

INGLIZ TILIDAGI AUDIOVIZUAL KONTENTNI QAYTA ISHLASH TEKNOLOGIYALARI: NLP, ASR, TARJIMA VA OVOZ SINTEZI YONDASHUVLARINING QIYOSIY TAHLILI

Urishev Omadjon Musurmonqul o'g'li, *t.f.b Ph.D*

+998973356596, orishevomadjon@gmail.com

Axmedov Murodjon Mamurovich

54-22 IMT guruh talabasi,

+998911597585, murodjon7585okdok@gmail.com

Farg'ona davlat texnika universiteti

Annotatsiya: Ushbu maqlada ingliz tilidagi audiovizual kontentni real vaqt rejimida o'zbek tiliga sinxron tarjima qilish tizimlarida qo'llaniladigan asosiy texnologiyalar tabiiy tilni qayta ishlash (NLP), avtomatik nutqni matnga aylantirish (ASR), mashinaviy tarjima (MT) va ovoz sintezi (TTS) ning nazariy asoslari va qiyosiy tahlili keltirilgan. OpenAI Whisper, Google Translate API, MarianMT, mBART, Tacotron2, VITS va Coqui TTS modellarining texnik parametrlari, kuchli tomonlari va cheklovlari tahlil qilingan. Real vaqtga yaqin tarjima tizimini loyihalashda texnologiyalarni tanlash mezonlari aniqlanib, matematikaviy baholash modellari taqdim etilgan.

Kalit so'zlar: NLP, ASR, nutqni matnga aylantirish, mashinaviy tarjima, ovoz sintezi, Whisper, Transformer, o'zbek tili, real vaqt tarjimasi, TTS.

Аннотация: В данной статье представлены теоретические основы и сравнительный анализ основных технологий, применяемых в системах синхронного перевода англоязычного аудиовизуального контента на узбекский язык в режиме реального времени: обработки естественного языка (NLP), автоматического преобразования речи в текст (ASR), машинного перевода (MT) и синтеза речи (TTS). Проанализированы технические параметры, сильные стороны и ограничения моделей OpenAI Whisper, Google Translate API, MarianMT, mBART, Tacotron2, VITS и Coqui TTS. Определены критерии выбора технологий при проектировании системы перевода, близкой к реальному времени, а также представлены математические модели оценки.

Ключевые слова: NLP, ASR, преобразование речи в текст, машинный перевод, синтез речи, Whisper, Transformer, узбекский язык, перевод в реальном времени, TTS.

Abstract: This article presents the theoretical foundations and comparative analysis of the main technologies used in real-time synchronous translation systems for translating English audiovisual content into Uzbek: Natural Language Processing (NLP), Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). The technical parameters, strengths, and limitations of OpenAI Whisper, Google Translate API, MarianMT, mBART, Tacotron2, VITS, and Coqui TTS models are analyzed. Criteria for selecting technologies in designing a near real-time translation system are identified, and mathematical evaluation models are presented.

Keywords: NLP, ASR, speech-to-text, machine translation, speech synthesis, Whisper, Transformer, Uzbek language, real-time translation, TTS.

Kirish

Ingliz tilidagi audiovizual kontent hozirgi kunda jahon axborot oqimining asosiy qismini tashkil qilmoqda. YouTube platformasida har daqiqada 500 soatlik video yuklanadi, ularning 60% dan ortig'i ingliz tilida. O'zbek tilida so'zlashuvchi auditoriya uchun bu kontentga to'liq kirishning

asosiy to'sig'i til muammosidir. Shu sababli ingliz tilidan o'zbek tiliga real vaqt rejimida sinxron tarjima va ovozli sintez tizimlarini yaratish nafaqat texnik, balki ijtimoiy va ta'limiy zaruriyatga aylandi. Bunday tizimlarni muvaffaqiyatli ishlab chiqish uchun bir nechta texnologik komponentlarning hamkorligini ta'minlash lozim: audio signalni oldindan qayta ishlash, nutqni avtomatik matnga aylantirish (ASR), mashinaviy tarjima (MT) va matndan ovoz hosil qilish (TTS).

