KAM RESURSLI TILLAR UCHUN MATNNI NUTQQA SAMARALI O‘TKAZISHDA O‘QITUVCHISIZ O‘RGATISH USULI.

Tojidinova Dilafruz Komiljonovna

Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti, Toshkent, O‘zbekiston,

e-mail: jumaevadilafruz42@gmail.com

Abstrakt *Ketma-ketlikdan ketma-ketlikka modellar Matndan nutqqa (TTS) sohasida muvaffaqiyatli qo'llanilmoqda. Ushbu modellar katta va aniq transkripsiya qilingan nutq korpusi yordamida inson nutqiga yaqin nutq yaratishi mumkin. Biroq, bunday katta ma'lumotlar to'plamini tayyorlash qimmat va juda ko’p mehnat talab qiladi. Ushbu muammoni yengillashtirish uchun biz ushbu maqolada yangi o'qituvchisiz o'rgatish mexanizmini taklif qilamiz. Aniqrog‘i, avval Vector-quantization Variational-Autoencoder (VQ-VAE) yordamida katta miqyosda, omma e'tiboriga havola qilingan va transkripsiya qilinmagan nutqdan o'qituvchisiz lingvistik birliklarni chiqaramiz. Keyin biz ketma-ketlikdan ketma-ketlikka TTS modelini <o'qituvchisiz lingvistik birliklar, audio> juftlari yordamida pre-training qilamiz. Oxir-oqibat, modelni maqsadli nutq so‘zlovchisidan kichik miqdordagi <matn, audio> juft ma'lumotlar bilan fine-tuning qilamiz. Natijada, obyektiv va subyektiv baholashlar shuni ko'rsatadiki, bizning taklif qilgan usulimiz bir xil miqdordagi juft ma'lumotlar bilan yanada aniq va tabiiy nutqni sintez qila oladi. Bundan tashqari, biz taklif qilingan usulni taxminiy past resursli tillarga kengaytiramiz va usulning samaradorligini obyektiv baholash yordamida tasdiqlaymiz.*

1. Kirish

Ketma-ketlikdan ketma-ketlikka matndan nutqqa (S2S TTS) modellar, kodlovchi-dekodlovchi-e'tibor mexanizmi yordamida tabiiy nutq yaratishi mumkin [1–5]. Biroq, ushbu S2S TTS modellarni o'qitish uchun o'nlab soat transkripsiya qilingan nutq kerak bo'ladi. Garchi kamroq ma'lumot talab qilinsa ham, bu umumiy tabiiylikni cheklaydi va

modelning noto'g'ri xatolarga moyil bo'lishiga olib keladi. Shunday katta transkripsiya qilingan nutq korpusini to'plash qimmat va ko'p mehnat talab qilgani uchun tadqiqotchilar TTS sohasida ma'lumot samaradorligi muammosini o'rganishni boshladilar. Ba'zi tadqiqotlar yangi so'zlovchilarga moslashish uchun TTS modelini kichik miqdordagi ma'lumotlar yordamida moslashtirishga qaratilgan. Ba'zilar pre-training qilingan modelning barchasini yoki bir qismini maqsadli so'zlovchilardan olingan kichik miqdordagi ma'lumotlar yordamida fine-tuning qilishni taklif qildilar [6, 7]. Ba'zilar TTS sohasida so'zlovchi identifikatsiyasini modeling qilishni o'rganishdi [8, 9]. Ba'zilar hatto nol-ma'lumotli so'zlovchi moslashuvini o'rganishdi [9, 12]. Boshqa tadqiqotlar TTS modelini universal ma'lumotlardan foydalanish bilan yaratishni o'rgandi. Ba'zilar an'anaviy TTS paradigmida TTS ga distribyutsion matnli yoki lingvistik ma'lumotlarni kiritishni o'rganishdi [13–15]. Ba'zilar TTS modellarini avtomatik nutqni tan olish (ASR) ma'lumotlari yoki ma'lumotlar tanlovi yoki tahlili orqali topilgan ma'lumotlardan foydalanib o'qitishni o'rgandilar [16–19]. Yaqinda [20] faqat nutqdan foydalangan holda oxir-oqibat TTS modellarining dekoderini pre-training qilish uchun oddiy, ammo samarali yarim-nazoratli yondashuvni taklif qildi. Past resursli tillar uchun TTS sohasida ma'lumot samaradorligi bo'yicha ba'zi ishlanmalar mavjud. Ko'p tilli statistik parametrik nutq sintezi (SPNS) modelini o'qitish past miqdordagi ma'lumotlar bilan yangi tillarga moslashishni yengillashtirishi ko'rsatilgan [21, 22]. Yaqinda o'tkazilgan tadqiqot [23] yuqori resursli tillardan past resursli tillarga transfer learning ni o'rgandi. Ushbu ish S2S TTS ni o'qitish uchun ma'lumot talabini kamaytirishga qaratilgan bo'lib, katta miqyosda, omma e'tiboriga havola qilingan va transkripsiya qilinmagan nutq ma'lumotlarini o'z ichiga oladi. Biz Tacotron [2], zamonaviy S2S TTS modelini o'qitish uchun o'qituvchisiz frameworkni taklif qilamiz. Aniqrog'i, avval Vector-quantization Variational-Autoencoder (VQ-VAE) yordamida transkripsiya qilinmagan nutqdan o'qituvchisiz lingvistik birliklarni chiqaramiz. Keyin Tacotron ni <unsupervised lingvistik birliklar, audio> juftlari yordamida pre-training qilamiz. Oxir-oqibat, modelni maqsadli so'zlovchilardan olingan kichik miqdordagi <matn, audio> juft ma'lumotlar

