American Journal of Alternative Education
ISSN: 2997-3953 (Online)
Vol. 2, No. 9, Sep 2025,
Available online at: https://scientificbulletin.com/index.php/AJAE

Criteria for Selecting and Analyzing Audio Materials in Corpus Linguistics and Speech Technology

Urazaliyeva Mavluda Yangiboyevna
Independent researcher at the National University of Uzbekistan, urazaliyeva_m@nuu.uz

Article information:
Manuscript received: 25 Jul 2025; **Accepted:** 26 Aug 2025; **Published:** 27 Sep 2025

Abstract: This paper examines the fundamental criteria for selecting and analyzing audio materials in the context of corpus linguistics and speech technology, with particular reference to low-resource languages such as Uzbek. Audio corpora play a crucial role in advancing Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and linguistic research, yet their effectiveness depends on the quality and representativeness of the data they contain. Drawing on international best practices and a case study of Kun.uz audio news materials, five key criteria are identified and evaluated: authenticity and naturalness of speech; technical quality and recording standards; sociolinguistic representativeness, including demographic and dialectal diversity; multilayer annotation and metadata enrichment; and ethical and legal responsibility in data collection and sharing. The results demonstrate that while available Uzbek audio resources provide valuable input for ASR development, they remain limited in thematic scope and regional coverage, which restricts their broader applicability. By comparing these findings with established corpora such as TIMIT, SWITCHBOARD, LibriSpeech, and Common Voice, the study highlights both strengths and shortcomings in current practices. The paper concludes with recommendations for expanding thematic domains, improving annotation depth, and ensuring ethical standards, thereby contributing to the creation of more robust, inclusive, and sustainable audio corpora for linguistic and technological innovation.

Keys words: audio corpora, speech technology, automatic speech recognition (ASR), uzbek language, corpus linguistics, data selection criteria.

Introduction

In recent decades, digital technologies have ushered in a new era in linguistics and computer science, where audio corpora, alongside written corpora, have become indispensable tools of language research. While written texts enable the study of lexical and syntactic structures, they cannot adequately reflect the phonetic, prosodic, and temporal variability of speech. For this reason, audio materials today stand at the crossroads of linguistics, computer science, psychology, and media studies, serving both theoretical inquiries and practical applications—most notably Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) (Bird & Simons, 2003).

Historically, early phonetic laboratories—for instance, those led by Peter Ladefoged—relied on small and controlled speech samples, which were valuable for theory but insufficient to capture full representativeness (Ladefoged, 1967). By the 1980s–1990s, the development of systematic speech corpora became inevitable. The **TIMIT** corpus pioneered methods of time-aligned phonetic transcription (Garofolo et al., 1993). Subsequently, **SWITCHBOARD** collected large-scale telephone

conversations, providing access to natural speech (Godfrey et al., 1992). Later, corpora such as **Libri Speech** (over 1000 hours of read audiobooks) (Panayotov et al., 2015) and **Mozilla Common Voice** (Ardila et al., 2020) expanded scale, multilingualism, and openness. These projects demonstrate that corpus quality depends not only on size but also on authenticity, technical precision, sociolinguistic representativeness, depth of annotation, and adherence to ethical norms.

Today's state-of-the-art ASR systems—such as **wav2vec 2.0** and **Whisper**—are trained on massive corpora. Yet models built solely on read speech recorded in studio conditions perform poorly in everyday conversations, dialectal variation, and noisy environments (Baevski et al., 2020; Radford et al., 2023). Thus, every decision in corpus design—whether to capture spontaneous speech, set recording quality standards, or include dialectal diversity—plays a decisive role in the performance of neural models.

On this basis, five key criteria are proposed for selecting and analyzing audio materials:

Authenticity – While read speech ensures clarity, it lacks natural conversational features such as pauses, repetitions, and intonation shifts. Authentic speech is essential for linguistic reliability and ASR robustness (Godfrey et al., 1992).