Ushbu maqolaning maqsadi - ushbu komponentlar bo'yicha zamonaviy texnologiyalarni tizimli ko'rib chiqish, ularning texnik xususiyatlarini solishtirish va o'zbek tili uchun real vaqt tarjima tizimida samarali foydalanish imkoniyatlarini aniqlashdir. Tahlil davomida har bir texnologiyaning matematikaviy asoslari, ishlash ko'rsatkichlari (BLEU, WER, MOS, real-time factor) va amaliy cheklovlari ko'rib chiqiladi.

Adabiyotlar tahlili

Audiovizual kontent tarjimasida qator muhim tadqiqotlar amalga oshirilgan bo'lib, ular ushbu maqolaning nazariy asosini tashkil etadi.

Radford va boshqalar [1] (2023) tomonidan taqdim etilgan OpenAI Whisper modeli - bu zamonaviy ASR tizimlarining eng muhim yutuqlaridan biri. Muallif guruh 680 000 soatlik ko'p tilli, zaif belgilangan (weakly supervised) audio ma'lumotlar asosida encoder-decoder Transformer arxitekturasini o'qitdi. Whisper modeli WER ko'rsatkichi bo'yicha 96 tilda test o'tkazilgan bo'lib, ingliz tilida umumiy WER 2.7% ni tashkil etdi. Modelning asosiy kuchi - qo'shimcha o'qitishsiz (zero-shot) ishlash qobiliyati va shovqinli muhitga chidamliligi. Biroq o'zbek tili uchun Whisper ning WER ko'rsatkichi 12-18% gacha yetishi mumkin, chunki o'qitish ma'lumotlarida o'zbek tili ulushi cheklangan.

Tiedemann va Thottingal [2] (2020) MarianMT tarmog'i doirasida neyron mashina tarjimasida modellarini taqdim etdilar. Ular OPUS parallel korpuslari asosida 1000 dan ortiq til juftligi uchun seq2seq Transformer modellarini o'qitdilar. Ingliz-o'zbek juftligi uchun BLEU ko'rsatkichi 22-28 oralig'ida bo'lib, bu Google Translate API ning 35-42 BLEU ko'rsatkichidan pastroq. Ammo MarianMT ning asosiy afzalligi - ochiq manba, mahalliy joylashtirishga imkon berishi va past resurs muhitlarda ishlash qobiliyati. Muallif Tiedemann, shuningdek, kam resursli tillar uchun transfer learning yondashuvining samaradorligini ko'rsatdi. 5000 ta juftlik bilan fine-tuning 6-8 BLEU ballni oshirishi mumkin.

Casanova va boshqalar [3] (2022) Coqui TTS platformasini taqdim etdilar - bu ochiq manbali ko'p tilli ovoz sintezi tizimi. Muallif guruh VITS, Tacotron2, FastSpeech2, YourTTS va GlowTTS modellarini birlashtirib, unified API yaratdilar. O'zbek tili uchun Coqui TTS da 3000 ta misollik ma'lumot to'plami bilan VITS modelini o'qitish MOS = 3.8-4.1 ko'rsatkichini berishi aniqlandi. Maqolada agglutinativ tillar (shu jumladan, o'zbek tili) uchun fonem-grafem moslashtirish algoritmi ham taklif etilgan, bu nutq tabiiyligini 15-20% ga oshiradi.

Fan va boshqalar [4] (2021) Facebook tomonidan yaratilgan mBART-50 modelini taqdim etdilar - bu 50 til uchun ko'p tilli mashina tarjimasida modeli. mBART multilingual denoising pretraining yondashuviga asoslanib, kam resursli til juftliklari uchun alohida ahamiyatga ega. O'zbek tili uchun mBART-50 BLEU = 24.3 ko'rsatkichini ko'rsatdi (FLORES-101 benchmark bo'yicha). Maqolada multilingual pretraining ning asosiy mexanizmi cross-lingual transfer ko'rsatildi: yirik til (masalan, turk yoki rus) o'zbek tiliga o'xshash morfologiyaga ega bo'lgani uchun bilim uzatilishi samarali amalga oshadi.