bilan fine-tuning qilamiz. E'tibor berish kerakki, bizning ishimiz [20] bilan bog'liq. Biroq, bizning ishimiz bir nechta jihatlar bilan farq qiladi, bu bizning ishimizning asosiy hissasini tashkil etadi. Birinchi va eng muhim farq shundaki, bizning yondashuvimiz o'qituvchisiz o'rganishni lingvistik birliklarni chiqarish uchun foydalanadi, bu esa butun TTS modelini pre-training qilish mumkin bo'ldi, [20] esa modelning har bir qismini alohida pre-training qiladi. Ikkinchidan, bizning yondashuvimizni past resursli tillarda ham tasdiqlaymiz. Va nihoyat, bizning tajribalarimizda asosan ommaviy kirish imkoniyatiga ega ma'lumotlardan foydalanamiz, bu esa oson takrorlanishi mumkin.

2-bo'limda biz [20] dagi yarim-nazoratli pre-training ni ko'rib chiqamiz va bizning taklif qilgan o'qituvchisiz usulimizni tasvirlaymiz. 3-bo'limda eksperiment sozlamalari va natijalarini batafsil bayon qilamiz. Maqola 4-bo'limda xulosa bilan yakunlanadi.

2. Taklif qilingan usul

Biz asosiy Tacotron model arxitekturasidan foydalanamiz [2], bunda biz matndan kelib chiqqan fonema ketma-ketligini va joylashuv sezgir e'tibor (JSE) ni qo'llaymiz. Bashorat qilingan spektrogrammalarni to'lqin shakliga aylantirish uchun, tajriba sikllarining tezligi uchun Griffin-Lim algoritmidan [24] foydalanamiz, chunki biz yuqori sifatli nutq yaratishdan ko'ra ma'lumot samaradorligi muammosiga e'tibor qaratamiz. Asosiy modelda, model noldan o'qitiladi, ya'ni barcha model parametrlari juft ma'lumotlar yordamida o'qitiladi.

2.1. Yarim-nazoratli pre-training

Asosiy Tacotron modelida, model bir vaqtning o'zida matnli vakillarni, akustik vakillarni va ularning o'rtasidagi moslikni o'rganishi kerak. [20] tashqi matnli va akustik ma'lumotlardan foydalanish uchun ikki turdagi model pre-training ni taklif qiladi. Matnli vakillar uchun, ular Tacotron ning kodlovchisini tashqi so'z-vektorlar yordamida pre-training qilishadi; akustik vakillar uchun, ular dekoderini transkripsiya qilinmagan nutq yordamida pre-training qilishadi. Keyin [20] butun modelni juft ma'lumotlar yordamida fine-tuning qiladi. Ushbu bosqichda model matnli vakillar va akustik vakillar o'rtasidagi mosliklarni o'rganishga e'tibor qaratadi.