Technical quality – Low-frequency (e.g., 8 kHz) or noisy recordings introduce acoustic distortions and can double the Word Error Rate (WER). High-quality recordings of at least 16 kHz are scientifically necessary (Panayotov et al., 2015).

Sociolinguistic representativeness – Corpora must capture not only the central dialect of Tashkent but also the varieties of Samarkand, Fergana, Khorezm, and Karakalpakstan. Otherwise, findings reflect only one dominant variety (Musaev et al., 2021).

Annotation – Orthographic transcription alone is insufficient. Multilayer annotation—including phonetic, morphological, and prosodic tiers—enhances reusability and scientific value (Garofolo et al., 1993; Bird & Simons, 2003).

Ethics – Since speech materials involve human participants, informed consent, anonymization, and open licensing (e.g., Creative Commons) are indispensable for legitimacy (Ardila et al., 2020).

These criteria are particularly urgent for low-resource languages such as Uzbek. Current ASR systems for Uzbek exhibit WER rates of 15–20%, considerably higher than those for English or Mandarin (Mukhamadiyev et al., 2022). The primary cause is not architecture but corpus limitations and thematic imbalance. Applying international standards can substantially improve Uzbek speech technologies (Bekarystankyzy et al., 2024).

Audio corpora represent a turning point in modern linguistics and speech technology. Their effectiveness depends on the scientifically grounded selection and analysis of materials. Authenticity, technical quality, representativeness, annotation, and ethical responsibility must guide every stage of development. This study seeks to establish the theoretical and practical foundations for analyzing Uzbek audio materials within these criteria and in comparison with international best practices.

Methods

The methodology of this study was designed to ensure scientific reliability, transparency, and replicability at every stage—from data collection to analysis and interpretation. Since the primary objective was to identify the criteria for selecting and analyzing Uzbek audio materials, the methodological framework relied on international best practices while also addressing the specific challenges of Uzbek as a low-resource language.

At the first stage, audio news items regularly published on the *Kun.uz* platform throughout 2024 were used. This platform served as the primary source due to the consistent technical quality of its audio reports. Each file was systematically collected and stored in a dedicated electronic database, accompanied by metadata including file name, article title, hyperlink, hashtags, and a short transcription.

In this way, a structured and research-ready dataset was established.

However, relying solely on *Kun.uz* would have limited the scope of the analysis. For comparative benchmarking, internationally recognized corpora were also incorporated. These included the phonetic corpus **TIMIT** (Garofolo et al., 1993), the conversational **SWITCHBOARD** (Godfrey et al., 1992), the large-scale **LibriSpeech** (Panayotov et al., 2015), the crowdsourced **Mozilla Common Voice** (Ardila et al., 2020), and the multilingual **MLS** corpus (Pratap et al., 2020). This comparative approach allowed the Uzbek dataset to be contextualized globally, highlighting both its unique characteristics and its gaps.

Once the dataset had been compiled, the central question became which criteria should guide the selection of audio materials. Five interrelated dimensions were applied:

Authenticity – Most *Kun.uz* materials are read speech, which ensures clarity but lacks natural conversational features such as pauses, hesitations, and prosodic variation. Similar limitations were noted in the TIMIT corpus, whose artificiality prompted the later development of SWITCHBOARD to capture spontaneous dialogue (Godfrey et al., 1992).

Technical quality – Even linguistically rich data lose scientific value if recording quality is poor. A minimum standard of 16 kHz sampling rate was adopted, with background noise not exceeding 30 dB. Prior studies confirm that higher noise levels can nearly double the Word Error Rate (WER) in ASR systems (Pratap et al., 2020).

Sociolinguistic representativeness – Analysis of hashtags revealed thematic imbalance: weather and international politics dominated, while domains such as healthcare and education were underrepresented. Similarly, geographical representation was skewed: Tashkent appeared frequently, whereas other regions—such as Karakalpakstan and Khorezm—were scarcely mentioned. This reduced the corpus's representativeness.

