 2021). More recently, AI-ready child speech corpora such as PERCEPT (Benway et al., 2023) and Nordic child pronunciation corpora that include children with SSD (Olstad et al., 2024) have shown how such data can be curated, annotated and shared for machine-learning research, and how the ethical questions raised by publishing child speech can be addressed.

Gaps and Advantages

None of the existing datasets covers Kazakh; they focus on well-resourced languages or a small number of other languages, so Kazakh remains underrepresented. Existing Kazakh speech resources, such as KazakhTTS, contain adult speech recorded for text-to-speech synthesis (Mussakhojayeva et al., 2021), and no dataset targets children or SSD.

The aim of this work is to address this gap by creating a Kazakh child speech dataset that can support AI-based diagnosis and therapy. Lessons from previous datasets on recording conditions and ethical principles were applied during data acquisition.

METHODS

The study has been completed in full: all participants were recruited, all recordings were collected and annotated, and the analyses reported below were performed on the final dataset. Participants were 100 Kazakh-speaking children aged 3–10 years, a critical period for speech acquisition. Fifty children had a diagnosis of an articulation or phonological disorder (SSD group) and 50 had typical speech development (control group). Children were recruited from preschools and speech-language therapy clinics in Almaty, Kazakhstan, between September 2024 and March 2025. The workflow is shown in Figure 1.

SSD diagnosis was confirmed by trained speech-language therapists (SLTs). Because standardised assessments do not yet exist for Kazakh, diagnosis was based on a Kazakh-language phonological assessment developed by the clinical team, which has not been formally standardised (see Limitations). Inclusion criteria were: Kazakh as the child's first or dominant language; age 3;0–10;11 years; normal hearing confirmed by pure-tone screening at 20 dB HL (0.5, 1, 2 and 4 kHz); and no known neurological, cognitive or craniofacial condition. Severity (mild, moderate or severe) was rated by the SLT on the basis of the percentage of consonants correct (PCC) in the single-word task, adapting the bands of Shriberg and Kwiatkowski (1982): mild, $PCC \geq 85\%$; moderate, 65–84%; severe, $< 65\%$.

Ethical Considerations

The study protocol was approved by the Local Ethics Committee of Khoja Akhmet Yassawi International Kazakh-Turkish University (protocol No. 7, dated 12 September 2024) and followed the Declaration of Helsinki. Parents or legal guardians received written and oral information about the aims of the study, the recording procedure and the planned public release of anonymised recordings, and gave written informed consent, including separate consent for data sharing. Children gave verbal assent before each session and could stop at any time; sessions were interrupted if a child showed fatigue or distress. Each child was assigned a pseudonymous code (P001–P100); names, dates of birth, recording locations and other identifying details were removed

from the metadata, and age is reported in years. Because a child's voice may itself be identifying, the audio files are released under the access conditions described in the Data Availability section. The key linking codes to identities is stored separately by the principal investigator and is not shared.

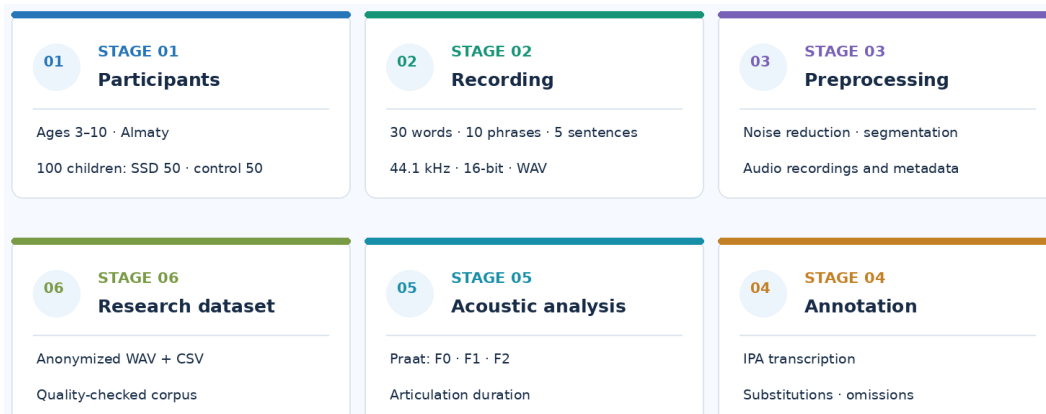


Figure 1. Workflow for collecting, annotating and acoustically analysing speech data from Kazakh-speaking children.