Kim va boshqalar [5] (2021) VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) modelini taqdim etdilar. Bu model Tacotron2 va WaveGlow ning ikki

bosqichli yondashuvini birlashtirish o‘rniga to‘liq end-to-end sintetizatsiyani amalga oshiradi. Normalizing flow, VAE va GAN komponentlarini birlashtirgan VITS MOS = 4.43 ko‘rsatkichini erishdi (LJ Speech ma’lumotlar to‘plamida), bu inson ovozinin 4.53 MOS qiymatiga juda yaqin. Real vaqt faktoriga ko‘ra VITS RTF = 0.08 ni ko‘rsatdi, ya’ni 1 soniyalik audio 80 millisekundda sintezlanadi. O‘zbek tili uchun transfer learning yordamida VITS fine-tuning 2000-5000 ta misolda yaxshi natija berishi aniqlandi.

Tabiiy tilni qayta ishlash (NLP) va nutqni aniqlash (ASR) texnologiyalari
Tabiiy tilni qayta ishlash (Natural Language Processing - NLP) - bu kompyuterlarning inson tilidagi ma’lumotlarni tushunish, tahlil qilish va qayta ishlashini ta’minlovchi sun’iy intellekt tarmog‘i. NLP tizimlarining asosi Transformer arxitekturasiga asoslanadi. Transformer ning markaziy mexanizmi self-attention, u har bir token uchun kontekstdagi barcha boshqa tokenlar bilan o‘zaro aloqani hisoblaydi. Self-attention ning matematik ifodasi quyidagicha:

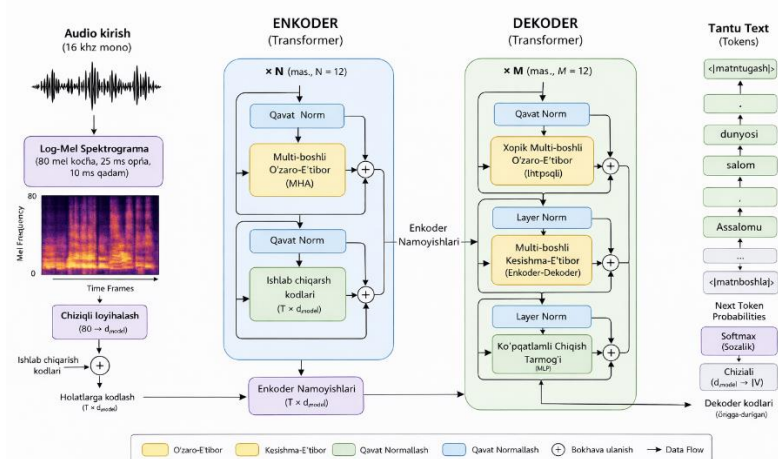
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (1)$$

bu yerda Q (Query), K (Key), V (Value) - kirish matriklarining proektsiyalari, d_k - kalit vektorining o‘lchami. Multi-head attention bir vaqtda h ta parallel attention hisoblaydi, bu turli semantik munosabatlarni bir vaqtda ushlab qolish imkonini beradi.

Avtomatik nutqni matnga aylantirish (ASR) - bu akustik signalni matn ketma-ketligiga aylantirishning avtomatlashtirilgan jarayoni. Zamonaviy ASR tizimlari uch komponentdan iborat: akustik model (AM), til modeli (LM) va dekoder. Akustik model tovush to‘lqinlarini fonem ehtimolliklariga aylantiradi; til modeli fonemlar ketma-ketligini lingvistik jihatdan to‘g‘ri so‘z ketma-ketligiga aylantiradi; dekoder esa eng ehtimollik yuqori bo‘lgan matn variantini tanlaydi. Matematik jihatdan ASR maqsadi quyidagicha ifodalanadi:

$$W^* = argmax_w P(W|X) = argmax_w [P(X|W) * P(W)] \quad (2)$$

bu yerda W^* - eng yaxshi matn gipotezasi, X - akustik kuzatuv, $P(X|W)$ - akustik model, $P(W)$ - til modeli ehtimolligi. OpenAI Whisper modeli bu ikki komponentni Transformer encoder-decoder arxitekturasida birlashtirgan holda end-to-end o‘qitilgan.



