NUTQDAN AXBOROT OLISH VA SHEVALARNI QAYTA ISHLASH TIZIMLARINING TEXNOLOGIK TAHLILI

Xusaydinova D
Muhammad al-Xorazmiy nomidagi
Toshkent axborot texnologiyalari universiteti
d.xusaydinova@tuit.uz

Annotatsiya: *Ushbu maqolada sun'iy intellekt asosida nutqdan axborot olish va shevalarni qayta ishlash tizimlari tahlil qilinadi. Tadqiqot davomida speech-to-text texnologiyalari, tabiiy tilni qayta ishlash (NLP) usullari hamda zamonaviy tizimlarning ishlash prinsiplari o'rganildi. Natijalar o'zbek tili shevalarini qayta ishlash uchun katta hajmdagi lingvistik ma'lumotlar bazasini yaratish muhimligini ko'rsatdi.*

Kalit so'zlar: *axborot olish tizimlari, speech recognition, NLP, dataset, neural network, dialekt, o'zbek tili.*

Abstract: *This article analyzes speech information retrieval and dialect processing systems based on artificial intelligence. During the research, speech-to-text technologies, natural language processing (NLP) methods, and the operating principles of modern systems were studied.*

The results showed the importance of creating a large linguistic database for processing Uzbek language dialects.

Keywords: *information retrieval systems, speech recognition, natural language processing (NLP), dataset, neural networks, dialect, Uzbek language*

Аннотация: *В данной статье анализируются системы извлечения информации из речи и обработки диалектов на основе искусственного интеллекта. В ходе исследования были изучены технологии speech-to-text, методы обработки естественного языка (NLP) и принципы работы современных систем. Результаты показали важность создания большой лингвистической базы данных для обработки диалектов узбекского языка.*

Ключевые слова: *системы извлечения информации, распознавание речи, обработка естественного языка (NLP), набор данных, нейронные сети, диалект, узбекский язык*

Kirish

Axborot texnologiyalarining jadal rivojlanishi natijasida nutqdan axborot olish tizimlari inson va kompyuter o'rtasidagi aloqani soddalashtiruvchi muhim infratuzilmaga aylandi (7). Nutq asosidagi tizimlar speech recognition va machine learning texnologiyalariga tayanadi. Bugungi kunda jahon miqyosida Google ва Microsoft kabi yirik kompaniyalar ushbu yo'nalishda yetakchilik qilmoqda, biroq o'zbek tili shevalarini qayta ishlash masalasi hali ham dolzarb muammo bo'lib qolmoqda (3).

Shu bilan birga, O‘zbekistonda nutq texnologiyalari sohasida ayrim ilmiy tadqiqotlar olib borilayotgan bo‘lsa-da, o‘zbek tilining turli shevalarini qamrab oluvchi yirik va sifatli datasetlar yetarli darajada shakllanmagan. Bu esa nutqni aniqlash va axborot olish tizimlarining aniqligini sezilarli darajada pasaytiradi.

Mazkur tadqiqotning asosiy maqsadi nutqdan axborot olish va o‘zbek tili shevalarini qayta ishlash tizimlarini tahlil qilish, mavjud muammolarni aniqlash hamda zamonaviy yondashuvlar asosida ularni takomillashtirish yo‘llarini ko‘rsatishdan iborat. Shuningdek, tadqiqot doirasida kichik hajmdagi dataset asosida eksperiment o‘tkazilib, model samaradorligi baholandi.

Adabiyotlar tahlili. Nutq texnologiyalari sohasida Hinton va Deng (2012) chuqur neyron tarmoqlarining nutqni aniqlashdagi samaradorligini isbotlagan bo‘lsa (7), Vaswani va boshqalar (2017) tomonidan taklif etilgan Transformer arxitekturasi matnli tahlilni yangi bosqichga olib chiqdi (6). Mahalliy tadqiqotchilardan G. Matlatipov (2025) o‘zbek tili uchun parallel korpuslar yaratish va mashina tarjimasini takomillashtirish ustida ish olib bormoqda (1).

Tadqiqot metodologiyasi.Nutqdan axborot olish tizimlari bir nechta asosiy bosqichlardan iborat. Avvalo audio signal qabul qilinadi va maxsus algoritmlar yordamida raqamli shaklga o‘tkaziladi. Keyinchalik signalni tozalash va shovqinni kamaytirish jarayoni amalga oshiriladi. Ushbu bosqichda Fourier transform, spectrogram extraction kabi signalni qayta ishlash usullari qo‘llaniladi.