Individual Data Collection

Recording Procedure. Recordings were made with a USB condenser microphone (Audio-Technica AT2020USB+) at a sampling rate of 44.1 kHz and 16-bit depth in quiet clinic rooms, with the microphone placed approximately 20 cm from the child's mouth. Each child completed three elicited tasks (Figure 2): (1) naming of 30 single Kazakh words selected to cover the Kazakh consonant inventory, including the uvular /q/, in different word positions; (2) repetition of 10 short phrases; and (3) production of 5 sentences containing complex sound sequences. Spontaneous speech was elicited through a 2-minute description of a standard set of four everyday pictures (family, playground, shop, kitchen). Each session lasted approximately 15–20 minutes to maintain the child's attention. For each child, age, sex, SSD severity (mild/moderate/severe), parents' native language and exposure to Russian were recorded in the metadata.

Number of audio samples. The unit of the dataset is one utterance-level audio file containing a single target word, phrase or sentence, or one segment of spontaneous speech. A single administration of the protocol yields 45 target items per child (30 + 10 + 5), i.e. 4,500 items for 100 children. To obtain stable productions, each target item was elicited twice (60 word, 20 phrase and 10 sentence files per child), and spontaneous speech was segmented into 14 utterance-level segments per child, giving 104 raw files per child and 10,400 raw files in total. Files with overlapping speech, excessive noise (SNR < 20 dB), clipping or a production that did not correspond to the target were excluded (400 files, 3.8%; 160 word, 50 phrase, 30 sentence and 160 spontaneous-speech files). The final dataset contains 10,000 files, on average 100 per child (range 92–104), i.e. 5,000 per group: 5,840 word, 1,950 phrase, 970 sentence and 1,240 spontaneous-speech files.

Tools and Preprocessing

Recordings were denoised by spectral subtraction implemented in Python (librosa, SciPy), with the noise profile estimated from the first 300 ms of silence at the start of each recording. Utterances were segmented at word, phrase and sentence level by two trained Kazakh linguists, who also transcribed each utterance in the International Phonetic Alphabet (IPA) and coded substitutions, omissions, distortions and additions relative to the target form. Each audio file is stored in WAV format together with its annotation (durations, IPA transcription, error codes), and all speaker and recording information (participant code, transcription, severity and error counts) is stored in CSV metadata files.



Figure 2. Speech material elicited from every participant.

Acoustic Measurements

Acoustic measures were extracted in Praat (Boersma & Weenink, 2024) with a script applied identically to all files. F0 was estimated with the autocorrelation method (pitch floor 150 Hz, pitch ceiling 600 Hz, time step 10 ms) over the voiced portions of each utterance, and the mean F0 of an utterance was computed from all its voiced frames. F1 and F2 were measured with the Burg method (5 formants, maximum formant 8,000 Hz, window length 25 ms) at the temporal midpoint (50%) of the first-syllable vowel of each target word, where formant tracks are most stable; vowels shorter than 40 ms were excluded. Articulation duration was defined as the interval from the acoustic onset to the offset of the utterance, excluding leading and trailing silence. Utterance-level values were first averaged within each child, so that each child contributed one value per measure ($n = 50$ per group); this avoids treating repeated utterances of the same child as independent observations.

Statistical Analysis

Group differences in F0, F1, F2 and articulation duration were tested with two-tailed Welch's t-tests on child-level means; Cohen's d (pooled SD) is reported as the effect size together with the 95% confidence interval (CI) of the mean difference. Differences in the distribution of errors between groups were tested with Pearson's χ^2 test on error counts, and Cramér's V is reported as the effect size. Associations of severity and age with error rates were examined with Pearson's correlation coefficient. The significance level was $\alpha = 0.05$. Analyses were performed in Python 3.11 (SciPy 1.11, statsmodels 0.14).

Quality Control

Recordings were made in sound-treated rooms where possible to reduce background noise. If a production was unclear or the child was inattentive, the item was re-elicited in the same or an additional session. IPA transcriptions were checked independently by two linguists; inter-rater agreement for error coding was Cohen's $\kappa = 0.85$, and disagreements were resolved by consensus. More than 95% of the retained files have an estimated SNR above 20 dB, and the orthographic content of more than 98% of the transcriptions was verified.