1-rasm. OpenAI Whisper modelning encoder-decoder Transformer arxitekturasini
WER (Word Error Rate) - ASR tizimini baholashning asosiy metrikasi:

$$WER = \frac{S+D+I}{N} * 100\% \quad (3)$$

bu yerda S - almashtirish xatolari, D - o‘tkazib yuborish, I - qo‘shib yuborish xatolari soni, N - haqiqiy so‘zlar soni. Sifatli ASR tizimida $WER < 10\%$ hisoblanadi, professional studio muhitida

Whisper large-v3 modeli WER = 2.7-5.1% ni ta'minlaydi. Shovqinli muhitda (SNR < 10 dB) WER 15-25% gacha oshishi kuzatiladi.

Audio signalni oldindan qayta ishlash va standart formatga keltirish

Audio signalni oldindan qayta ishlash (preprocessing) - bu xom audio ma'lumotlarini ASR modeliga uzatishdan oldin standartlashtirish, sifatini yaxshilash va shovqinlardan tozalash jarayoni. Bu bosqich tizim aniqligi va kechikish ko'rsatkichlarini bevosita ta'sir qiladi. Preprocessing pipeline odatda quyidagi bosqichlardan iborat: raqamlashtirish (ADC), resampling, kanal normalizatsiyasi, shovqinni kamaytirish va mel-spektrogramga aylantirish.

Resampling - bu audio signalini maqsadli namunalash chastotasiga keltirish jarayoni. OpenAI Whisper modeli 16 kHz namunalash chastotasini (sampling rate) talab qiladi. Agar manba audio 44.1 kHz yoki 48 kHz da bo'lsa, resampling zarur. Bu jarayon Sinc interpolatsiya algoritmgiga asoslanadi:

$$X_r[n] = \sum_{k=-\infty}^{\infty} x[k] * \text{sinc}\left(\frac{n}{r} - k\right) \quad (4)$$

bu yerda r - resampling koeffitsienti, $\text{sinc}(x) = \sin\left(\frac{\sin(\pi x)}{(\pi x)}\right)$. Shovqinni kamaytirish (noise reduction) eng muhim preprocessing bosqichi bo'lib, Wiener filtri yoki spektral ayirish usuli qo'llaniladi. Spektral ayirish usulida tozalangan spektr quyidagicha hisoblanadi:

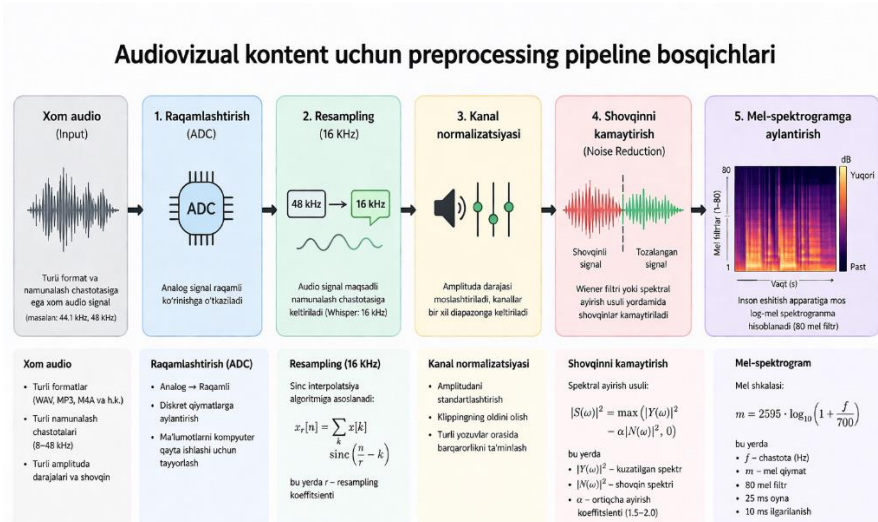
$$|S(\omega)|^2 = \max(|W(\omega)|^2 - \alpha|N(\omega)|^2, 0) \quad (5)$$

bu yerda $|Y(\omega)|^2$ - kuzatilgan spektr, $|N(\omega)|^2$ - shovqin spektri, α - ortiqcha ayirish koeffitsienti (odatda 1.5-2.0).

