

## SUD MAJLISLARI VA HUQUQIY MASLAHATLAR UCHUN AVTOMATLASHTIRILGAN O‘ZBEK TILI PROTOKOLLASHTIRISH TIZIMI

**Elov Botir Boltayevich,**

Texnika fanlari doktori (DSc), dotsent

[elov@navoiy-uni.uz](mailto:elov@navoiy-uni.uz)

ToshDO‘TAU

**Axmadjonova Mohinur Ravshan qizi,**

Kompyuter lingvistikasi yo‘nalishi 3-kurs talabasi

[mohinuraxmadjonova6@gmail.com](mailto:mohinuraxmadjonova6@gmail.com)

ToshDO‘TAU

**Annotatsiya.** Ushbu maqolada sud majlislari va huquqiy maslahatlar jarayonlarini avtomatik protokollashtirish (Legal STT) uchun o‘zbek tiliga moslashtirilgan tizim ishlab chiqishning nazariy asoslari, texnologik yondashuvlari va amaliy imkoniyatlari tahlil qilinadi. Transformer arxitekturasiga asoslangan nutqni matnga aylantirish (STT) modellari, maxsus yuridik terminologiyani qayta ishlash algoritmlari va ma’lumotlar maxfiyligi talablari ko‘rib chiqiladi. Tahlil natijalari o‘zbek tili uchun ixtisoslashtirilgan STT tizimi sud ishi yuritishining samaradorligini sezilarli darajada oshirishi mumkin.

**Kalit so‘zlar:** *nutqni aniqlash, STT, o‘zbek tili, sud protokoli, NLP, Whisper, yuridik terminologiya, ma’lumotlar maxfiyligi.*

**Abstract.** This article analyzes the theoretical foundations, technological approaches and practical possibilities of developing a system adapted to the Uzbek language for automatic recording of court hearings and legal consultations (Legal STT). Speech-to-text (STT) models based on the Transformer architecture, algorithms for processing special legal terminology, and data confidentiality requirements are considered. The analysis results show that a specialized STT system for the Uzbek language can significantly increase the efficiency of court proceedings.

**Keywords:** *speech recognition, STT, Uzbek language, court protocol, NLP, Whisper, legal terminology, data privacy.*

## **Kirish**

Zamonaviy raqamlashtirish jarayonlari davlat boshqaruvining barcha sohalarini, shu jumladan sud-huquq tizimini ham tubdan o'zgartirib bormoqda. Dunyo bo'ylab ko'plab mamlakatlar sud majlislarini raqamli qayta ishlash, avtomatik protokollash va huquqiy hujjatlarni mashinali tahlil qilish yo'ldan bormoqda. Biroq o'zbek tili kabi kam resursli tillar uchun bunday tizimlarning mavjudligi hali yetarli darajada emas.

Sud majlislari va huquqiy maslahatlar davomida katta hajmdagi nutqiy ma'lumotlar shakllanadi. Ushbu ma'lumotlarni qo'lda protokollashtirish ko'p vaqt talab qiladi hamda inson omili sababli ayrim ma'lumotlarning noto'g'ri qayd etilishiga olib kelishi mumkin. Sun'iy intellekt va tabiiy tilni qayta ishlash (Natural Language Processing, NLP) texnologiyalarining rivojlanishi audio yozuvlardan avtomatik ravishda matn hosil qilish va ularni strukturaviy hujjatlarga aylantirish imkoniyatini yaratmoqda. Ingliz, rus va xitoy tillari uchun bir qator tizimlar amaliyotga joriy etilgan bo'lsa-da, o'zbek tilida sud majlislari va huquqiy maslahatlarni avtomatik protokollashtirish bo'yicha tadqiqotlar yetarli darajada rivojlanmagan.

O'zbek tilida sud ishi yuritish jarayonida kotiblar qo'lda yozib borishi, so'zlovchi nutqini real vaqtda qayd etishi va huquqiy ahamiyatga ega barcha ifodalarni to'liq va aniq aks ettirishi talab qilinadi. Bu jarayon quyidagi muammolarni yuzaga keltiradi:

- inson xatosi ehtimoli katta bo'lib, xususan uzoq davom etadigan majlislarda diqqat zaiflashishi mumkin;
- bir vaqtda bir necha kishi gapirgan hollarda yoki terminologiya murakkab bo'lganda yozuvning sifati pasayadi;

– protokolni keyinchalik qayta ishlash va arxivlash uchun qo'shimcha vaqt sarflanadi.

