NUTQNI TANIB OLISHDA CHUQUR O‘QITISH MODELLARI TAHLILI

Primqulov Oybek Dilmurot o‘g‘li[1], Sherqo‘ziyeva Shahnozabonu Baxromjon qizi[1]

[1]Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Email: primqulovoybek1@gmail.com, shahnozabonusherqoziyeva@gmail.com

K E Y W O R D S

avtomatik nutqni tanib olish, Self-Supervised Learning, Wav2Vec2.0, tashqi qurilmalar, Transformer, kvantlash, Deep Speech 2, Whisper, nutq texnologiyalari, Whisper

A B S T R A C T

Ushbu maqolada avtomatik nutqni tanib olish sohasida qo‘llanilayotgan zamonaviy chuqur o‘qitish modellari tahlil qilingan. Tadqiqotda Self-Supervised Learning yondashuviga asoslangan Wav2Vec 2.0 modelining ishlash tamoyillari, arxitekturasi va amaliy imkoniyatlari ko‘rib chiqilgan. Modelning kam hajmdagi annotatsiyalangan ma’lumotlar bilan yuqori aniqlikka erishish qobiliyati LibriSpeech natijalari asosida baholangan. Shuningdek, Wav2Vec 2.0 modelining Raspberry Pi kabi cheklangan resurslarga ega qurilmalardagi samaradorligi, kvantlashtirish usullarining hisoblash xarajatlariga ta’siri ham tahlil qilingan. Tadqiqot davomida Deep Speech 2 va Whisper modellari bilan qiyosiy tahlil o‘tkazilib, ularning afzallik va cheklovlari yoritilgan. Olingan natijalar Wav2Vec 2.0 modelining kam hajmdagi nutq ma’lumotlari bilan ham samarali ishlashini, hisoblash resurslariga bo‘lgan talabni optimallashtirish imkoniyatiga ega ekanligini hamda turli amaliy sohalarda qo‘llash uchun istiqbolli yechim sifatida qaralishini ko‘rsatadi.

KIRISH

Sun’iy neyron tarmoqlarning rivojlanishi, katta hajmdagi ma’lumotlar to‘plamlarining paydo bo‘lishi va hisoblash quvvatining ortishi hisobiga so‘nggi o‘n yil ichida zamonaviy avtomatik nutqni tanib olish (ASR) tizimlari tabiiy tilga qayta ishlov berishda yuqori aniqlikka erishdi [1]. Biroq an’anaviy yondashuvlar minglab soat vaqt ketadigan yorliqli (annotatsiyalangan) nutq ma’lumotlarini talab qiladi. Bu esa dunyo tillarining aksariyati uchun jiddiy muammo hisoblanadi, chunki bugungi kunda 7000 dan ortiq til mavjud bo‘lib, ularning ko‘pchiligida yetarli hajmdagi nutq korpuslari shakllanmagan [2]. Shu sababli so‘nggi yillarda ASR tizimlarida katta hajmdagi yorliqli ma’lumotlarga bo‘lgan ehtiyojni kamaytirishga qaratilgan Self-Supervised Learning yondashuvi muhim yo‘nalishlardan biriga aylandi. Ushbu usul modelga avval yorliqsiz xom audio ma’lumotlardan nutqning muhim akustik xususiyatlarini mustaqil o‘rganish, keyinchalik esa kam hajmdagi yorliqli ma’lumotlar yordamida aniq vazifaga moslashish imkonini beradi. Wav2Vec 2.0 modeli ushbu yondashuvning eng muvaffaqiyatli namunalaridan biri bo‘lib, u Alexei Baevski va hammualliflar tomonidan taklif etilgan [3]. Model atigi 10 daqiqalik yorliqli ma’lumot bilan ham avvalgi yarim nazoratli yondashuvlarga nisbatan yaxshiroq natijalarni namoyon qildi. Shu bilan birga, so‘nggi yillarda ASR tizimlarini bulut serverlardan bevosita tashqi qurilmalarga - smartfonlar, aqlli soatlar va Raspberry Pi kabi platformalarga moslashtirishga bo‘lgan qiziqish ortib bormoqda. Biroq zamonaviy ASR modellari cheklangan resursli qurilmalarda qo‘llash uchun katta hisoblash quvvatini talab qiladi. Shu sababli ushbu tadqiqotda Wav2Vec 2.0 modelining ishlash tamoyillari, tashqi qurilmalardagi samaradorligi hamda boshqa zamonaviy ASR yondashuvlari bilan taqqosiy jihatlari tahlil qilindi.