Annotation – *Kun.uz* audio reports were provided only with orthographic transcripts, lacking phonetic, prosodic, or morphological layers of annotation. This contrasts with corpora like Common Voice and SWITCHBOARD, which offer multilayer annotation and thus greater reusability (Ardila et al., 2020).

Ethics – Although *Kun.uz* materials are publicly available, future corpus building involving spontaneous speech must prioritize participant consent, anonymization, and open licensing, such as Creative Commons, to ensure legitimacy (Bird & Simons, 2003).

Analysis proceeded in three stages:

Quantitative frequency analysis – Hashtags were extracted and their frequencies calculated. Results were presented in tables and charts to identify dominant topics.

Comparative benchmarking – Thematic distributions in Uzbek audio were compared with international corpora. For instance, while SWITCHBOARD emphasized spontaneous conversations, *Kun.uz* materials primarily reflected media discourse. This revealed strengths and weaknesses of the dataset.

ASR performance estimation – Although neural models were not trained directly, predictions were made based on existing literature regarding WER and Character Error Rate (CER). It was estimated that narrow domains (e.g., weather, political news) might yield WER around 8–10%, whereas underrepresented domains could reach 20–25%.

All stages of the study were carefully documented. Data were stored in structured spreadsheets, files were systematically archived, and frequency analyses with visualizations were supported by reproducible scripts. This ensures that other researchers can replicate or extend the study. References to international corpora were drawn exclusively from academic sources, guaranteeing methodological credibility.

In sum, the methodology encompassed: systematic collection of *Kun.uz* (2024) audio data, application of five selection criteria, quantitative frequency analysis with visualization, benchmarking against

international corpora, and estimation of ASR performance via WER/CER.

Through these stages, a scientifically grounded, transparent, and replicable methodology was established, offering both theoretical and practical foundations for building an Uzbek audio corpus.

The *Kun.uz* dataset analyzed within this study contained 200 recordings published in 2024, comprising 321 unique hashtags. Although this scale is modest compared to major international corpora, it provides empirical insights into the current state of Uzbek audio resources. The analysis covered overall frequency distribution, leading domains, visualization, thematic clustering, and implications for ASR systems.

Results

The analysis of the *Kun.uz* 2024 audio dataset, comprising 200 audio messages and 321 associated hashtags, demonstrates significant thematic clustering, with several domains dominating the overall distribution. To provide a clear picture of this distribution, both tabular data and visual representations are included.

1. Frequency distribution of domains

Table 1 presents the most frequent thematic domains identified in the dataset. The leading categories are *ob-havo* (weather) and *O‘zgidromet* (the Uzbek Hydrometeorological Service), each occurring 10 times. They are followed by *Toshkent* (9), *Rossiya* (8), *Ukraina* (8), and *elektr* (8). Mid-range frequencies include *gaz* and *Farg‘ona* (5 each), while lower-frequency but notable domains include *Yaponiya, soliq, bojxona, havo,* and *ekologiya* (4 each).

Domain	Frequency	Percentage (%)
ob-havo	10	7.7
O‘zgidromet	10	7.7
Toshkent	9	6.9
Rossiya	8	6.2
Ukraina	8	6.2
elektr	8	6.2
gaz	5	3.8
Farg‘ona	5	3.8
Yaponiya	4	3.1
soliq	4	3.1
bojxona	4	3.1
havo	4	3.1
ekologiya	4	3.1

Table 1. Frequency of top domains in Kun.uz 2024 audio dataset

2. Visualization of thematic distribution

The **bar chart (Figure 1)** illustrates the frequency of the top domains. The predominance of meteorological categories (*ob-havo, O‘zgidromet*) is immediately apparent, while political (Rossiya, Ukraina) and infrastructural (elektr, gaz) themes also appear prominently.