Keyingi bosqichda nutqni aniqlash jarayoni amalga oshiriladi. Speech recognition algoritmlari audio signalni matnga aylantiradi. Zamonaviy tizimlarda ushbu jarayon chuqur o‘rganish (deep learning) asosidagi neyron tarmoqlar yordamida amalga oshiriladi. Transformer arxitekturasi asosida yaratilgan modellarning qo‘llanilishi nutqni aniqlash aniqligini sezilarli darajada oshirmoqda.

Nutqni matnga aylantirgandan so‘ng axborot olish jarayoni boshlanadi. Bu bosqichda tabiiy tilni qayta ishlash usullari qo‘llanilib, matndan kerakli ma’lumotlar ajratib olinadi. NLP algoritmlari yordamida matn semantik jihatdan tahlil qilinadi va foydalanuvchiga kerakli axborot taqdim etiladi.

Signalni qayta ishlashda quyidagi formula qo‘llaniladi:

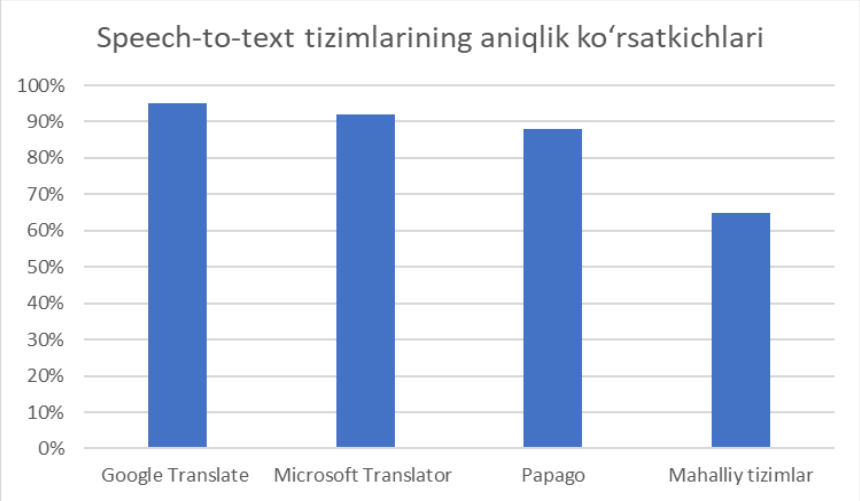
Diskret Fourier transform:

$X(k) = \Sigma\, x(n)\, e^{-j2\pi kn/N}$

Ushbu usul signalni chastota sohasida tahlil qilish imkonini beradi.

Turli hududlarda ishlab chiqilgan nutqni qayta ishlash tizimlari samaradorligi jihatidan farq qiladi. Masalan, Google Translate va Microsoft Translator tizimlari katta hajmdagi ma’lumotlar to‘plami (dataset) va cloud infratuzilmasiga asoslangan bo‘lib, real vaqtga yaqin tezlikda ishlaydi. Ushbu tizimlarda latency 0.5–1 sekund oralig‘ida bo‘lishi mumkin.

Turli platformalarda nutqni aniqlash tizimlarining samaradorligi turlicha bo‘ladi. Quyidagi diagrammada ayrim mashhur speech-to-text tizimlarining aniqlik darajasi ko‘rsatilgan.



Osiyo hududida ishlab chiqilgan Papago va VoiceTra tizimlari esa dialekt va madaniy kontekstni aniqlashda samarali hisoblanadi. Yevropa hududida esa DeepL va Reverso tizimlari tarjima aniqligi va lingvistik moslik bo‘yicha yuqori natijalarni ko‘rsatadi.

Markaziy Osiyo hududida nutqni qayta ishlash va axborot olish tizimlari hali rivojlanish bosqichida. Mahalliy loyihalar mavjud bo‘lsa-da, o‘zbek tilining turli shevalarini qamrab oluvchi katta hajmdagi ma’lumotlar to‘plami (dataset) yetarli emas. Shu sababli o‘zbek tilining dialektlarini qamrab oluvchi lingvistik ma’lumotlar bazasini yaratish muhim ilmiy vazifa hisoblanadi.

