O‘ZBEK TILI NUTQINI ANIQLASH (ASR) TIZIMLARI UCHUN MA’LUMOTLAR TO‘PLAMINI FILTRLASH KONVEYERINI (PIPELINE) ISHLAB CHIQISH VA SAMARADORLIGINI BAHOLASH

Mullaboyeva Xurmatoy Murodilovna,
1-bosqich magistrant,
hurmatoy@gmail.com
ToshDO‘TAU

Annotatsiya. Automatic Speech Recognition (ASR) tizimlarining aniqligi va barqarorligi bevosita trening dataset sifati bilan bog‘liq. Ochiq manbalardan yig‘ilgan audio va matn ma’lumotlari ko‘pincha noto‘g‘ri transkripsiyalar, mos kelmaydigan audio-matn juftliklari hamda nomutanosib namunalarni o‘z ichiga oladi. Ayniqsa low-resource agglutinative tillarda, jumladan o‘zbek tilida, bunday shovqinli ma’lumotlar modelda gallyutsinatsiya, noto‘g‘ri token generatsiyasi va Word Error Rate (WER) ko‘rsatkichining oshishiga olib keladi. Ushbu tadqiqotda Whisper asosidagi ASR modellar uchun dataset filtrlash pipelineni taklif etiladi. Taklif etilgan yondashuv audio-matn juftliklarini tekshirish hamda audio davomiyligi va transkripsiya uzunligi mutanosibligini nazorat qilish bosqichlarini o‘z ichiga oladi. Tajribalar natijasida 70 mingga yaqin audio-matn juftliklari orasidan 61 081 ta yuqori sifatli juftliklar ajratib olindi. Eksperimental natijalar o‘zbek tili uchun optimal mutanosiblik 1 soniya audio uchun maksimal 4 ta so‘z ekanligini ko‘rsatdi. Taklif etilgan filtering strategiyasi ASR modeldagi gallyutsinatsiyalarni kamaytirib, WER ko‘rsatkichini sezilarli yaxshiladi.

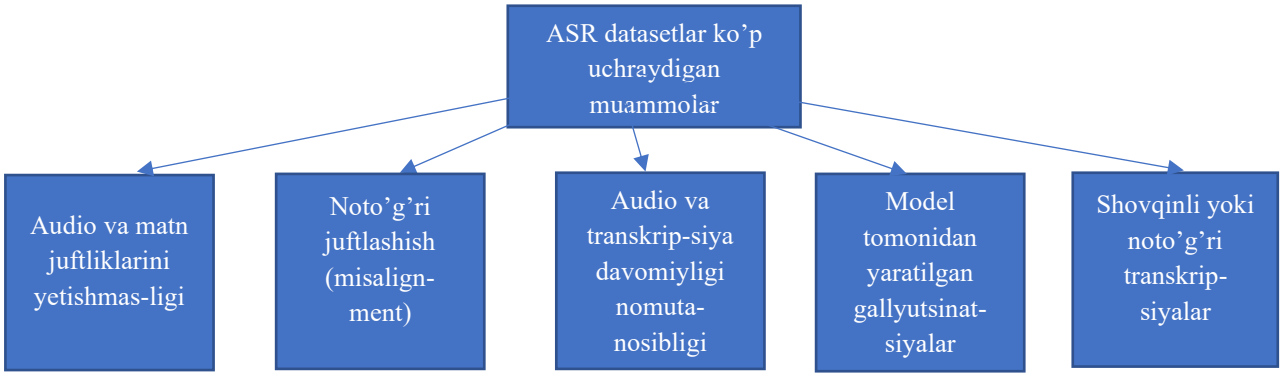
Kalit so‘zlar: *Automatic Speech Recognition, ASR, dataset filtrlash, whisper, kam resursli tillar, o‘zbek tili, nutqni avtomatik tanish, ma’lumotlarni tozalash, word error rate, WER, audio-matn moslashtirish, gallyutinatsiyani kamaytirish*.

Abstract - The accuracy and robustness of Automatic Speech Recognition (ASR) systems are directly related to the quality of the training dataset. Audio and text data collected from open sources often contain incorrect transcriptions, mismatched audio–text pairs, and imbalanced samples. In low-resource

agglutinative languages, including Uzbek, such noisy data may lead to hallucinations, incorrect token generation, and an increased Word Error Rate (WER). This study proposes a dataset filtering pipeline for Whisper-based ASR models. The proposed approach includes stages for validating audio–text pairs and controlling the proportional balance between audio duration and transcription length. As a result of the experiments, 61 081 high-quality pairs were selected from approximately 70 000 audio - text pairs. The experimental results showed that the optimal proportional threshold for Uzbek is a maximum of four words per one second of audio. The proposed filtering strategy reduced hallucinations in the ASR model and significantly improved the WER score.