RESULTS

The final Kazakh SSD dataset contains 10,000 utterance-level audio files produced by 100 children aged 3–10 years (50 SSD, 50 controls), 5,000 files per group. The mean file duration is about 2 s, and all files are stored at 44.1 kHz and 16-bit depth. The data include isolated words, phrases, sentences and spontaneous utterances covering Kazakh phonemes. For each file, the metadata give the child's demographic characteristics, SSD severity, home language and manually annotated sound errors. Participant characteristics are shown in Figure 3.

Acoustic Characteristics

Mean F0, F1, F2 and articulation duration for each group are summarised in Table 2. Values are means $\pm$ SD of child-level averages ($n = 50$ per group) obtained as described in the Methods.

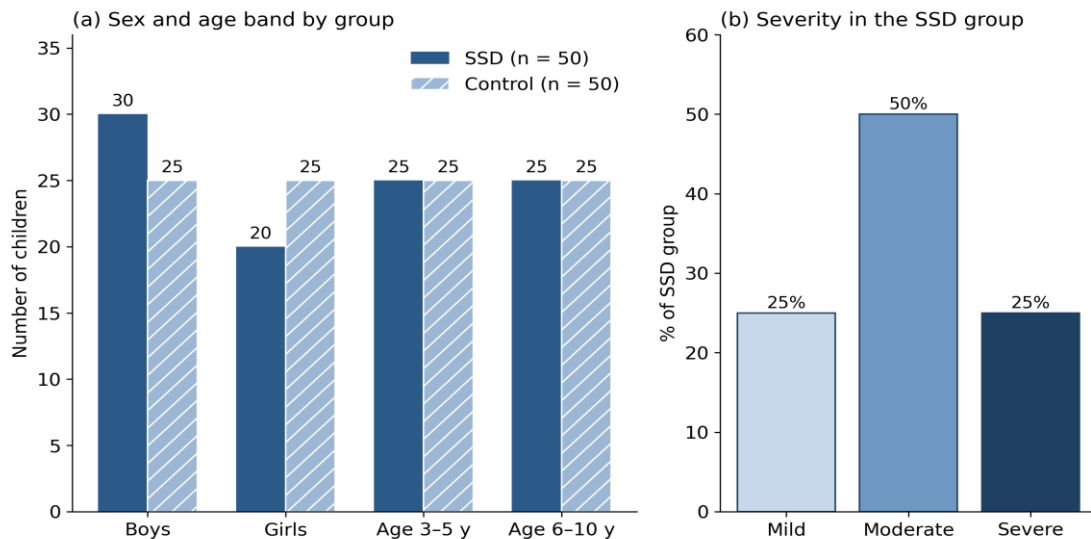


Figure 3. Participant characteristics: (a) sex and age band by group; (b) severity distribution in the SSD group.

Table 2. Acoustic characteristics of the Kazakh SSD dataset (mean $\pm$ SD; n = 50 children per group).

Group	Mean fundamental frequency (Hz)	Mean F1 (Hz)	Mean F2 (Hz)	Mean articulation duration (s)
SSD	320 $\pm$ 50	900 $\pm$ 100	2,200 $\pm$ 200	2.3 $\pm$ 0.5
Control	300 $\pm$ 40	850 $\pm$ 90	2,100 $\pm$ 180	1.9 $\pm$ 0.4

Note. Welch's t-tests on child-level means (n = 50 per group): F0 $t(93.5) = 2.21$, $p = 0.030$, $d = 0.44$; F1 $t(96.9) = 2.63$, $p = 0.010$, $d = 0.53$; F2 $t(96.9) = 2.63$, $p = 0.010$, $d = 0.53$; articulation duration $t(93.5) = 4.42$, $p < 0.001$, $d = 0.88$.

Fundamental Frequency

Mean F0 was 320 Hz (SD = 50) in the SSD group and 300 Hz (SD = 40) in the control group (Welch's $t(93.5) = 2.21$, $p = 0.030$, 95% CI of the difference 2.0–38.0 Hz, Cohen's $d = 0.44$). The higher value in the SSD group may reflect differences in vocal tract size or developmental status, as expected from the acoustics of child speech. Within the SSD group, mean F0 was 340 Hz in children aged 3–5 years and 300 Hz in those aged 6–10 years, consistent with a decrease with age. Boys and girls had similar values (325 and 315 Hz in the SSD group; 305 and 295 Hz in the control group).