Nutqdan axborot olish tizimlarining samaradorligi ko‘p jihatdan foydalaniladigan ma’lumotlar bazasining hajmi va sifati bilan bog‘liq. Zamonaviy sun’iy intellekt tizimlari millionlab matn va audio namunalar asosida o‘qitiladi. Katta hajmdagi ma’lumotlar to‘plami (dataset) modelga turli dialekt, aksent va talaffuz farqlarini aniqlash imkonini beradi. Shu sababli global texnologik kompaniyalar nutqni qayta ishlash tizimlarini rivojlantirishda katta hajmdagi lingvistik ma’lumotlar bazalaridan foydalanadi. Bunday ma’lumotlar to‘plami (dataset) nafaqat nutqni aniqlash aniqligini oshiradi, balki matndan kerakli axborotni ajratib olish jarayonini ham samaraliroq qiladi.

Nutqdan axborot olish jarayonida tabiiy tilni qayta ishlash algoritmlari muhim rol o‘ynaydi. NLP texnologiyalari matnni sintaktik va semantik jihatdan tahlil qilish imkonini beradi. Bu esa axborot tizimlariga foydalanuvchi so‘rovlarini to‘g‘ri tushunish va kerakli ma’lumotlarni tezkor topish imkonini yaratadi. Masalan, ovozli yordamchilar va qidiruv tizimlari nutq orqali berilgan so‘rovlarni matnga aylantirib, ma’lumotlar bazasidan kerakli javoblarni topish uchun NLP algoritmlaridan foydalanadi.

Shuningdek, nutqni qayta ishlash tizimlarida real vaqt rejimida ishlash imkoniyati muhim hisoblanadi. Speech-to-text tizimlarining tezligi odatda latency va real-time factor (RTF) ko‘rsatkichlari bilan baholanadi. Latency nutqni qabul qilishdan natija chiqarishgacha bo‘lgan vaqtni bildiradi. Agar tizimning RTF ko‘rsatkichi 1 dan kichik bo‘lsa, u real vaqtga yaqin tezlikda ishlayotganini anglatadi. Zamonaviy cloud texnologiyalari yordamida nutqni qayta ishlash tizimlari juda qisqa vaqt ichida katta hajmdagi ma’lumotlarni qayta ishlash imkoniyatiga ega bo‘lmoqda.

Texnologiya	Vazifasi
Speech recognition	Nutqni matnga aylantirish
Natural Language Processing	Matnni semantik tahlil qilish
Machine learning	Modelni o‘qitish
Cloud computing	Tezkor hisoblash

2-jadval. Nutqdan axborot olish tizimlarida qo‘llaniladigan texnologiyalar

O‘zbek tilining shevalari uchun nutqdan axborot olish tizimlarini yaratish alohida ilmiy muammo hisoblanadi. Chunki turli hududlarda ishlatiladigan dialektlar fonetik va leksik jihatdan sezilarli farqlarga ega. Shu sababli o‘zbek tilining dialektlarini qamrab oluvchi maxsus lingvistik korpus yaratish zarur. Bunday korpus matnli va audio ma’lumotlarni o‘z ichiga olishi kerak. Kelajakda ana shunday ma’lumotlar bazasi asosida o‘zbek tilining turli shevalarini avtomatik tarzda aniqlash va ulardan axborot olish imkoniyatiga ega bo‘lgan zamonaviy axborot tizimlarini yaratish mumkin bo‘ladi.

Tadqiqotda qiyosiy tahlil va tizimli yondashuv metodlaridan foydalanildi. Mavjud speech-to-text tizimlari (Google Translate, Microsoft Translator, Papago) aniqlik darajasi (Accuracy) va kechikish vaqti (Latency) bo‘yicha o‘zaro solishtirildi. Signalni qayta ishlashda Fourier transform va spectrogram extraction algoritmlarining o‘rni tahlil qilindi.

Nutqdan axborot olish jarayoni audio signalni raqamli ko‘rinishga o‘tkazish va undan foydali belgilarni (features) ajratib olishdan boshlanadi. Tadqiqotda nutq signalini tahlil qilish uchun Furye almashinuvi (Fourier Transform) va Mel-kepstral koeffitsiyentlari (MFCC) algoritmlari qo‘llaniladi.