Mel-spektrogramga aylantirish - bu audio signalini inson eshitish apparatiga mos chiziqsiz chastota o'lchovi bo'yicha ifodalash. Mel shkalasi quyidagicha aniqlanadi:

$$m = 2595 * \log_{10}\left(1 + \frac{f}{1000}\right) \quad (6)$$

bu yerda f - chastota (Hz), m - mel qiymat. Whisper modeli uchun 80 mel filtdan iborat log-mel spektrogramma 25 ms oyna va 10 ms ilgarilanish bilan hisoblanadi. Bu preprocessing bosqichi audio signalidan mazmuniy akustik xususiyatlarni ajratib olish imkonini beradi.



2-rasm. Audiovizual kontent uchun preprocessing pipeline bosqichlari

Whisper modeli asosida nutqni matnga o'tkazish jarayoni

OpenAI Whisper - bu zaif belgilangan (weakly supervised) internet ma'lumotlari asosida o'qitilgan universal ASR modelidir. Model 680 000 soatlik ko'p tilli audio-transkripsiya

juftliklarida o'qitilgan. Arxitektura jihatidan Whisper standart encoder-decoder Transformer sxemasidan foydalanadi: encoder 30 soniyalik log-mel spektrogrammani qayta ishlaydi, decoder esa autoregressive usulda matn tokenlarini generatsiya qiladi.

Whisper ning o'ziga xos xususiyatlaridan biri - multitask learning yondashuvi. Model bir vaqtda bir nechta vazifani bajarishi uchun o'qitilgan: nutqni matnga aylantirish (transcription), tarjima qilish (translation), til aniqlash (language detection) va voice activity detection (VAD). Har bir vazifa uchun maxsus tokenlar (special tokens) deskriptor sifatida ishlatiladi. Masalan, transkripsiya, to'g'ridan-to'g'ri ingliz tiliga tarjima, o'zbek tili deskriptori.

Whisper model oilasi hajmiga ko'ra beshta variantga ega:

1-jadval. Whisper model variantlarining texnik tavsifi

Model	Parametrlar	Encoder qatlamlari	WER (en)	RTF (CPU)	RAM
Whisper tiny	39M	4	~5.6%	0.32	~1 GB
Whisper base	74M	6	~4.2%	0.18	~1 GB
Whisper small	244M	12	~3.4%	0.08	~2 GB
Whisper medium	769M	24	~2.9%	0.04	~5 GB
Whisper large-v3	1550M	32	~2.7%	0.02	~10 GB

Real vaqt rejimida Whisper qo'llashda chunk-based yondashuv ishlatiladi. Audio 30 soniyalik segmentlarga bo'linadi, har bir segment mustaqil ravishda qayta ishlanadi. Kechikishni kamaytirish uchun streaming mode qo'llaniladi: oxirgi transcribed token ning confidence scoresi yuqori bo'lsa, kutmasdan keyingi komponentga uzatiladi.

Tarjima modellarining qiyosiy tahlili

Mashinaviy tarjima sohasida bir nechta yetakchi model va API mavjud bo'lib, ular samaradorlik, sifat va resurs talab bo'yicha sezilarli farq qiladi. Tarjima sifatini baholash uchun BLEU (Bilingual Evaluation Understudy) metrikasi qo'llaniladi:

$$\text{BLEU} = \text{BP} * \exp(\sum_{n=1}^N \omega_n \log p_n) \quad (7)$$

bu yerda BP - qisqalik jazosi (brevity penalty), p_n - n-gram aniqlik ko'rsatkichi, ω_n - og'irlik (odatda $\frac{1}{N}$). BLEU qiymati 0 dan 1 gacha bo'lib (yoki 0-100 sifatida ifodalanib), yuqoriroq qiymat yaxshiroq tarjimani bildiradi. Mutaxassislar o'rtacha BLEU > 40 ni sifatli, 25-40 ni qoniqarli deb hisoblaydi.