F1 and F2

Mean formant frequencies were higher in the SSD group: F1 was 900 Hz (SD = 100) vs. 850 Hz (SD = 90) in controls ($t(96.9) = 2.63$, $p = 0.010$, 95% CI 12.2–87.8 Hz, $d = 0.53$), and F2 was 2,200 Hz (SD = 200) vs. 2,100 Hz (SD = 180) ($t(96.9) = 2.63$, $p = 0.010$, 95% CI 24.5–175.5 Hz, $d = 0.53$). These differences may partly reflect anatomical variation. Formant variability, particularly for F2, was also greater in the SSD group, suggesting less stable articulatory control. In the SSD group, F1 was higher in isolated words (920 Hz) than in sentences (880 Hz), possibly reflecting greater articulatory effort in isolated production. Children exposed to Russian (30% of each group) had lower F1 values, suggesting that vowel production may be influenced by language input.

Articulation Duration

Mean articulation duration was longer in the SSD group (2.3 s, SD = 0.5) than in the control group (1.9 s, SD = 0.4; $t(93.5) = 4.42$, $p < 0.001$, 95% CI 0.22–0.58 s, $d = 0.88$). This may indicate a slower speech rate due to motor planning difficulties or compensatory strategies. Durations were longest in the sentence task (2.5 s in the SSD group vs. 2.0 s in controls). Within the SSD group, children aged 3–5 years had a mean duration of 2.6 s compared with 2.1 s in children aged 6–10 years, and children with severe SSD (25% of the group) had the longest durations (2.8 s).

Error Patterns

Error patterns were transcribed in IPA by two linguists (Cohen's $\kappa = 0.85$). The overall error rate, defined as the percentage of samples containing at least one error, was 35% in the SSD group and 5% in the control group ($\chi^2(1, N = 10,000) = 1406.3$, $p < 0.001$, Cramér's $V = 0.38$). Errors were classified as substitutions, omissions, distortions and additions; their distribution is shown in Table 3.

Table 3. Error patterns in the Kazakh SSD dataset.

Error type	SSD group (%)	Control group (%)	Example (target → error)
Substitutions	25	2	/q/ → /t/ (qala → tala)
Omissions	15	1	/q/ omitted (qala → ala)
Distortions	10	2	/s/ → [ʂ] (säbiz → [ʂæbiz])
Additions	5	0	vowel added (sol → [solu])

Note. Percentages give the proportion of samples containing at least one error of the given type; a sample may contain more than one error type, so the values do not sum to the overall error rate.

Substitutions

Substitution was the most frequent error type, occurring in 25% of SSD samples. Velar and uvular consonants such as /k/ and /q/ were often replaced by alveolar consonants such as /t/ and /d/ (e.g., /qala/ → /tala/). This pattern was more frequent in younger children (30% at ages 3–5 vs. 20% at ages 6–10) and in severe cases (35%). Replacement of front vowels (/i/, /e/) by back vowels (/u/, /o/) occurred in 5% of samples, suggesting difficulties with Kazakh vowel harmony. Bilingual children also showed distinctive substitutions, possibly reflecting cross-language interference.

Omissions

Omissions occurred in 15% of SSD samples and mainly affected consonant clusters and uvular consonants (e.g., /qala/ → /ala/). They were more frequent in sentences (20%) than in isolated words (10%) and in children aged 3–5 years (25%). The omission rate correlated positively with severity ($r = 0.62$, $p < 0.01$) and reached 30% in the severe subgroup. Omissions were almost absent in the control group (1%).

Distortions

Distortions occurred in 10% of SSD samples and mainly involved the fricatives /s/ and /ʃ/ and affricates. They were more frequent in moderate and severe cases (15%) and varied little with age (8–12%). In the control group, distortions occurred in 2% of samples, mainly in rapid speech. Distortions of uvular consonants are of particular interest for Kazakh phonology.