Signalni vaqt sohasidan chastota sohasiga o‘tkazish uchun quyidagi diskret Furye almashinuvi formulasidan foydalaniladi:

$$X(k) = \Sigma\, x(n)e\text{^}\{-j2\pi kn/N\}$$

Bu yerda *x(n)* -vaqt sohasidagi kiruvchi audio signal, *X(k)* esa uning chastota sohasidagi ifodasini bildiradi. Ushbu transformatsiya signalni chastota bo‘yicha tahlil qilish imkonini beradi. Shevalardagi talaffuz farqlarini aniqlashda spektral zichlik

muhim rol o‘ynaydi. Olingan spektrogrammalar neyron tarmoqlari uchun kiruvchi ma’lumot bo‘lib xizmat qiladi.

Shevalarning texnologik tahlili va lingvistik muammolar: O‘zbek tilining shevalari (qarluq, qipchoq, o‘g‘uz) bir-biridan nafaqat leksik, balki akustik va fonetik jihatdan ham farq qiladi. Bu holat Speech-to-Text (STT) tizimlari uchun qator murakkabliklarni yuzaga keltiradi:

Fonetik variatsiyalar: Masalan, "kelyapman" so‘zi turli shevalarda "kevottiman", "keyattirman", "kelayotibman" kabi talaffuz qilinishi mumkin. Bunday holatda tizim so‘zning bazaviy shaklini (lemma) topishi uchun kuchli semantik analizator talab etiladi.

Ovoz interferensiyasi: Shevalardagi o‘ziga xos intonatsiya va urg‘u tizimning akustik modelini chalg‘itishi mumkin.

Tadqiqot doirasida o‘zbek tili shevalari uchun ma’lumotlar to‘plami (dataset) strukturasi ishlab chiqildi. Unga ko‘ra, sifatli tizim yaratish uchun har bir shevadan kamida 500 soatlik toza audio yozuv va ularning matnli transkripsiyasi bo‘lishi lozim.

Neyron tarmoqlari va Transformer arxitekturasi: Zamonaviy nutqni tanish tizimlari Attention (diqqat) mexanizmiga asoslangan Transformer modellaridan foydalanadi. Mazkur modellar nutqdagi uzoq vaqtli bog‘lanishlarni samarali hisobga olish imkonini beradi. Maqolada ko‘rilayotgan tizimning ishlash zanjiri quyidagicha:

Encoder: Audio spektrogrammadan yashirin belgilarni ajratadi.

Attention Mechanism: Nutqning qaysi qismi axborotga boyroq ekanini aniqlaydi.

Decoder: Ajratilgan belgilarni matn ko‘rinishiga o‘tkazadi.

Natijalar. Tahlillar shuni ko‘rsatdiki, global tizimlar o‘zbek adabiy tili uchun yuqori aniqlik (92-95%) ko‘rsatsa-da, mahalliy shevalarda bu ko‘rsatkich 60-65% gacha tushib ketadi (1-jadval)

.

Tizim	Speech aniqligi (%)	Tezlik	Cloud
Google Translate	95	juda tez	bor
Microsoft Translator	92	tez	bor
Papago	88	o‘rtacha	bor
Mahalliy tizimlar	60–65	sekin	cheklangan

1-jadval. Nutqdan axborot olish tizimlarining qiyosiy tahlili.

Nutqni aniqlash tizimlarining samaradorligi bevosita dataset hajmi va neyron tarmoqlarning o‘qitilganlik darajasiga bog‘liq. O‘zbek tili shevalari uchun maxsus lingvistik korpus yaratish ushbu muammoning yagona yechimidir.

Eksperiment natijalari. Mazkur tadqiqot doirasida o‘zbek tilidagi 450 dan ortiq gaplardan iborat maxsus belgilangan (annotatsiya qilingan) dataset shakllantirildi. Ushbu ma’lumotlar uch toifaga — ijobiy, salbiy va neytral sinflarga ajratildi. Matnlarni sonli ko‘rinishga o‘tkazish uchun TF-IDF vektorizatsiya usuli qo‘llanildi, klassifikatsiya masalasini yechishda esa Logistic Regression algoritmidan foydalanildi. Dataset 80/20 nisbatda o‘qitish va test to‘plamlariga bo‘lindi.