Keywords: *Automatic Speech Recognition, ASR, dataset filtering, Whisper, low-resource languages, Uzbek language, speech recognition, data cleaning, Word Error Rate, WER, audio text alignment, hallucination reduction*.

1. Kirish. Automatic Speech Recognition (ASR) tizimlari so‘nggi yillarda transformer arxitekturasi va self-supervised learning metodlari rivojlanishi natijasida sezilarli taraqqiyotga erishdi. Whisper, wav2vec2 va Conformer kabi modellar ko‘p tilli nutqni tanishda yuqori natijalar ko‘rsatmoqda. Biroq model arxitekturasi qanchalik kuchli bo‘lmasin, trening dataset sifati ASR tizimining yakuniy aniqligini belgilovchi asosiy omillardan biri bo‘lib qolmoqda. Ochiq manbalardan yig‘ilgan ASR datasetlar ko‘pincha bir qancha muammolarni o‘z ichiga oladi. 1-rasmda ASR datasetlar ko‘p uchraydigan muammolar keltirilgan.



1-rasm. ASR datasetlar ko‘p uchraydigan muammolar

Bunday muammolar ayniqsa kam resursli va agglutinative tillarda sezilarli ta’sir ko‘rsatadi. O‘zbek tilida so‘zlarning morfologik murakkabligi va qo‘shimchalarning ko‘pligi sababli noto‘g‘ri transkripsiyalar modelning attention mexanizmiga salbiy ta’sir qiladi.

Ushbu tadqiqotning asosiy maqsadi Whisper arxitekturasiga asoslangan avtomatik nutqni tanish modellari uchun ma’lumotlar to‘plamini filtrlash metodologiyasini shakllantirish hamda audio va matn muvofiqligining model samaradorligiga ta'sirini eksperimental tadqiq etish orqali ma’lumotlarni optimallashtirish bo‘yicha texnologik yechimlarni ishlab chiqishdan iborat.

Tadqiqot doirasida erishilgan asosiy ilmiy natijalar va amaliy takliflar quyidagilardan iborat:

1. Audio-matn juftliklarining o‘zaro muvofiqligini filtrlash algoritmi;
2. Akustik signalning davomiyligi va transkripsiya hajmi o‘rtasidagi korrelyatsiyaga asoslangan ma’lumotlar sifatini nazorat qilish filtri;
3. O‘zbek tili nutq korpuslari uchun “sekund-so‘z” nisbatining optimal koeffitsiyentlarini aniqlash va metodologik asoslash;
4. Ma’lumotlar to‘plamini saralash (dataset filtering) jarayonining modelning xatolik ko‘rsatkichlari (WER) hamda gallyutsinatsiya holatlari chastotasiga ta’sirini eksperimental baholash.

2. Adabiyotlar sharhi

ASR tizimlari uchun dataset sifati masalasi ko‘plab tadqiqotlarda muhim omil sifatida ko‘rib chiqilgan. B. Elov va boshqalar (4:1-9) tadqiqotlarida o‘zbek tili korpusi matnlarini qayta ishlashda dastlabki preprocessing va shovqinli ma’lumotlarni tozalash bosqichlari umumiy NLP nuqtayi nazaridan ko‘rib chiqilgan. B. Elov va A. Abdullayevlar (1:1-7) o‘zbek tilida sentiment tahlil uchun katta hajmdagi dataset yaratish jarayonida xom ma’lumotlarni tozalash,

normalizatsiya, tokenizatsiya va lemmatizatsiya bosqichlari orqali mos va strukturallashgan dataset shakllantirilishini ta’kidlaydi.

Radford va boshqalar datasetdagi noto‘g‘ri transkripsiyalar va alignment xatolari modelda gallyutsinatsiyalarni yuzaga keltirishi mumkinligini qayd etadi[10:1-41].

Kahn va boshqalar [6:1-7] yaxshi tozalanmagan datasetlar *ASR* tizimlarida *Word Error Rate* ko‘rsatkichining sezilarli oshishiga olib kelishini ko‘rsatgan.

Shuningdek, Manohar va boshqalar audio-text alignment filtering orqali training datasetdagi xatolarni kamaytirish mumkinligini ko‘rsatgan [7:1-5].

Park va boshqalar trening ma’lumotlarining sifati augmentation samaradorligiga kuchli ta’sir qilishini ta’kidlaydi [8:1-6].