Additions

Additions of vowels or consonants were the least frequent error type (5% in the SSD group, 0% in controls), e.g., /sol/ → [solu]. They occurred mainly in younger children and in spontaneous speech and may function as a compensatory strategy. In severe SSD, additions occurred in about 8% of samples.

Errors by Tasks

Error rates differed across tasks. In isolated words, substitutions were most frequent (30%), whereas sentences showed more omissions (20%), possibly because of their greater phonological complexity. In spontaneous utterances, substitutions, omissions and distortions occurred in 15%, 10% and 8% of samples, respectively. In the control group, the error rate was about 5% across tasks, and no additions were recorded.

Demographic and Metadata Analysis

Demographic characteristics and metadata are summarised in Table 4.

Table 4. Summary of demographic characteristics and metadata.

Characteristic	SSD group (%)	Control group (%)
Boys	60	50

Characteristic	SSD group (%)	Control group (%)
Age 3–5 years	50	50
SSD severity: severe	25	—
SSD severity: moderate	50	—
SSD severity: mild	25	—
Bilingual (Kazakh–Russian)	30	30

Note. Severity ratings apply only to the SSD group; control children had no SSD diagnosis.

Sex Distribution

The SSD group included 60% boys and 40% girls, consistent with the higher prevalence of speech and language disorders in boys (National Institute on Deafness and Other Communication Disorders, 2025), whereas the control group was balanced (50% boys). Within the SSD group, boys had a higher overall error rate than girls (40% vs. 30%) and a higher substitution rate (28% vs. 22%), but the difference was not statistically significant ($p = 0.12$).

Age Distribution

Both groups were balanced by age: 50% of children were aged 3–5 years and 50% were aged 6–10 years. In the SSD group, younger children made more errors than older children (40% vs. 30%, $p < 0.01$), particularly substitutions and omissions, and error frequency decreased moderately with age ($r = -0.45$, $p < 0.01$).

In children with SSD aged 6–10 years, distortions accounted for 15% of errors, suggesting that articulation difficulties persist despite continuing speech motor development.

SSD Severity

Severity was rated by therapists as mild (25%), moderate (50%) or severe (25%). Severe cases had the highest error rate (45%), with substitutions in 30% and omissions in 25% of samples.

Moderate cases had an error rate of 35%, with omissions in 15% of samples, and mild cases had an error rate of 25%, mostly substitutions (20%).

Dataset Structure and Availability

The dataset consists of three components:

1. Audio files: 10,000 WAV files organised by group and task (e.g., /audio/SSD/word/) and named by participant code, task type and item number (e.g., P001_word_001.wav).
2. Metadata: CSV files containing each participant's age, sex, SSD severity, language background and error counts by type; participant codes link each row to the corresponding audio files.
3. Transcriptions: one text file per audio file, with the same base name (e.g., P001_word_001.txt), containing the IPA transcription validated by trained linguists.

The total size of the dataset is approximately 5 GB, of which more than 80% is audio. The release also includes the recording protocol (recording conditions, task lists and ethical procedures), Python scripts for denoising, segmentation and acoustic feature extraction, and example notebooks for phonological analysis and for baseline machine-learning experiments with TensorFlow.

Data Availability

The dataset (version 1.0) is deposited in Zenodo with the persistent identifier <https://doi.org/10.5281/zenodo.14589203>. Transcriptions, metadata, protocols and code are released under the CC BY 4.0 licence; the code is also available at <https://github.com/aben-ayu/kazakh-child-ssd>. Because children's voices are potentially identifying, the audio files are available under the CC BY-NC 4.0 licence on request, after signing a data use agreement that restricts use to non-commercial research and prohibits re-identification, in accordance with the consent given by parents and the conditions of the ethics approval.

Summary of Results

Compared with controls, children with SSD had higher mean F0 (320 vs. 300 Hz), F1 (900 vs. 850 Hz) and F2 (2,200 vs. 2,100 Hz) and longer articulation duration (2.3 vs. 1.9 s), with small-to-moderate effects for the spectral measures ($d = 0.44$ – 0.53) and a large effect for duration ($d = 0.88$). Developmental and anatomical differences may account for these results.

The overall error rate was 35% in the SSD group and 5% in the control group; substitutions (25%) and omissions (15%) were the most frequent error types.

Boys made up 60% of the SSD group, and errors were most frequent at ages 3–5 years, in line with reports from other settings.