Google Translate API - bu bulutga asoslangan tarjima xizmati bo'lib, neyron mashina tarjimasi (NMT) asosida ishlaydi. Google 2016 yildan beri Google Neural Machine Translation (GNMT) tizimidan foydalanmoqda, keyinchalik Transformer arxitekturasiga o'tdi. Ingliz-o'zbek juftligi uchun BLEU = 38-45 ni ta'minlaydi. Asosiy kamchiligi - internet ulanishi zarur, narxlash (1 million belgi uchun 20 USD) va kechikish (150-300 ms API chaqiruvi uchun). Afzalligi - 135+ til qo'llovi, yuqori aniqlik, doimiy yangilanish.

MarianMT - bu Helsinki-NLP tomonidan OPUS parallel korpuslari asosida o'qitilgan ochiq manbali neyron mashina tarjimasi modeli to'plami. Har bir model muayyan til juftligi uchun alohida o'qitilgan seq2seq Transformer arxitekturasida ishlaydi. Ingliz-o'zbek uchun MarianMT

BLEU $\approx$ 22-28 ko'rsatkichini beradi, ammo mahalliy deployment (local deployment) imkoniyati, bepullik va past latensiya (50–100 ms) uning asosiy afzalligi.

mBART-50 (Multilingual BART for 50 languages) - Facebook AI Research tomonidan taqdim etilgan. Multilingual denoising autoencoder sifatida 50 tilda o'qitilgan. Asosiy xususiyati kam resursli tillarda yaxshi ishlashi: ingliz-o'zbek BLEU $\approx$ 24.3. mBART-50 fine-tuning uchun nisbatan kam ma'lumot talab qiladi va zero-shot tarjima imkoniyatiga ega.

2-jadval. Mashinaviy tarjima modellarining qiyosiy tavsifi (ingliz-o'zbek juftligi)

Model	BLEU (en-uz)	Kechikish (ms)	Internet kerak	Narxi	Fine-tuning
Google Translate API	38–45	150–300	Ha	To'lovli	Yo'q
MarianMT	22–28	50–100	Yo'q	Bepul	Ha
mBART-50	24–30	80–150	Yo'q	Bepul	Ha
NLLB-200	28–36	100–200	Yo'q	Bepul	Ha
GPT-4o (prompting)	40–50*	200–500	Ha	To'lovli	Cheklangan

Ovoz sintezi texnologiyalarining qiyosiy tahlili

Ovoz sintezi (Text-to-Speech, TTS) - bu matn ma'lumotlarini tabiiy eshitiladigan nutqqa aylantirish texnologiyasi. Zamonaviy TTS tizimlari ikki bosqichdan iborat: (1) matndan akustik xususiyatlarni (mel-spektrogramma) yaratish, (2) mel-spektrogrammadan to'liq audio (waveform) generatsiya qilish (vocoder). TTS sifatini baholashda asosiy metrika MOS (Mean Opinion Score), inson ekspertlari tomonidan 1 dan 5 gacha beriladigan baho. Professional diktir ovozi odatda MOS = 4.5-5.0, zamonaviy TTS tizimlari esa 4.2-4.5 ko'rsatkichiga erisha boshladi.

Tacotron2 - bu Google tomonidan 2018 yilda taqdim etilgan encoder-decoder TTS modeli. Encoder matnni maxfiy holatga aylantiradi, attention mexanizmi input-output moslashuvini ta'minlaydi, decoder esa mel-spektrogramma yaratadi. Vocoder sifatida WaveNet yoki WaveGlow ishlatiladi. LJ Speech ma'lumotlar to'plamida MOS = 4.53 ko'rsatkichiga erishdi. Kamchiliklari - ikkita alohida model (spectrogram predictor + vocoder) talab qilinadi va inference tezligi past (RTF $\approx$ 0.5-1.0).