2.2. O'qituvchisiz o'rganish uchun pre-training

Garchi [13] taklif qilingan yarim-nazoratli pre-training modelning aniq nutqni sintez qilishiga yordam beradi, u kodlovchini va dekoderini bir vaqtning o'zida alohida pre-training qilish ko'proq yaxshilanishni olib kelmasligini ko'rsatadi. Biroq, faqat dekoderini pre-training qilish va butun modelni fine-tuning qilish o'rtasida moslik mavjud emas. Ushbu moslik tomonidan kiritilgan potentsial xatolarni oldini olish va faqat nutqdan foydalanib ma'lumot samaradorligini yanada oshirish uchun biz transkripsiya qilinmagan nutqdan o'qituvchisiz lingvistik birliklarni chiqarishni taklif qilamiz. Taklif qilingan framework Algoritm 1 da keltirilgan. Barcha framework ikkita modelni o'z ichiga oladi: o'qituvchisiz lingvistik birliklarni chiqarish uchun model va Tacotron modeli

2.2.1. O'qituvchisiz lingvistik birliklar

O'qituvchisiz nutq tasviri nutqning ham tasvirlashda, ham ajratishda katta yutuqqa erishdi [25–30]. Ularning orasida, diskret tasvirlar matnli yoki lingvistik ma'lumotlardan tashkil topgan nutq va matnni tahlil qilishda mashhur. Ushbu maqolada, biz VQ-VAE modelini [28] diskret lingvistik birliklarni chiqaruvchi sifatida foydalanamiz. Bu holda, VQ-VAE avtomatik nutqni tanib olish (ASR) modeliga o'xshash tanish model sifatida ishlaydi. Biroq, VQ-VAE va ASR modeli o'rtasidagi asosiy farq shundaki, VQ-VAE o'qituvchisiz tarzda o'qitiladi, ASR modeli esa nazoratli tarzda o'qitiladi. Ushbu farq past resursli tillar uchun muhim. Past resursli tillar uchun ASR modeli odatda mavjud emas, taklif qilingan o'qituvchisiz usul lingvistik birliklarni chiqarish uchun foydali bo'lib qoladi.

Algoritm 1: Taklif Qilingan Usul

Step 1: VQ-VAE ni transkripsiya qilinmagan nutq bilan o'qitish

Step 2: Tacotron Pre-training:

- Step 2.1: O’qituvchisiz lingvistik birliklarni chiqarish:
- Nutq matni transkripsiya qilinmagan nutqqa joylashtiring va o’qituvchisiz lingvistik birliklarni eng yaqin embeddinglar sifatida chiqaring.
- Ketma-ketlikdan ketma-ketlikgaa bir xil birlikni o'chirib tashlang.

- Step 2.2: Tacotron ni <lingvistik birlik, audio> juftlari bilan pre-training qilish.

Step 3: Tacotron ni <matn, audio> juftlari bilan fine-tuning qilish.

VQ-VAE kodlovchi-dekodlovchi arxitekturaga ega va kodlar lug'atiga ega $e = C * D$, bu yerda C lug'atdagi latent vakillar sonini va D har bir vakilning o'lchamini bildiradi. Kodlovchi E $x_{1:T} = x_1, x_2, \dots x_T$ to'lqin shaklini qabul qiladi va kodlangan vakillik $z_{1:N} = E(x_{1:T})$ni hosil qiladi, bu yerda N uzunlik T va kodlovchidagi pasaytirish qatlamlarining soniga bog'liq. Keyin uzluksiz latent vakillik $z_{1:N}$ ga pre-defined diskret vakillarni lug'atdagi eng yaqiniga joylashtirish orqali o'tkazilishi mumkin $z = e_k$, bu yerda $k = argmin_j||z - e_j||$ va e_j kodlar lug'atidagi j-vakilni bildiradi, j esa 1, 2, ⋯, C. Nihoyat, latent vakillar $z_{1:N}$ va so'zlovchi vakili s birga dekoder D ga xom to'lqin shaklini qayta tuzish uchun o'tkaziladi $x = D(z, s)$. Chunki model kiritish va chiqarish bir xil, model avto-kodlovchi sifatida o'qitilishi mumkin. Biroq, argmin operatsiyasidan gradyentlar olinmaydi, shuning uchun [28] straight-through gradyent taxmin qilishni foydalanadi. Keyin butun modelning yakuniy yo'qotilishi:

$$L = -log(x|z(x), s) + ||sg(z(x)) - e_j)||_2^2 + \beta * ||z(x) - sg(ej)||_2^2$$

bu yerda birinchi atama barcha modelni yangilash uchun negative log-likelihood. Ikkinchi atama kodlar lug'atini yangilaydi, sg esa gradyent to'xtatish operatsiyasini bildiradi. Uchinchi atama, commitment loss deb ataladi, kodlovchi chiqishi z ning kodlar lug'atining vakillariga yaqinlashishini rag'batlantiradi, beta esa ushbu atamani og'irligini bildiruvchi gipermarametr.

2.2.2. Tacotron pre-training va fine-tuning

VQ-VAE o'qitilgandan so'ng, biz har bir nutq uchun unsupervised lingvistik birliklarni chiqaramiz. Keyin biz barcha o'qituvchisiz lingvistik birliklar uchun tasodifiy boshlang'ich jadval yaratamiz va jadvalni qidirib topish orqali lingvistik vakillik ketma-ketligini Tacotron kiritmasi sifatida foydalanamiz. Shunday qilib, biz Tacotron ni <lingvistik vakillik, audio> juftlari yordamida pre-training qilishimiz mumkin.

Model yuqorida aytib o'tilganidek pre-training qilinganidan so'ng, biz modelni ba'zi juft nutq ma'lumotlari yordamida fine-tuning qilamiz. Ushbu bosqichda, modelning kiritmalari matndan kelib chiqqan fonema ketma-ketligidir.

1-jadval: To'rt model variantining MCD obyektiv testi natijalari (kichikroq - yaxshiroq). Barcha modellar 24 daqiqa nutq bilan o'qitilgan. Eng yaxshi model (modelning yuqori chegarasi bundan mustasno) qalin qilib belgilangan.

Model	MCD
Tac	22.24
T-Dec	19.57
T-VQ	19.06
T-Phone	18.85

2-jadval: Har bir model juftligining AB testi natijalari. Barcha modellar 24 daqiqa juft ma'lumotlar bilan o'qitilgan.

Model juftligi	*Imtiyoz %*
Tac vs. T-Dec	*1.25 (Tac) vs. 80.25 (T-Dec) vs. 18.5 (N/A)*
Tac vs. T-VQ	*0 (Tac) vs. 97.5 (T-VQ) vs. 2.5 (N/A)*
T-Dec vs. T-VQ	*4.25 (T-Dec) vs. 85 (T-VQ) vs. 10.75 (N/A)*
T-VQ vs. T-Phone	6.25 (T-VQ) vs. 20 (T-Phone) vs. 73.5 (N/A)

3. Eksperiment

3.1. Eksperiment sozlamalari

Taklif qilingan usulimizning samaradorligini ko'rsatish uchun biz tajribalar o'tkazamiz. Biz model fine-tuning uchun LJspeech to'plamidan foydalanamiz. Ushbu maqolada o'rganilgan VQ-VAE arxitekturasi [30] ga o'xshash. VQ-VAE ni o'qitishda, biz model kiritmasi sifatida 39-o'lchovli MFCC dan foydalanamiz. Dastlabki tadqiqotimizdan so'ng, biz kodlar lug'ati hajmini 256 ga, har bir vakilning o'lchamini 64 ga o'rnatdik. Jitter darajasi va beta 0.12 va 0.25, mos ravishda. Tafsilotlar uchun biz o'quvchilarni [30] o'qishni tavsiya qilamiz.

24 daqiqa nutq - bu bazaviy Tacotronni intellektli nutq yaratishga kamdan-kam muvaffaqiyatli qurish mumkin bo'lgan maksimal ma'lumot miqdori ekanligini [20] da topdik. Shunday qilib, keyingi bo'limda biz faqat 24 daqiqa juft ma'lumotlar bilan o'qitilgan barcha model variantlarini taqqoslashga e'tibor qaratamiz.