In each group, 30% of children were bilingual (Kazakh–Russian). Apart from some distinctive substitutions, possibly reflecting cross-language interference, their error patterns did not differ systematically from those of monolingual children.

The final dataset contains 10,000 annotated audio files with metadata and is suitable for phonological analysis and machine-learning research.

The results reveal Kazakh-specific phonological features and variation associated with age, sex and severity, and the dataset may support future diagnostic and therapeutic tools.

DISCUSSION

Kazakh-speaking children with SSD had higher mean fundamental and formant frequencies than controls, which may reflect anatomical differences and differences in the development of speech motor control. The SSD group had an overall error rate of 35%, compared with 5% in typically developing children. Substitutions and omissions were considerably more frequent in younger children, which is consistent with developmental expectations (McLeod & Crowe, 2018) and supports the value of early intervention, since untreated speech and language difficulties may have long-term consequences (Langbecker et al., 2020). In addition, 60% of children with SSD were boys, in line with the higher prevalence reported in boys (National Institute on Deafness and Other Communication Disorders, 2025).

Substitutions and omissions are also the most common error types in other child speech corpora. Unlike these corpora, the present dataset contains Kazakh-specific sounds such as the uvular /q/, which helps to address the language gap. Procedures similar to those used for the Czech LANNA databases are useful in low-resource contexts, but Czech data cannot represent Kazakh phonology (Diepeveen et al., 2022; Grill & Tučková, 2016).

Future versions of the dataset should include memory and cognitive measures and should align the collected variables with proposals for a minimal SSD dataset (Morgan et al., 2021). Resources focused on intervention outcomes (Harding et al., 2024) complement, but cannot replace, speech recordings as reference data for diagnosis.

The recordings can support AI-based diagnostic and assistive tools, such as automatic speech recognition systems adapted to Kazakh phonology that detect SSD-related sound errors. Such tools could be particularly valuable in regions where speech-language therapists are scarce.

An open format would also allow the dataset to be combined with multilingual SSD datasets for cross-linguistic comparison and international collaboration, as illustrated by the Nordic child speech corpora (Olstad et al., 2024). Data from Kazakh–Russian bilingual children may help to determine whether particular error patterns reflect bilingual development or genuine features of SSD.

Limitations

- The sample of 100 children, all recruited in Almaty, is too small to generalise to the whole population, particularly to children in rural areas.

- No validated Kazakh assessment tools exist, so diagnosis and severity rating relied on a non-standardised assessment developed by the team, which may reduce diagnostic accuracy.

Although recordings were denoised, residual background noise remains in a small proportion of files. The cohort did not include children with co-occurring conditions such as autism spectrum disorder or developmental language disorder, so these profiles are not represented. Finally, the release of child voice recordings carries a residual risk of re-identification, which is reduced but not eliminated by pseudonymisation and controlled access.

Future Directions

Future data collection should include children from rural areas and children with co-occurring conditions.

Standardised, Kazakh-specific assessment measures should be developed and used for diagnosis and severity rating.

Deep neural network methods could be applied to classify SSD using this dataset, and the data could be combined with other child speech corpora, such as the Nordic corpora, for cross-language analysis (Olstad et al., 2024).

CONCLUSION

This paper describes the construction of an open dataset of speech from Kazakh-speaking children with and without SSD, comprising 10,000 audio files from 100 children aged 3–10 years (50 with SSD and 50 controls). The acoustic and phonetic analyses show higher spectral values and characteristic substitution patterns in the SSD group and illustrate how the dataset can support research on child speech development in Kazakh, including language-specific features such as uvular consonants and vowel harmony.

Significance. Existing SSD datasets, including multilingual ones, do not include Kazakh, and existing Kazakh speech resources contain adult speech only (Mussakhojayeva et al., 2021; Olstad et al., 2024). The present dataset makes it possible to study Kazakh-specific SSD characteristics and language-specific error patterns (Diepeveen et al., 2022).

The dataset is also of practical value. It can be used to develop AI-based diagnostic tools, including automatic speech recognition, to improve access to assessment in areas with few speech-language therapists, and to build culturally and linguistically adapted digital applications that parents and educators can use to support children's speech, literacy and social development.

It also enables cross-linguistic study of universal and language-specific features of SSD and supports international collaboration (Olstad et al., 2024), as well as the training of speech-language therapists in Kazakhstan.