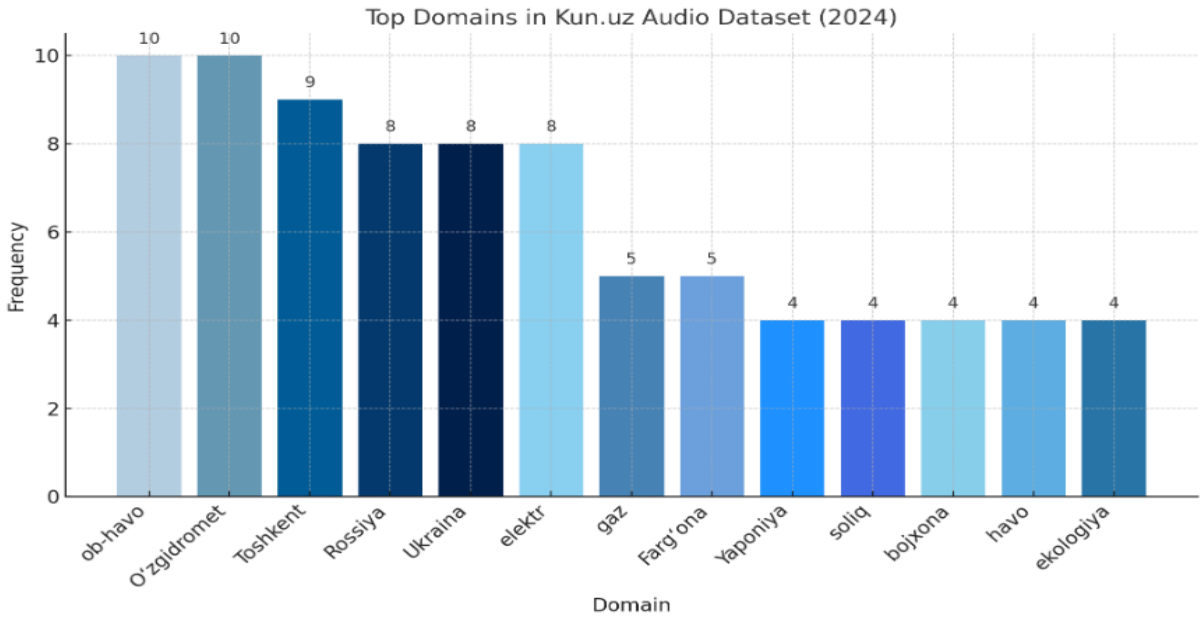

Figure 1. Bar chart of top domains in the Kun.uz 2024 dataset

The **pie chart (Figure 2)** highlights the proportional distribution of these domains. Meteorological topics together account for more than **15%** of the dataset, while international politics (Russia, Ukraine, Japan) comprise around **16%**. Domestic issues (Toshkent, soliq, bojxona) account for roughly **13%**, whereas energy and infrastructure (elektr, gaz) make up approximately **10%**.

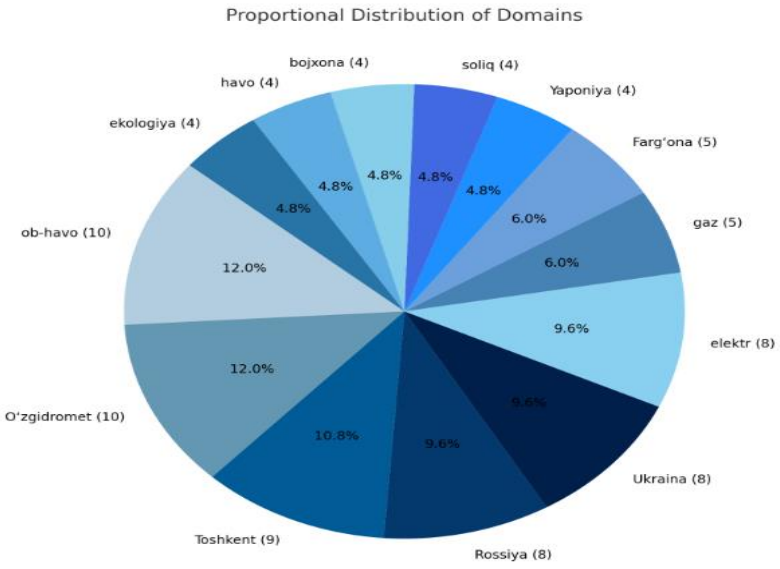

Figure 2. Proportional distribution of domains in the dataset

3. Thematic clusters

The dataset shows four clear thematic clusters:

Meteorology and climate – *ob-havo, O'zgidromet, havo, ekologiya* (~21%).