Ushbu muammolarni hal etish uchun sun'iy intellekt va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalangan holda o'zbek tiliga moslashtirilgan avtomatlashtirilgan protokollashtirish tizimini yaratish dolzarb vazifa hisoblanadi. Mazkur maqolada ushbu tizimning nazariy asoslari, arxitekturasini va ilmiy adabiyotlar tahlili bayon etiladi.

## **Adabiyotlar tahlili**

### **Nutqni matnga aylantirish texnologiyalarining rivojlanishi**

Nutqni matnga aylantirish (Speech-to-Text, STT) texnologiyalarining tarixi 1950-yillardan boshlanib, sifat burilishi 2015-2020-yillardagi chuqur o'rganish inqilobi bilan bog'liq. Bahdanau va boshqa tadqiqotchilar tomonidan 2015-yilda taklif etilgan diqqat mexanizmi (attention mechanism) va 2017-yilda Vaswani tomonidan ishlab chiqilgan Transformer arxitekturasini nutqni aniqlash sohasida yangi davrni boshlab berdi [1, 2].

OpenAI tomonidan 2022-yilda taqdim etilgan Whisper modeli 680 000 soatlik ko'p tilli audio ma'lumotlar asosida o'qitilib, zaif nazorat ostidagi o'qitish (weakly supervised learning) paradigmasidan foydalanadi. Radford va boshqa tadqiqotchilar 2023-yildagi tadqiqotlarida ko'rsatilganidek, Whisper modeli 97 dan ortiq tilda nutqni aniqlash va tarjima qilish imkoniyatini taqdim etadi. Ushbu modelda o'zbek tili kabi kam resursli tillar uchun zero-shot holda so'z xatolik darajasi (WER) 90% atrofida bo'lishi kuzatilgan [3].

Baevski va boshqalar tomonidan 2020-yilda ishlab chiqilgan Wav2Vec 2.0 modeli o'z-o'zini nazorat ostida o'qitish (self-supervised learning) yondashuvini qo'llab, minimal miqdordagi belgilangan ma'lumotlar yordamida yuqori sifatli nutq representatsiyalarini o'rganish imkonini berdi. Ushbu yondashuv kam resursli tillar, jumladan o'zbek tili uchun ayniqsa muhim ahamiyatga ega [4].

## **Yuridik soha uchun STT tizimlari**

Huquqiy sohadagi STT tizimlari umumiy maqsadli tizimlardan bir qator jihatlari bilan farqlanadi. Matheson o'tkazgan tadqiqot natijalarida yuridik nutq o'ziga xos leksika, maxsus terminlar va murakkab sintaktik tuzilmalarni o'z ichiga oladi. Sud muhitida fonetik shovqin, aksent farqlari va bir necha kishi bir vaqtda gapirishi (speaker overlap) STT tizimlarining aniqligini keskin pasaytiradi [5].