- Wav2Vec 2.0 arxitekturasi ishlash uslubi;
- Modelning tashqi qurilmalarda samaradorlik ko‘rsatkichlari;

(cc) BY

- Whisper va Deep Speech 2 kabi yondashuvlarga nibatan ustunligi;

TADQIQOT METODOLOGIYASI

Tadqiqot davomida wav2vec 2.0 modelining samaradorligi, moslashuvchanligi va amaliy qo‘llanish istiqbollari boshqa zamonaviy ASR yondashuvlari bilan taqqosiy ravishda o‘rganildi. Tadqiqotning nazariy asosini Wav2Vec 2.0, Whisper va Deep Speech 2 modellariga bag‘ishlangan zamonaviy ilmiy maqolalar tashkil etdi. Self-Supervised Learning asosidagi Wav2Vec 2.0 arxitekturasi, modelning tashqi qurilmalardagi ishlash samaradorligi, sud-tibbiyot yo‘nalishidagi qo‘llanilishi hamda chuqur o‘rganishga asoslangan an’anaviy ASR tizimlari bilan bog‘liq ilmiy tadqiqotlar tahlil qilindi. Tahlil jarayonida modellar samaradorligi bir nechta asosiy ko‘rsatkichlar asosida baholandi. Nutqni tanib olish aniqligi WER (Word Error Rate) orqali, qayta ishlash tezligi RTF (Real-Time Factor) yordamida tahlil qilindi. Shuningdek, energiya sarfi, operativ xotira talabi, annotatsiyalangan ma’lumotlarga bog‘liqlik darajasi hamda shovqinli muhitdagi barqaror ishlash qobiliyati ham taqqosiy mezon sifatida ko‘rib chiqildi.

Wav2Vec 2.0 arxitekturasi va Self-Supervised Learning

Wav2Vec 2.0 xom audio signalni bevosita kirish sifatida qabul qilib, ko‘p qatlamli konvolyutsion neyron tarmoq (CNN) yordamida undan nutqning muhim akustik xususiyatlarini ajratib oladi va ularni model qayta ishlashi uchun qulay bo‘lgan ichki ifodalarga aylantiradi. Keyingi bosqichda ajratib olingan akustik xususiyatlar Transformer tarmog‘iga uzatiladi va u yerda nutqning umumiy konteksti hamda vaqt bo‘yicha bog‘liqliklarini hisobga olgan holda chuqurroq ifodalarga aylantiriladi. O‘qitish jarayonida ushbu ma’lumotlarning ayrim qismlari vaqtincha yashiriladi (masking), model esa yashirilgan qismlarni atrofidagi kontekst asosida tiklashni o‘rganadi. Bu yondashuv tabiiy tilni qayta ishlashdagi BERT modelining ishlash tamoyiliga yaqin hisoblanadi. Arxitekturada ikkita asosiy konfiguratsiyadan iborat: BASE va LARGE. BASE nisbatan ixchamroq bo‘lib, kamroq hisoblash resurslarini talab qiladi, LARGE esa ko‘proq parametrlar va chuqurroq qatlamlar hisobiga yuqoriroq aniqlikni ta’minlaydi. Shu sababli BASE modeli resurslari cheklangan qurilmalar uchun mos bo‘lsa, LARGE modeli yuqori unumdorlik talab qilinadigan vazifalarda samaraliroq hisoblanadi [3]. Wav2Vec 2.0 arxitekturasidagi feature encoder xom audio signalni bir necha konvolyutsion qatlamlar orqali qayta ishlab, undan nutqning asosiy akustik xususiyatlarini ajratib oladi. Ushbu xususiyatlar modelning keyingi bosqichlarida yanada chuqur tahlil qilinadi. Modelning eng muhim ustunligi kam hajmdagi annotatsiyalangan ma’lumot bilan ham yuqori aniqlikni ta’minlay olishi hisoblanadi. Jadval 1 da turli hajmdagi ma’lumotlar asosida olingan WER natijalari keltirilgan.

LibriSpeech da turli miqdordagi yorliqli ma’lumot bilan WER ko‘rsatkichlari

Model	Yorliqli ma’lumot	Test-clean WER (%)	Test-other WER (%)
BASE (LS-960)	10 daqiqa	9.1	15.6
LARGE (LV-60k)	10 daqiqa	4.8	8.2
BASE (LS-960)	1 soat	5.5	11.3
LARGE (LV-60k)	1 soat	2.9	5.8
BASE (LS-960)	100 soat	3.4	8.0
LARGE (LV-60k)	960 soat	1.8	3.3

Jadval natijalari wav2vec 2.0 modelining kam hajmdagi annotatsiyalangan ma’lumot bilan ham yuqori aniqlikka erisha olishini ko‘rsatadi. Xususan, LARGE konfiguratsiyasi atigi 10 daqiqalik ma’lumot asosida ham past WER ko‘rsatkichlarini namoyon etib, modelning ma’lumotdan samarali foydalanish qobiliyatini tasdiqlaydi. Annotatsiyalangan ma’lumot hajmi

(cc) BY

oshgani sari tanib olish aniqligi yanada yaxshilanib, 960 soatlik ma'lumotda model juda past xatolik darajasiga erishadi. Bu esa SelfSupervised Learning yondashuvi yordamida oldindan o'rganilgan ichki ifodalar keyingi moslashtirish bosqichida yuqori samaradorlik berishini ko'rsatadi hamda modelning to'liq nazoratli yondashuvlar bilan bemalol raqobatlasha olishini tasdiqlaydi[3].