International politics – *Rossiya, Ukraina, Yaponiya* (~16%).

Domestic governance – *Toshkent, soliq, bojxona* (~13%).

Energy and infrastructure – *elektr, gaz* (~10%).

These clusters highlight a thematic imbalance: domains such as health, culture, and education are significantly underrepresented, limiting the corpus's sociolinguistic breadth.

4. Implications

Such thematic skew has consequences for linguistic research and Automatic Speech Recognition (ASR) systems. Models trained on this dataset may achieve low Word Error Rates (WER) in meteorological or political contexts but perform poorly in underrepresented domains. Similarly, the urban-centric focus (e.g., *Toshkent*) limits regional and dialectal representation, which restricts the corpus's utility for building inclusive Uzbek language technologies.

Discussion

The results of the present study, illustrated in **Table 1**, **Figure 1**, and **Figure 2**, reveal clear thematic tendencies within the *Kun.uz* 2024 audio dataset. Weather-related topics (*ob-havo, O'zgidromet*) and international politics (*Rossiya, Ukraina*) dominate the corpus, while crucial domains such as health, education, and culture remain underrepresented. This discussion expands upon these findings by situating them within broader international research on speech corpora, analyzing the consequences for Automatic Speech Recognition (ASR), and outlining directions for future corpus development in Uzbek.

Well-established corpora such as **TIMIT** (Garofolo et al., 1993), **SWITCHBOARD** (Godfrey & Holliman, 1997), **LibriSpeech** (Panayotov et al., 2015), and **Common Voice** (Ardila et al., 2020) have set benchmarks in terms of size, diversity, and annotation quality. For instance, SWITCHBOARD contains over 2,400 telephone conversations between American English speakers, capturing authentic dialogues across diverse regions and demographics. By contrast, the *Kun.uz* dataset is heavily skewed toward read news speech, limiting natural interactional features such as hesitations, overlaps, or dialectal variation.

Similarly, Libri Speech, derived from audiobooks, achieves consistency in technical quality but is criticized for lacking spontaneous speech. This comparison highlights that the Uzbek dataset currently mirrors Libri Speech more closely than SWITCHBOARD, with strong technical clarity but reduced linguistic authenticity. Meanwhile, large-scale multilingual corpora like MLS (Pratap et al., 2020) and Common Voice embrace crowdsourcing strategies to incorporate thousands of volunteers worldwide. Such models demonstrate pathways for Uzbekistan to expand its resources by involving community participation, thereby ensuring dialectal and demographic inclusivity.

The thematic imbalance observed in the dataset has direct implications for ASR development. Meteorological and political news, being overrepresented, may lead to artificially low **Word Error Rate (WER)** scores when models are tested within these limited domains. However, once exposed to conversational speech in education, healthcare, or cultural contexts, the error rate could increase dramatically. International benchmarks suggest that domain-restricted corpora can achieve WER below 10%, while performance may degrade to 25–30% in underrepresented registers. For Uzbek, where ASR resources are already scarce, such imbalance risks reinforcing systemic weaknesses in speech technology.

Furthermore, **Character Error Rate (CER)**, particularly relevant for morphologically rich languages like Uzbek, may be inflated in domains absent from training data. Without sufficient lexical coverage across varied fields, ASR models are prone to misrecognize inflections, derivational morphology, and named entities outside political and meteorological topics. This finding aligns with Radford et al.'s (2023) observation that even multilingual models like Whisper exhibit domain-specific weaknesses without balanced corpora.