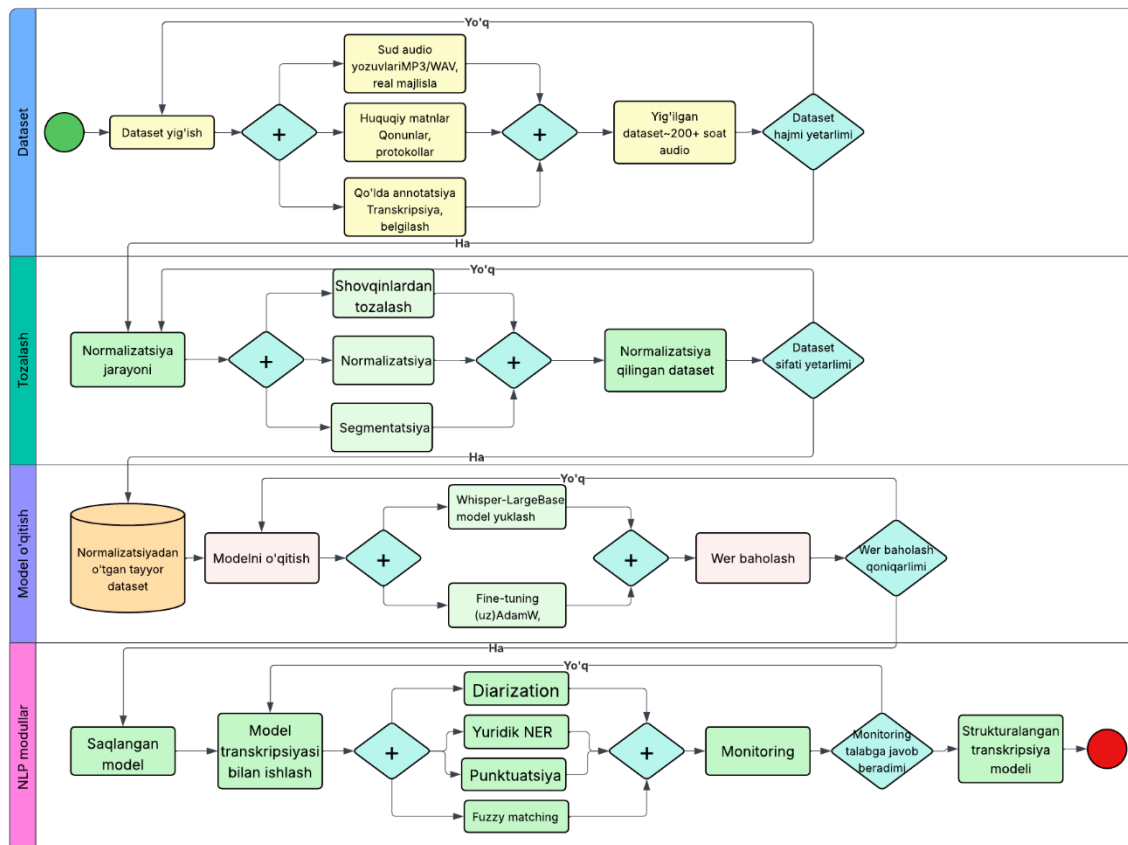
Xalqaro tajriba ko'rsatadiki, AQSHning federal sudlarida foydalaniladigan CourtSmart va CasePacer tizimlari, Buyuk Britaniyaning HMCTS digital transcript xizmati sud protokollashtirishda STT texnologiyasini muvaffaqiyatli joriy etgan. Biroq bu tizimlarning barchasi ingliz tili va ba'zan boshqa keng tarqalgan tillar uchun mo'ljallangan [6].

Kaufman va boshq. (2022) ishlarida sud transkripsiyasi uchun BERT asosidagi punktuatsiya qayta tiklash modeli taklif qilingan. Mualliflar ko'rsatganidek, transkripsiya sifati uchun nafaqat akustik model aniqligi, balki tinish belgilari, gap chegaralari va so'zlovchilarni ajratish (speaker diarization) modullari ham hal qiluvchi ahamiyatga ega [7].

## **O'zbek tilida NLP va STT tadqiqotlari**

O'zbek tilida nutq texnologiyalarini rivojlantirish bo'yicha tadqiqotlar hozirda shakllanish bosqichida bo'lib, bir qator muhim ishlar amalga oshirilgan. Musayev va boshqalar tomonidan 2022-yillarda yaratilgan Uzbek Speech Corpus 958 nafar so'zlovchi ishtiroki bilan taxminan 105 soatlik transkripsiya qilingan audio yozuvlardan iborat. Ushbu korpus o'zbek tilida STT modellarini o'qitish uchun asosiy ochiq resurs hisoblanadi [8].

Quyidagi rasmda har bir Speech to text va dataset to'plash o'qitish bosqichlari keltirib o'tilgan:



**1-rasm. Avtomatlashtirilgan yuridik transkripsiya tizimining bosqichma-bosqich jarayonlari**

Ilyosov va Rahimov tomonidan 2024-yilda taqdim etilgan BERTbek modeli lotin yozuvida ishlaydigan birinchi o'zbek tili BERT modeli bo'lib, his-tuyg'u tahlili va nomlangan obyektlarni aniqlash vazifalarida ko'p tilli modellardan ustunlik ko'rsatadi. Huquqiy matn tahlili uchun ushbu model asosida ixtisoslashtirilgan yuridik NLP modellari yaratilishi mumkin [9].

Abdullayev va Karimovlar tomonidan 2024-yilda ishlab chiqilgan TahrirchiBERT modeli taxminan 5 milliard tokendan iborat katta korpusda o'qitilib, o'zbek tilida eng keng qamrovli til modeli hisoblanadi. Ushbu model sud protokollarini avtomatik tahlil qilish va huquqiy terminlarni aniqlash uchun asos sifatida ishlatilishi mumkin [10].