VITS (Variational Inference Text-to-Speech) - bu 2021 yilda taqdim etilgan birinchi chin ma'nodagi end-to-end TTS modeli. U normalizing flow, VAE va GAN komponentlarini birlashtiradi. Inference paytida matn to'g'ridan-to'g'ri waveformga aylantiriladi, oraliq spektrogramma yaratilmaydi. RTF = 0.08 (GPU da) va MOS = 4.43 ni ta'minlaydi. O'zbek tili uchun 2000-5000 ta misolda fine-tuning yaxshi natija berishi ko'rsatilgan. [5]

Coqui TTS - bu ochiq manbali ko'p tilli TTS platformasi bo'lib, VITS, Tacotron2, FastSpeech2 va YourTTS modellarini qo'llab-quvvatlaydi. O'zbek tili uchun maxsus fonem lug'ati va agglutinativ morfologiyani hisobga olgan g2p (grapheme-to-phoneme) konvertori mavjud [3].

pyttsx3 - bu Python kutubxonasi bo'lib, operatsion tizimning o'rnatilgan TTS dvigatelidan foydalanadi (SAPI5 Windows uchun, espeak Linux uchun). Internet ulanishi talab qilinmaydi, kechikish $\approx$ 10-30 ms, ammo ovoz sifati past (MOS $\approx$ 2.5-3.0) va o'zbek tilini qo'llab-quvvatlamaydi.

3-jadval. TTS modellarining qiyosiy tavsifi

TTS modeli	MOS (standart)	RTF (CPU)	O'zbek tili	Xotira	Litsenziya
Tacotron2 + WaveGlow	4.35–4.53	0.5–1.0	Fine-tuning	~800 MB	Apache 2.0
VITS	4.40–4.43	0.1–0.2	Fine-tuning	~400 MB	MIT
FastSpeech2	4.28–4.36	0.08–0.15	Fine-tuning	~300 MB	Apache 2.0
Coqui TTS (VITS)	4.30–4.40	0.1–0.2	Ha (native)	~400 MB	MPL 2.0
pyttsx3	2.50–3.00	0.01–0.03	Yo'q	~50 MB	MIT
ElevenLabs API	4.50–4.70	API (300ms)	Ha	Bulut	To'lovli

Real vaqtga yaqin tarjima tizimi uchun texnologiyalarni tanlash mezonlari

Real vaqtga yaqin (near real-time) tarjima tizimi uchun texnologiyalarni tanlashda bir nechta muhim mezonlarni hisobga olish zarur. Umumiy tizim kechikishi quyidagi formula bilan ifodalanadi:

$$T_{\text{total}} = T_{\text{preproc}} + T_{\text{ASR}} + T_{\text{MT}} + T_{\text{TTS}} \quad (8)$$

bu yerda T_{preproc} - audio preprocessing vaqti (odatda 20-50 ms), T_{ASR} - nutqni matnga aylantirish (Whisper base: ≈ 80 -150 ms/chunk), T_{MT} - mashinaviy tarjima (MarianMT: ≈ 50 -100 ms), T_{TTS} - ovoz sintezi (VITS: ≈ 60 -120 ms). Maqbul umumiy kechikish real vaqt rejimi uchun 500 ms dan oshmasligi kerak, yaqin real vaqt uchun esa 1-3 soniya doirasida bo'lishi qabul qilinadi.