Another key limitation is the strong urban bias: *Toshkent* appears prominently as a domain, while other regions remain marginal (*Farg'ona* with only 5 occurrences). This imbalance overlooks the linguistic richness of Uzbek dialects, including variations in vowel harmony, consonantal shifts, and prosodic patterns across regions such as Samarqand, Khorezm, and Surkhandarya. International experience shows

that neglecting dialectal diversity undermines corpus representativeness. For example, while SWITCHBOARD ensured geographic spread within the United States, the Russian National Corpus incorporates dialectal materials to preserve sociolinguistic breadth.

For Uzbek, capturing rural speech patterns and generational variation is not only a matter of linguistic completeness but also essential for technological inclusivity. ASR systems trained solely on Tashkent-centered speech risk marginalizing users from peripheral regions, echoing the "urban bias" problem documented in many low-resource language corpora.

Annotation depth is another crucial factor. Unlike international corpora that provide multi-layer annotation—phonetic transcription, prosodic marking, morphological and syntactic tags—the *Kun.uz* dataset remains largely unannotated beyond metadata and thematic labels. This limitation restricts its reusability for broader linguistic research, such as prosody, discourse analysis, or phonetic studies.

Ethical responsibility further complicates corpus development. While news broadcasts avoid privacy issues, expanding to spontaneous speech will require informed consent, anonymization, and transparent licensing. Bird and Simons (2003) emphasized that portability and reusability depend not only on technical design but also on ethical guarantees. For Uzbekistan, establishing national guidelines for audio data collection—similar to the consent protocols in Common Voice—would be a decisive step toward sustainable corpus growth.

Building upon these observations, several strategies emerge for strengthening Uzbek audio corpora:

- **Balanced thematic coverage** – Expanding beyond meteorology and politics to include health, education, arts, daily life, and cultural domains.
- **Dialectal and demographic diversity** – Incorporating rural, youth, and elderly speakers to capture sociolinguistic variation.
- **Hybrid data types** – Combining read news speech with spontaneous dialogues, interviews, and user-generated content to enhance authenticity.
- **Multilayer annotation** – Adding phonetic, prosodic, and morphosyntactic layers to facilitate interdisciplinary research.
- **Ethical frameworks** – Developing informed consent, anonymization, and open licensing policies for wider reusability.
- **Crowdsourcing and AI-assisted annotation** – Leveraging public participation and machine learning tools to accelerate corpus growth.
- **Multimodality** – Incorporating video and gesture alongside audio to reflect natural communication contexts, aligning with current trends in multimodal NLP.

The *Kun.uz* 2024 dataset provides a valuable starting point for Uzbek audio corpus development, yet it exhibits significant thematic, regional, and methodological limitations. Compared to international exemplars, it excels in technical quality but lags behind in authenticity, representativeness, and annotation. Addressing these gaps is critical not only for advancing Uzbek ASR performance but also for ensuring that digital language technologies reflect the full sociolinguistic diversity of Uzbek society.

Conclusion

The present study set out to examine the criteria for selecting and analyzing audio materials in the context of building a sustainable Uzbek audio corpus, drawing on the *Kun.uz* 2024 dataset as a case study. By applying five interrelated criteria—authenticity, technical quality, sociolinguistic representativeness, multilayer annotation, and ethical responsibility—we have been able to highlight both the strengths and the limitations of the current resource.

The analysis revealed that while the dataset demonstrates strong technical quality—clear recordings, standardized sampling, and professional production values—it lacks authentic spontaneous speech,

which limits its ability to capture the natural rhythms and variability of Uzbek communication. Thematic imbalance is another key limitation: meteorological and political news dominate, while everyday domains such as education, healthcare, and culture remain marginalized. This skew restricts the representativeness of the corpus and may inflate ASR performance in narrow fields while undermining it in underrepresented contexts.

A further finding concerns regional and dialectal coverage. The dataset is highly urban-centric, with *Toshkent* appearing disproportionately, whereas peripheral regions and dialects are insufficiently represented. Without deliberate inclusion of rural and generational voices, ASR systems risk reinforcing linguistic inequality and failing to serve the broader Uzbek-speaking population. In addition, the absence of multilayer annotation—phonetic, prosodic, morphological, and syntactic—limits the dataset's reusability for both linguistic and computational research.