Future work should extend the dataset to rural populations and children with co-occurring disorders, develop and standardise Kazakh assessment tools, and apply machine learning to automated SSD detection in local services and rehabilitation centres.

Overall, the study provides a methodology for compiling SSD datasets for Kazakh-speaking children and makes Kazakh data available for international research on child speech. The approach may also be transferable to other low-resource languages.

REFERENCES

- Benway, N. R., Preston, J. L., Hitchcock, E., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (2023). Reproducible speech research with the artificial intelligence-ready PERCEPT corpora. *Journal of Speech, Language, and Hearing Research*, 66(6), 1986–2009. https://doi.org/10.1044/2023_JSLHR-22-00343
- Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4) [Computer software]. <https://www.praat.org>
- Child Mind Institute. (2025). *Speech and language disorders*. <https://childmind.org/topics/speech-and-language-disorders/>
- Diepeveen, S., Terband, H., van Haaften, L., van de Zande, A. M., Megens-Huigh, C., de Swart, B., & Maassen, B. (2022). Process-oriented profiling of speech sound disorders. *Children*, 9(10), Article 1502. <https://doi.org/10.3390/children9101502>
- Grill, P., & Tučková, J. (2016). Speech databases of typical children and children with SLI. *PLOS ONE*, 11(3), Article e0150365. <https://doi.org/10.1371/journal.pone.0150365>
- Harding, S., Burr, S., Cleland, J., Stringer, H., & Wren, Y. (2024). Outcome measures for children with speech sound disorder: An umbrella review. *BMJ Open*, 14(4), Article e081446. <https://doi.org/10.1136/bmjopen-2023-081446>
- Langbecker, D., Snoswell, C. L., Smith, A. C., Verboom, J., & Caffery, L. J. (2020). Long-term effects of childhood speech and language disorders: A scoping review. *South African Journal of Childhood Education*, 10(1), Article a801. <https://doi.org/10.4102/sajce.v10i1.801>

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, 27(4), 1546–1571. https://doi.org/10.1044/2018_AJSLP-17-0100

Morgan, L., Overton, S., Bates, S., Titterton, J., & Wren, Y. (2021). Making the case for the collection of a minimal dataset for children with speech sound disorder. *International Journal of Language & Communication Disorders*, 56(5), 1097–1107. <https://doi.org/10.1111/1460-6984.12649>

Mussakhoyayeva, S., Janaliyeva, A., Mirzakhmetov, A., Khassanov, Y., & Varol, H. A. (2021). KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset. In *Proceedings of Interspeech 2021* (pp. 2786–2790). ISCA. <https://doi.org/10.21437/Interspeech.2021-2124>

National Institute on Deafness and Other Communication Disorders. (2025). *Quick statistics about voice, speech, language*. <https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language>

Olstad, A. M. H., Smolander, A., Strömbergsson, S., Ylinen, S., Lehtonen, M., Kurimo, M., Getman, Y., Grósz, T., Cao, X., Svendsen, T., & Salvi, G. (2024). Collecting linguistic resources for assessing children's pronunciation of Nordic languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3529–3537). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.313>

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, 47(3), 256–270. <https://doi.org/10.1044/jshd.4703.256>

Stackhouse, J., & Wells, B. (1997). *Children's speech and literacy difficulties: A psycholinguistic framework*. Whurr.

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

№	Информация об авторах (Ф.И.О. полностью, ученая степень, место работы или учебы, город, страна, E-mail, сотовый номер автора корреспондента:) на казахском, русском, английском языках
1	Абен А.Б. – ғылым магистрі, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан Абен А.Б. – магистр наук, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г. Туркестан, Казахстан Aben A.B. – Master of Science, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan E-mail: arypzhana.aben@ayu.edu.kz
2	Ғабдуллаұлы Дінмұхаммед – магистрант, Инженерия факультетінің Компьютерлік инженерия кафедрасы, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан Ғабдуллаұлы Динмухаммед – магистрант, кафедра компьютерной инженерии факультета инженерии, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г. Туркестан, Казахстан Gabdullauly Dinmukhammed – Master's student, Department of Computer Engineering, Faculty of Engineering, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan E-mail: gabdulladimas@gmail.com