Pratap va boshqalar ko‘p tilli ASR modeli uchun datasetni tayyorlashda matnlarni NFKC normalizatsiyasidan o‘tkazib, barcha tinish belgilarini olib tashlagan hamda har bir tilning orfografiyasiga mos valid Unicode belgilar ro‘yxati asosida ushbu diapazondan tashqaridagi belgilarni o‘z ichiga olgan so‘zlarni filtrlab tashlagan [9:1-5]

Kam resursli tillar uchun dataset filtering masalasi hali yetarlicha o‘rganilmagan bo‘lib, ayniqsa o‘zbek tili uchun audio va matn mutanosibligiga oid tadqiqotlar deyarli mavjud emas.

3. Metodologiya

3.1 Dataset. Tadqiqot uchun ochiq manbalardan yig‘ilgan audio va avtomatik transkripsiyalardan tashkil topgan datasetdan foydalanildi. Dataset podkastlar, yangiliklar, audio kitoblar va ijtimoiy media materiallarini o‘z ichiga olgan 71 289 ta audio-matn juftliklaridan iborat.

3.2 Juftliklar orasidagi moslikka asoslangan filtrlash.

Avtomatik transkripsiya jarayonida ayrim audio yoki matn fayllari juftliksiz qolishi kuzatildi. Bunday namunalar model treningiga salbiy ta’sir qilishi sababli

datasetdan chiqarib tashlandi. Filtering algoritmi audio fayllar va matn fayllarning bazaviy nomlarini taqqoslash asosida ishlaydi.

Quyidagi kod juftliksiz audio fayllarni aniqlash uchun ishlatildi:

```
audio_dir = "/home/corvax/NLP/Mullaboyeva-
             STT/data/train/street/audio"
text_dir = "/home/corvax/NLP/Mullaboyeva-
            STT/data/train/street/text"
bad_dir = "/home/corvax/NLP/Mullaboyeva-
           STT/data/train/street/bad"
os.makedirs(bad_dir, exist_ok=True)
audio_files = {
        os.path.splitext(f)[0]: f
        for f in os.listdir(audio_dir)
        if f.endswith(".wav")
}
text_files = {
        os.path.splitext(f)[0]
        for f in os.listdir(text_dir)
        if f.endswith(".txt")
}

for name, filename in audio_files.items():
    if name not in text_files:
      shutil.move(
                  os.path.join(audio_dir, filename),
                  os.path.join(bad_dir, filename)
      )
```

Mazkur bosqich noto‘g‘ri alignment va yetishmayotgan transkripsiyalarni kamaytirishga xizmat qildi.

3.3 Audio va matnning mutanosibligi filtri.

Audio va matn juftligi mavjud bo‘lgan holatda ham ular bir-biriga mazmuniy jihatdan mos kelmasligi mumkin. Ayrim hollarda quyidagilar kuzatildi:

a) uzun audio uchun juda qisqa matn;

b) qisqa audio uchun juda uzun transkripsiya;

c) model gallyutsinatsiyasi;

d) noto‘g‘ri kesilgan segmentlar

Shu sababli audio davomiyligi va matndagi so‘zlar soni orasidagi mutanosiblik eksperimental tahlil qilindi. Tajribalarda quyidagi nisbatlar sinovdan o‘tkazildi:

1-jadval. 1 soniyadagi so‘zlar sonining werga ta’siri

Nisbat	Natija
1 soniya ≤ 10 so‘z	WER > 25%, yuqori gallyutsinatsiya
1 soniya ≤ 8 so‘z	22–25% WER
1 soniya ≤ 6 so‘z	18–22% WER
1 soniya ≤ 4 so‘z	15–18% WER, gallyutsinatsiya kuzatilmadi
1 soniya ≤ 2 so‘z	WER>20%, deletion xatolar
1 soniya = 1 so‘z	WER>30%, yuqori deletion rate

Natijalarga ko‘ra, optimal qiymat:1 second audio≈4 words ekanligi aniqlandi. Filtering algoritmi quyidagicha amalga oshirildi:

```
dataset = load_from_disk(
    "/home/corvax/NLP/Mullaboyeva-STT/prepared_dataset"
)

bad = 0
total = 0
for row in dataset["train"]:
    audio, sr = librosa.load(row["audio_path"], sr=16000)
    seconds = len(audio) / sr
    words = len(row["text"].split())
    if words > seconds * 4:
        bad += 1
    total += 1
print("Bad samples:", bad)
print("Total:", total)
print("Ratio:", bad/total)
```

Natijalar

Filtering pipeline natijasida dataset sifati sezilarli yaxshilandi va shu bilan birgalikda quyidagi natijalarga erishildi:

1. Gallyutsinatsiyalar kamaydi.

2. Noto‘g‘ri generatsiyalar yo‘qoldi.
3. Modelning stabil ishlashi yaxshilandi.

2-jadval. Filtrlashda dataset hajmi

Bosqich	Namunalar soni
Dastlabki juftliklar	~71 289
Pair filteringdan keyin	64 300
Proportional filteringdan keyin	61 081

Eksperimental trening natijalari filtering samaradorligini tasdiqladi.