Ethical considerations also emerge as central to future corpus development. While news-based datasets circumvent privacy issues, the collection of conversational and user-generated speech must adhere to principles of informed consent, anonymization, and transparent licensing. Establishing such frameworks will ensure both the scientific credibility and the social legitimacy of Uzbek corpora.

Looking forward, the study recommends several strategic directions: expanding thematic coverage beyond politics and meteorology; capturing dialectal and demographic diversity; integrating both read and spontaneous speech for authenticity; enriching annotation layers for interdisciplinary utility; and embedding ethical protocols into corpus design. Furthermore, crowdsourcing models (as demonstrated by Mozilla's Common Voice) and AI-assisted annotation offer practical pathways for scaling corpus development efficiently. In the long term, exploring multimodal corpora that integrate audio, video, and gesture will align Uzbek research with global advances in multimodal natural language processing.

In conclusion, the *Kun.uz* dataset constitutes a valuable first step toward an Uzbek audio corpus but cannot, in its current form, meet the broader needs of linguistic research or speech technology. By adopting international best practices and addressing current imbalances, Uzbekistan can build an audio corpus that is not only technically robust but also sociolinguistically inclusive, ethically sound, and globally interoperable. Such a resource would significantly improve ASR performance, strengthen linguistic scholarship, and ensure that the Uzbek language is fully represented in the digital age.

References

1. **Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G.** (2020). *Common Voice: A Massively-Multilingual Speech Corpus.* In: **Proceedings of LREC 2020**, pp. 4218–4222. European Language Resources Association.
2. **Baevski, A., Zhou, H., Mohamed, A., & Auli, M.** (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* In: **Advances in Neural Information Processing Systems (NeurIPS 33)**, pp. 12449–12460. proceedings.neurips.cc+2proceedings.neurips.cc+2
3. **Bird, S., & Simons, G.** (2003). *Seven dimensions of portability for language documentation and description.* **Language**, 79(3), 557–582. https://doi.org/10.1353/lan.2003.0149
4. **Godfrey, J. J., Holliman, E. C., & McDaniel, J.** (1992). *SWITCHBOARD: Telephone speech corpus for research and development.* In: **ICASSP-92** (pp. 517–520). IEEE. https://doi.org/10.1109/ICASSP.1992.225858
5. **Ladefoged, P.** (1967). *Three Areas of Experimental Phonetics: Stress and Respiratory Activity, the Nature of Vowel Quality, Units in the Perception and Production of Speech.* London: Oxford University Press.
6. **Panayotov, V., Chen, G., Povey, D., & Khudanpur, S.** (2015). *LibriSpeech: An ASR corpus based on public domain audiobooks.* In: **ICASSP 2015** (pp. 5206–5210). IEEE. https://doi.org/10.1109/ICASSP.2015.7178964

7. **Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R.** (2020). *MLS: A Large-Scale Multilingual Dataset for Speech Research.* In**: Interspeech 2020** (pp. 2757–2761). https://doi.org/10.21437/Interspeech.2020-2826
8. **Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I.** (2023). *Robust Speech Recognition via Large-Scale Weak Supervision.* In: **Proceedings of the 40th ICML**, PMLR 202, pp. 28492–28518.
9. **Mukhamadiyev, A., Khujayarov, I., Djuraev, O., & Cho, J.** (2022). *Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language.* **Sensors**, 22(10), 3683. https://doi.org/10.3390/s22103683
10. **Musaev, M., Mussakhojayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., & Varol, H. A.** (2021). *USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments.*
11. **Bekarystankyzy, A., Mussakhojayeva, S., Dzhusupov, K., et al.** (2024). *Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets.* **Scientific Reports**, 14, 14691.
12. **Ren, Z., Chen, G., & Li, X.** (2022). *Improving Hybrid CTC/Attention Architecture for Low-Resource ASR.* **Applied Sciences**, 12(19), 9554.