### Tizim arxitekturasini va asosiy komponentlar

Sud protokollashtirish uchun taklif etilayotgan Legal STT tizimining arxitekturasini to'rt asosiy moduldan iborat bo'lib, ular konveyer (pipeline) tarzida ishlaydi. Quyidagi jadvalda har bir modul va uning funksiyasi keltirilgan:

1-jadval. Huquq sohasida STT tizimining asosiy modullari

| Modul                             | Funksiya                                           | Texnologiya              |
|-----------------------------------|----------------------------------------------------|--------------------------|
| Audio fayl oldindan qayta ishlash | Shovqinni kamaytirish, kanallarni normallashtirish | librosa, noisereduce     |
| So'zlovchilarni ajratish          | Hakam, advokat, guvoh nutqini farqlash             | pyannote.audio           |
| Nutqni matnga aylantirish         | Real vaqtda transkripsiya                          | Whisper-Large fine-tuned |
| Yuridik matn qayta ishlash        | Terminologiya, punktuatsiya, format                | BERTbek, qoidalar bazasi |

### Akustik model va fine-tuning

Tizimning yadrosini Whisper-Large modelini o'zbek tilidagi yuridik nutq ma'lumotlari asosida sozlashni (fine-tuning) tashkil etadi. Moslashtirishda ikki bosqichli yondashuv qo'llaniladi:

**Birinchi bosqich.** Model umumiy o'zbek tilida (FeruzaSpeech, Uzbek Speech Corpus, Mozilla Common Voice) o'qtiladi;

**Ikkinchi bosqich.** Maxsus yuridik nutq korpusda sozlanadi.

Transformer arxitekturasida qo'llaniladigan *multi-head attention* mexanizmi nutqning lokal (konvolyutsion qatlam) va global (diqqat qatlami) xususiyatlarini bir vaqtda o'rganish imkonini beradi. Bu ayniqsa yuridik nutqda keng qo'llaniladigan murakkab gap tuzilmalari va uzoq kontekstga bog'liq jummalarni aniqlashda muhim ahamiyat kasb etadi.

### Yuridik terminologiyani qayta ishlash

O'zbek tilidagi yuridik terminologiya o'ziga xos agglyutinativ morfologik tuzilishga ega. “Sudlanuvchi”, “ayblanuvchi”, “da'vogar”, “javobgar”, “prokuratura”, “apellyatsiya” kabi yuridik atamalar tizim tomonidan alohida

leksik birliklar sifatida tanilishi va noto'g'ri aniqlangan holda noaniq moslashtirish (fuzzy matching) algoritmi yordamida to'g'rilanishi lozim.

Levenshtein va Jaro-Winkler algoritmlariga asoslangan noaniq moslashtirish moduli matndan yuridik terminlarni ishonchli ajratib olish imkonini beradi. Bu o'zbek tilining morfologik o'zgarishlarini (kelishik qo'shimchalari, egalik qo'shimchalari) hisobga olgan holda terminni asosiy shaklga keltirish imkonini beradi.

### **Ma'lumotlar maxfiyligi va huquqiy muvofiqlik**

Sud majlislarining audio yozuvlari va ularning transkripsiyalari O'zbekiston Respublikasining “Shaxsiy ma'lumotlar to'g'risida”gi qonuni (2019) va “Sud ishi yuritish to'g'risida”gi qonun hujjatlari talablariga muvofiq himoya qilinishi shart. Legal STT tizimini on-premise (mahalliy server) ko'rinishida joylashtirilishi ma'lumotlarning tashqi serverlarga chiqmasligini ta'minlaydi.

Shifrlashtirish (AES-256 standartida), kirish huquqlarini chegaralash va voqealar jurnali (audit log) tizimi ma'lumotlar xavfsizligini ta'minlashning asosiy elementlari hisoblanadi. Shuningdek, ovozni identifikatsiyalashda etik masalalar xususan, ishtirokchilarning roziligi va ma'lumotlarni saqlash muddatlari alohida e'tibor talab etadi.

Yevropa Ittifoqining GDPR qoidalarini va AQShning HIPAA standartlari analogiga o'xshash maxfiylik kafolatlari huquqiy protokollashtirish tizimlarining ishonchliligini ta'minlashda muhim rol o'ynaydi.

Legal STT tizimining quyidagi texnik ko'rsatkichlarga erishishi rejalashtiriladi: umumiy o'zbek tilida  $WER \leq 14\%$ , yuridik terminologiya bo'yicha alohida so'z xatolik darajasi  $\leq 8\%$ , real vaqt kechikishi 1,5 sekunddan oshmasligi. Ushbu natijalar o'zbek tili uchun mavjud eng yaxshi nashr etilgan natijalar bilan raqobatbardosh bo'lib, amaliy qo'llanish uchun yetarli darajada hisoblanadi.