Tashqi qurilmalarda ishlash ko'rsatkichlari

Wav2vec 2.0 modelining tashqi qurilmalardagi samaradorligini baholash natijalari uni optimallashtirish orqali resurslari cheklangan platformalarda ham samarali ishlatish mumkinligini ko'rsatadi. Raspberry Pi qurilmasida o'tkazilgan tajribalarda modelning kvantlashtirilgan, ya'ni xotira va hisoblash talabini kamaytirish uchun optimallashtirilgan, hamda kvantlashtirilmagan, ya'ni asl to'liq aniqlikdagi variantlari o'zaro taqqoslandi. Tahlillar til modeli bilan dekodirlash qo'llanilganda nutqni tanib olish aniqligi oshishini hamda WER darajasi sezilarli kamayishini ko'rsatdi. Biroq aniqlikning ortishi xotira sarfi hisobiga yuz beradi, chunki katta hajmli til modeli qo'llanilganda operativ xotira talabi keskin oshadi.

RTF (Real-Time Factor) turli yadro konfiguratsiyalari uchun

Model	1 yadro	2 yadro	3 yadro	4 yadro
Wav2Vec (kvantlashtirilgan)	2.30	1.35	1.07	0.98
Wav2Vec (kvantlashtirilmagan)	2.48	1.66	1.45	1.36

Jadval natijalari wav2vec 2.0 modelini optimallashtirish uning hisoblash samaradorligini oshirishini ko'rsatadi. Tajribalar davomida yadrolar soni ortgani sari modelning qayta ishlash tezligi ham yaxshilangani kuzatildi. Xususan, 4 yadroli konfiguratsiyada RTF ≈ 0.98 natijaga erishilishi modelning deyarli real vaqt rejimida ishlash imkoniyatiga ega ekanini bildiradi. Shuningdek, kvantlashtirish usuli energiya sarfi va xotira talabini kamaytirib, modelni cheklangan resurslarga ega qurilmalarda yanada samarali ishlatish imkonini beradi. Olingan natijalar Wav2Vec 2.0 modelini mobil qurilmalar, IoT platformalari hamda boshqa tashqi tizimlarga moslashtirish mumkinligini ko'rsatadi [1].

Deep Speech 2 va an'anaviy yondashuvlar

Deep Speech 2 modeli avtomatik nutqni tanib olishda chuqur o'rganishga asoslangan ilk muvaffaqiyatli yondashuvlardan biri bo'lib, to'liq nazoratli o'rganish paradigmasiga tayanadi. Model audio signalni log-spektrogram ko'rinishida qabul qilib, konvolyutsion va takroriy neyron tarmoqlar yordamida nutqdagi vaqt bo'yicha bog'liqliklarni o'rganadi hamda CTC mexanizmi orqali audio va matn o'rtasidagi moslikni aniqlaydi. Ushbu yondashuv o'z davri uchun yuqori samaradorlikni namoyon etib, LibriSpeech sinovlarida inson darajasiga yaqin, ayrim holatlarda esa undan ham yaxshiroq aniqlik ko'rsatkichlariga erishgan. Deep Speech 2 ning samaradorligi katta hajmdagi annotatsiyalangan ma'lumotlar mavjudligiga kuchli darajada bog'liq bo'lib, bu uni resurslari cheklangan tillar uchun murakkab yechimga aylantiradi. Bundan tashqari, zamonaviy self-supervised modellar bilan taqqoslaganda uning shovqinli akustik muhitlarga moslashuvchanligi nisbatan pastroq. Shu bilan birga, past kechikishli ishlash va amaliy joriy etish imkoniyatlari sababli Deep Speech 2 ASR tizimlarining keyingi rivojlanishiga muhim asos bo'lib xizmat qilgan model sifatida baholanadi [4].