3-jadval. Filtrlashning werga ta’siri

Dataset turi	WER
Filtrlanmagan dataset	24–27%
Pair filtering	20–22%
To‘liq filtering pipeline	15–18%

5. Muhokama

Natijalar shuni ko‘rsatdiki, dataset filtering ASR modellar sifati uchun muhim bosqich hisoblanadi. Ayniqsa agglutinative tillarda audio va matn mutanosibligi katta ahamiyatga ega.

Tadqiqot davomida quyidagilar aniqlandi:

1. Uzun transkripsiyalar modelni gallyutsinatsiyaga olib keladi.
2. Juda qisqa transkripsiyalar deletion xatolarni oshiradi.
3. Optimal balans attention mexanizmini stabil ishlashiga yordam beradi.

Taklif etilgan filtering pipeline Whisper-based ASR modellar bilan bir qatorda boshqa transformer-based speech recognition tizimlari uchun ham qo‘llanishi mumkin.

6. Xulosa

Mazkur tadqiqotda o‘zbek tilidagi ASR modellar uchun dataset filtering pipeline taklif qilindi. Pipeline audio-matn juftliklarini tekshirish hamda audio davomiyligi va matn uzunligi mutanosibligini nazorat qilish bosqichlarini o‘z ichiga oladi. Eksperimental natijalariga ko‘ra quyidagilarga erishildi:

1. Dataset sifatini yaxshilandi.

2. Gallyutsinatsiyalarni kamaytirdi.
3. WER ko‘rsatkichini pasaytirdi.

Kelajakda ishonchlilik ko‘rsatkichi asosida filtrlash, semantik moslashtirish, perpleksiya asosida filtrlash, o‘z-o‘zini nazorat qilish asosida ma’lumotlarni tozalash metodlarini qo‘shish rejalashtirilmoqda.

Foydalanilgan adabiyotlar ro‘yxati

1. Abdullayev A. Q., Elov B. B. Sentiment tahlil uchun katta hajmdagi datasetni yaratish bosqichlari. 2025.
2. Alayev R. H., Elov B. B., Hamdamov O‘. Ma’lumotlar to‘plamini o‘qitish, baholash va test to‘plamlariga ajratish usullari. 2024.
3. Elov B. B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem. 2023.
4. Elov B. B., Khamroeva Sh. M., Alayev R. H., Khusainova Z. Yu., Yodgorov U. S. Methods of Processing the Uzbek Language Corpus Texts. International Journal of Open Information Technologies. 2023.
5. Elov B. B., Khamroeva Sh. M., Dauletov A. Yu., Matyakubova N. Sh. Algorithm for Aligning Paragraphs and Sentences in Aligner Tool. 2024.
6. Kahn J. et al. Libri-Light: A Benchmark for ASR with Limited or No Supervision Proceedings of ICASSP 2020. 2020. arXiv:1912.07875.
7. Manohar V. et al. Semi-Supervised Training of Acoustic Models Using Lattice-Free MMI Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP. 2018. P. 4844 - 4848. DOI: 10.1109/ICASSP.2018.8462331.
8. Park D. S. et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition Proceedings of Interspeech 2019. DOI: 10.21437/Interspeech.2019-2680.

9. Pratap V. et al. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters Proceedings of Interspeech 2020. 2020. P. 4751 - 4755.

10. Radford A. et al. Robust Speech Recognition via Large-Scale Weak Supervision Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023. Vol. 202. P. 28492 - 28518. arXiv:2212.04356.

11. Mirdjonovna, K. S., Boltayevich, E. B., Habibovich, A. R., & Qizi, S. D. F. (2025, September). Morphotactic Models and Algorithms of the Uzbek Language. In 2025 10th International Conference on Computer Science and Engineering (UBMK) (pp. 1596-1601). IEEE. https://ubmk.org.tr/wp-content/uploads/2025/09/Ilk-Kisim-2025.pdf