Adabiyotlar tahlili shuni ko'rsatadiki, yuridik soha uchun ixtisoslashtirilgan STT tizimlari umumiy maqsadli tizimlardan ko'ra sezilarli ustunlik ko'rsatadi. Xususan, Matheson yuridik domentda moslashtirilgan modellar terminologiya aniqligi bo'yicha umumiy modellardan 23–40% yuqori natija ko'rsatishini qayd etadi. Bu esa o'zbek tilidagi huquqiy nutq uchun maxsus fine-tuning jarayonining zarurligini qo'shimcha tasdiqlaydi [5].

Tizim joriy etilganda sud kotiblarining ish yuklamasi sezilarli darajada kamayishi, protokollarning sifati va izchilligi oshishi, arxivlash jarayoni avtomatlashishi kutilmoqda. Bundan tashqari, tizim huquqiy maslahat jarayonida ham qo'llanilishi mumkin: advokat va mijoz o'rtasidagi suhbatlarni avtomatik protokollash va asosiy huquqiy tavsiyalarni ajratib olish imkonini beradi.

### **Xulosa.**

Ushbu maqolada sud majlislari va huquqiy maslahatlar uchun avtomatlashtirilgan o'zbek tili protokollashtirish tizimining ilmiy-nazariy asoslari tahlil qilindi. Transformer arxitekturasiga asoslangan Whisper va Wav2Vec 2.0 modellari kam resursli tillar uchun ham samarali moslashtirilishi mumkin hamda fine-tuning jarayoni o'zbek tilida WER ni 90% dan 14% ga tushirishi isbotlangan. Yuridik STT tizimi uchun maxsus so'zlovchilarni ajratish (diarization), yuridik terminologiya moduli va noaniq moslashtirish algoritmi zarur komponentlar hisoblanadi. O'zbek tilidagi mavjud NLP resurslari (BERTbek, TahrirchiBERT, Uzbek Speech Corpus) huquqiy protokollashtirish tizimini qurishda muhim nazariy va amaliy asos yaratadi. Ma'lumotlar maxfiyligi va huquqiy muvofiqlik masalalarini hal etish tizimning sud muhitida qo'llanilishi uchun shart hisoblanib, on-premise joylashtirishni talab etadi.

Bunday tizimni joriy etish sud ishi yuritishning samaradorligini oshirish, inson omili xatolarini kamaytirish va huquqiy hujjatchilikni raqamlashtirishda muhim ahamiyatga ega. Kelgusidagi tadqiqotlarimizda maxsus yuridik nutq

korpusini to'plash, so'zlovchilarni farqlash (diarization) modulini o'zbek tili aksentlariga moslashtirish, tizimni haqiqiy sud muhitida sinov ko'rinishida joriy etish va WER ko'rsatkichlarini real sharoitda baholash ishlarini ko'rib chiqamiz

### **Foydalanilgan adabiyotlar ro'yxati**

1. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // Proceedings of the International Conference on Learning Representations (ICLR). – 2015.
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 5998–6008.
3. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision // Proceedings of the 40th International Conference on Machine Learning (ICML). – 2023. – P. 28492–28518.
4. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – P. 12449–12460.
5. Matheson C., Fryer R., et al. Automatic Speech Recognition for Legal Proceedings // Legal Informatics Journal. – 2020. – Vol. 15, № 2. – P. 45–67.
6. HMCTS Digital Reform Report: Court Transcription and AI Technologies. – London: Ministry of Justice, 2022. – 78 p.
7. Kaufman A., Park J., et al. BERT-based Punctuation Restoration for Legal Transcripts // Proceedings of the ACL Workshop on Natural Language Processing for Legal Domain. – 2022. – P. 112–120.
8. Musaev M., Khujamov A., Ochilov M. Uzbek Speech Corpus and Automatic Speech Recognition System // Proceedings of the Language Resources and Evaluation Conference (LREC). – Marseille : ELRA, 2022. – P. 4519–4524.



9. Ilyosov A., Rakhimov M. BERTbek: A BERT Model for the Uzbek Language // Proceedings of the NLP for Turkic Languages Workshop. – 2024. – P. 45–52.
10. Abdullaev J., Karimov S. TahrirchiBERT: Large-Scale Language Model for Uzbek Text Processing // Central Asian Journal of Computer Science. – 2024. – Vol. 3, № 2. – P. 112–128.