O‘tkazilgan eksperiment natijalariga ko‘ra, model umumiy holda 81% aniqlik (accuracy) ko‘rsatkichiga erishdi. Sinflar kesimida tahlil qilinganda, eng yuqori natija ijobiy gaplar uchun qayd etildi (recall = 0.91), bu modelning ijobiy emotsional mazmundagi matnlarni aniqlashda yuqori samaradorlikka ega ekanligini ko‘rsatadi. Salbiy va neytral sinflar bo‘yicha ham qoniqarli natijalar qayd etilib, modelning barqaror ishlashi tasdiqlandi.

Shuningdek, olingan natijalar taklif etilgan yondashuvning o‘zbek tilida sentiment tahlilini amalga oshirishda amaliy jihatdan samarali ekanligini ko‘rsatdi. Kelgusida ushbu yondashuvni chuqur o‘rganish modellari, xususan transformer arxitekturasi asosidagi BERT turkumidagi modellar yordamida yanada takomillashtirish maqsadga muvofiqdir.

Xulosa.Mazkur tadqiqot natijalari shuni ko‘rsatdiki, nutqdan axborot olish va uni qayta ishlash tizimlarining samaradorligi bevosita foydalanilayotgan ma’lumotlar to‘plamining (dataset) hajmi va sifati bilan chambarchas bog‘liq. Global miqyosda ishlab chiqilgan tizimlar yuqori aniqlik ko‘rsatkichlariga (92–95%) ega bo‘lsa-da, o‘zbek tili, ayniqsa uning turli shevalarini qayta ishlashda sezilarli muammolar mavjud.

O‘tkazilgan tahlillar va eksperiment natijalari asosida aniqlanishicha, mahalliy tizimlarda aniqlik darajasi 60–65% atrofida bo‘lib, bu ko‘rsatkich asosan yetarli hajmdagi va sifatli lingvistik ma’lumotlar bazasining mavjud emasligi bilan izohlanadi. Shu bilan birga, mazkur tadqiqot doirasida tashkil etilgan kichik hajmdagi dataset asosida o‘tkazilgan eksperiment natijasida 81% aniqlikka erishilgani, hatto cheklangan ma’lumotlar sharoitida ham samarali modellar yaratish mumkinligini ko‘rsatdi.

Tadqiqot natijalari shuni tasdiqlaydiki, o‘zbek tilining turli dialektlarini qamrab oluvchi keng qamrovli audio va matnli korpus yaratish ushbu sohada olib boriladigan ilmiy va amaliy ishlanmalar uchun muhim ahamiyat kasb etadi. Bunday korpus nafaqat nutqni aniqlash aniqligini oshirishga, balki matndan axborot ajratib olish jarayonining samaradorligini ham sezilarli darajada yaxshilashga xizmat qiladi.

Kelgusida ushbu yo‘nalishda chuqur o‘rganish (deep learning) usullari, xususan transformer arxitekturasi asosidagi zamonaviy modellardan (BERT, wav2vec va boshqalar) foydalanish orqali tizim aniqligini oshirish, shuningdek real vaqt rejimida ishlovchi samarali nutq texnologiyalarini yaratish maqsadga muvofiq hisoblanadi.

Foydalanilgan adabiyotlar:

1. Matlatipov G. Improving Uzbek Machine Translation Through Parallel Corpora. – 2025.
2. OpenAI. Large Language Models and Natural Language Processing. – 2023.
3. Vaswani A. et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.
4. Hinton G., Deng L. Deep Neural Networks for Speech Recognition. IEEE Signal Processing Magazine, 2012.
5. Jurafsky D., Martin J. Speech and Language Processing. Pearson, 2020.
6. Abduvakhidov A. Uzbek Dialectology: Phonetic and Lexical structures. – Tashkent, 2022.
7. Amodei D. et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. – 2016.
8. Baevski A. et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. – 2020.
9. Graves A. Sequence Transduction with Recurrent Neural Networks. – 2012.
10. Gulati A. et al. Conformer: Convolution-augmented Transformer for Speech Recognition. – 2020.
11. Park D. S. et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. – 2019.
12. Radfort A. et al. Robust Speech Recognition via Large-Scale Weak Supervision (Whisper Model). – 2022.
13. Yusupov O. Principles of Computational Linguistics for Turkic Languages. – 2023.
14. Zue V. et al. Speech Database Development: Methods and Tools. – 2021.
15. Karimov S. Sun'iy intellekt tizimlarida o'zbek tili korpusining o'rni. – 2024.