(cc) BY

ASR modellarining qiyosiy tahlili

Mezon	Deep Speech 2 [4]	Wav2Vec 2.0 [3]	Whisper [2]
O‘qitish paradigmasi	To‘liq nazoratli	O‘z-o‘zini nazorat ostida + kalibratsiya	To‘liq nazoratli
Ma’lumotlar to‘plami	~12,000 soat	10 daqiqa — 960 soat	680,000 soat
Arxitektura	RNN + CNN + CTC	CNN + Transformer + CTC	Encoder-Decoder Transformer
LibriSpeech WER (clean)	5.15%	1.8%	~5.4%
Til ko‘lamlilik	Faqat maqsad til	128 til (XLS-R)	Ko‘p til, nol-tanqis
Tashqi qurilma muvofiqligi	o‘rta	Moslashuvchan	Cheklangan (katta model)
Shovqinga chidamliligi	o‘rta (2D conv yordamida)	o‘rta	Yuqori
Maxfiylik imkoniyati	Cheklangan	Federal o‘rganishga mos	Cheklangan

Ushbu taqqosiy tahlil har bir model muayyan senariy uchun afzalliklarga ega ekanini ko‘rsatadi. Wav2Vec 2.0 resurslari cheklangan (past resursli) tillar uchun va ma’lumot maxfiyligini talab qiladigan muhitlar uchun alohida qimmatli. Deep Speech 2 katta o‘qitish to‘plamlari mavjud bo‘lgan resurslarga boy (high-resource) tillarda yuqori aniqlikni ta’minlaydi. Whisper esa shovqinli muhitlarda va nol-tanqis (zero-shot) holatlarda ustunlik qiladi.

XULOSA

Ushbu tadqiqot davomida zamonaviy avtomatik nutqni tanib olish texnologiyalari ilmiy manbalar asosida tahlil qilinib, Wav2Vec 2.0 modelining arxitekturasi, ishlash tamoyillari va amaliy imkoniyatlari o‘rganildi. O‘tkazilgan tahlillar Self-Supervised Learning yondashuvi asosida qurilgan ushbu model katta hajmdagi annotatsiyalangan ma’lumotlarga bo‘lgan ehtiyojni sezilarli darajada kamaytirishini hamda kam miqdordagi yorliqli ma’lumotlar bilan ham yuqori aniqlikdagi natijalarga erisha olishini ko‘rsatdi. Bu esa Wav2Vec 2.0 modelini resurslari cheklangan tillar uchun nutqni tanib olish tizimlarini yaratishda istiqbolli yechim sifatida baholash imkonini beradi. Shuningdek, tadqiqotda modelning chetki qurilmalarda qo‘llanish imkoniyatlari ham ko‘rib chiqildi. Kvantlashtirish va optimallashtirish usullaridan foydalanish orqali Wav2Vec 2.0 modelini Raspberry Pi kabi cheklangan hisoblash resurslariga ega platformalarda deyarli real vaqt rejimida ishlatish mumkinligi aniqlandi. Bu esa energiya sarfi va xotira talablarini kamaytirish bilan birga, modelni amaliy tizimlarga joriy etish imkoniyatlarini kengaytiradi. Deep Speech 2 va Whisper modellari bilan o‘tkazilgan qiyosiy tahlillar natijasida har bir yondashuvning o‘ziga xos afzalliklari mavjudligi kuzatildi. Wav2Vec 2.0 kam resursli tillar va ma’lumotlar hajmi cheklangan sharoitlarda samarali natijalar ko‘rsatsa, Whisper modeli shovqinli akustik muhitlarda yuqoriroq aniqlikni ta’minlaydi. Tadqiqot natijalari Wav2Vec 2.0 modelining yuqori aniqlik, moslashuvchanlik va turli qurilmalarga integratsiya qilish imkoniyatlari bilan zamonaviy ASR tizimlari orasida muhim o‘rin tutishini ko‘rsatdi. Kelgusida modelni o‘zbek tili uchun moslashtirish, hisoblash xarajatlarini yanada kamaytirish hamda amaliy dasturiy yechimlar ishlab chiqish ushbu yo‘nalishdagi dolzarb tadqiqot vazifalaridan biri bo‘lib qoladi.

Foydalanilgan adabiyotlar

1. Santosh Gondi, Wav2Vec2.0 on the Edge: Performance Evaluation. Meta Inc. arXiv:2202.05993v1 [cs.SD] 12 Feb 2022. Sound (cs.SD); Human-Computer Interaction(cs.HC);

(cc) BY

Audio and Speech Processing (eess.AS) https://doi.org/10.48550/arXiv.2202.05993
2. Juan Camilo Vásquez-Correa, Aitor Álvarez Muniain. Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper. Journal of Sensors, Vol.23, Issue 4, 2023. https://doi.org/10.3390/s23041843
3. Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. https://arxiv.org/abs/2006.11477v3
4. Dario Amodei at all. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. https://arxiv.org/abs/1512.02595v1 8 Dec 2015.
5. Vassil Panayotov, Guoguo Chen, Daniel Povey, S. Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. Journal of IEEE International Conference on Acoustics, Speech, and Signal Processing.
6. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. https://arxiv.org/abs/2212.04356v1 6 Dec 2022

(cc) BY