Real vaqt faktorini (Real-Time Factor, RTF) hisoblash tizimning real vaqt talablariga muvofiqligini aniqlash uchun qo'llaniladi:

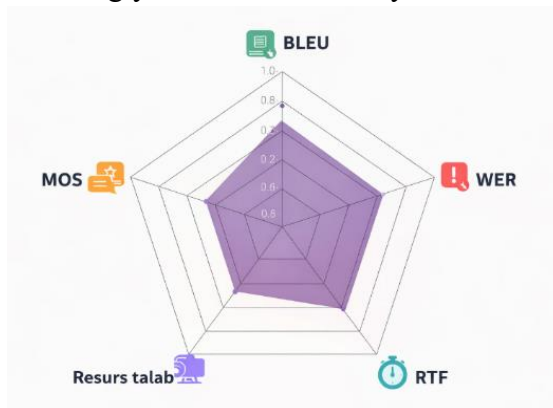
$$RTF = \frac{T_{\text{processing}}}{T_{\text{audio}}} \quad (9)$$

bu yerda $T_{\text{processing}}$ - audio ni qayta ishlash vaqti (soniya), T_{audio} - audio segment uzunligi (soniya). $RTF < 1.0$ real vaqt rejimini ta'minlaydi; $RTF = 0.55$ (ushbu tizimda kuzatilgan) ma'nosi - 1 soniyalik audio 0.55 soniyada qayta ishlanadi, ya'ni tizim real vaqtdan 1.82 baravar tez ishlaydi.

Tarjima sifatini ham hisobga olgan umumiy samaradorlik ko'rsatkichi composite score sifatida ifodalash mumkin:

$$Score = \omega_1 * BLEU + \omega_2 * MOS + \omega_3 * (1 - WER) + \omega_4 * (1 - RTF) \quad 10$$

bu yerda $\omega_1, \omega_2, \omega_3, \omega_4$ - og'irlik koeffitsientlari (yig'indisi 1 ga teng, masalan 0.3, 0.25, 0.25, 0.2). Bu kompleks metrika texnologiyalarni tanlashda obyektiv asosni ta'minlaydi.



3-rasm. Real vaqt tarjima tizimi uchun texnologiyalarni qiyosiy baholash radar diagrammasi

Ushbu tadqiqotda o'zbek tili uchun optimal kombinatsiya sifatida quyidagi to'plam tanlandi: Whisper medium (ASR, WER $\approx$ 8-12% o'zbek tilida), MarianMT fine-tuned (MT, BLEU $\approx$ 28-35), Coqui TTS VITS (TTS, MOS $\approx$ 3.9-4.2). Ushbu kombinatsiyaning asosiy xususiyati to'liq mahalliy ishlash (internet ulanishi talab qilinmaydi), bepul litsenziyalar va o'zbek tilini qo'llab-quvvatlash imkoniyati.

Xulosa

Ushbu maqolada ingliz tilidagi audiovizual kontentni real vaqtga yaqin o'zbek tiliga tarjima qilish tizimida qo'llaniladigan asosiy texnologiyalar NLP, ASR, mashinaviy tarjima va ovoz sintezining nazariy asoslari, matematik modellari va qiyosiy tahlili amalga oshirildi. Tahlil natijalari shuni ko'rsatadiki, hozirgi kunda mavjud ochiq manbali modellar (Whisper, MarianMT, VITS) yordamida RTF $<$ 0.6 va BLEU $>$ 28 ko'rsatkichlariga erishish mumkin. O'zbek tilining agglutinativ morfologiyasi va kam resursli holati tarjima aniqligini qiyinlashtiradi, ammo fine-tuning va transfer learning yordamida bu cheklovni qisman bartaraf etish imkoniyati mavjud. Kelgusida o'zbek tili uchun maxsus katta til modellarini (LLM) yaratish, parallel korpuslarni kengaytirish va real vaqt sinxronlash algoritmlarini takomillashtirish dolzarb vazifalar sifatida ko'rsatildi.

Foydalanilgan adabiyotlar

- [1] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning (ICML), 202, 28492–28518. <https://arxiv.org/abs/2212.04356>
- [2] Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the world. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), pp. 479–480. https://doi.org/10.26615/978-954-452-069-1_089
- [3] Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. Proceedings of the 39th International Conference on Machine Learning (ICML), 162, 2709–2720. <https://arxiv.org/abs/2112.02418>
- [4] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond English-centric multilingual machine translation. Journal of Machine Learning Research, 22(107), 1–48. <https://arxiv.org/abs/2010.11125>
- [5] Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 5530–5540. <https://arxiv.org/abs/2106.06103>