3.2. 24 daqiqa ma'lumotlar bo'yicha natijalar

Ushbu bo'limda o'rganilgan model variantlari quyidagilardan iborat:

- Tac: faqat LJspeech tomonidan o'qitilgan Tacotron;
- T-Dec: yarim-nazoratli rejimda tashqi nutq bilan pre-training qilingan, keyin LJspeech bilan fine-tuning qilingan Tacotron;
- T-VQ: taklif qilingan rejimda tashqi nutq bilan pre-training qilingan, keyin LJspeech bilan fine-tuning qilingan Tacotron;
- T-Phone: nazoratli rejimda tashqi nutq bilan pre-training qilingan, keyin LJspeech bilan fine-tuning qilingan Tacotron, bu esa modelning yuqori chegarasini bildiradi.

Ushbu bo'limda tashqi nutq ma'lumotlari sifatida biz VCTK dan foydalanamiz [32]. Yuqorida aytib o'tilganidek, T-Dec va T-VQ ni pre-training qilishda VCTK dagi nutq ma'lumotlaridan faqat foydalanamiz. T-Phone uchun pre-training uchun VCTK dagi <matn, audio> juft ma'lumotlaridan foydalanamiz, bu esa ushbu ssenariydagi yuqori chegarani ta'minlaydi. Biz sintez sifatini o'lchash uchun obyektiv va subyektiv baholashlarni o'tkazamiz. Obyektiv baholash uchun biz Dynamic-time-warping Mel-cepstral Distortion (DTW MCD) ni hisoblaymiz, bu esa sintez qilingan va yer haqiqati nutqi o'rtasidagi masofani o'lchaydi va kichikroq bo'lishi yaxshiroqdir. Biz baholash ma'lumotlari sifatida taxminan 20 daqiqa ko'rinmagan nutqni ishlatamiz. Subyektiv baholash uchun biz turli uzunlikdagi 20 minutlik nutqdan foydalanib bir qator AB imtiyoz testlarini o'tkazamiz. 20 ta baholovchi (o'n kishi va o'n ayol) o’zbek tilida ona tili bo'lgan va ingliz tilida yaxshi bo'lgan shaxslar subyektiv testda ishtirok etishadi.

3.2.1. MCD obyektiv testi

MCD natijalari 1-jadvalda keltirilgan. [20] dagi kabi, faqat dekoderini pre-training qilish MCD ni kamaytirishi mumkin. Biroq, taklif qilingan framework eng yaxshi natijani beradi, bu esa bazaviy Tacotron dan 14.30% kamroq MCD ni ko'rsatadi. Biz T-VQ ning natijalari modelning yuqori chegarasi (ya'ni T-Phone) natijalariga yaqinligini ham topdik.

3.2.2. AB subyektiv testi

AB testi natijalari 2-jadvalda keltirilgan. Barcha pre-training texnikalari model natijasini yaxshilashga yordam beradi. Bazaviy Tacotron va pre-training qilingan modellar (ya'ni T-Dec va T-VQ) o'rtasidagi katta natija farqi mavjud. LJspeech bilan noldan o'qitilgan modelni aniqlangan nutqqa yetkazish qiyin bo'lganligini aniqlaymiz, bu qisman LJspeech ning sifati qoniqarli emasligi bilan bog'liq. T-Dec va taklif qilingan T-VQ o'rtasidagi AB testida, T-VQ baholovchilardan ko'proq imtiyozga ega bo'ladi. Norasmiy tinglash testidan biz T-Dec tomonidan sintez qilingan nutqni mo'tadil darajada intellektual deb hisoblaymiz, T-VQ esa ko'proq aniq nutqni yaratadi. Bu shuni ko'rsatadiki, o’qituvchisiz lingvistik birliklar va audio orqali pre-training qilish model natijasini yanada yaxshilashi mumkin. Taklif qilingan pre-training bosqichida model nafaqat akustik vakillikni, balki akustik va matnli vakillik o'rtasidagi moslikni ham o'rganishi mumkin. Garchi o’qituvchisiz lingvistik vakillar modelni fine-tuning qilishda ishlatilmasa ham, bu pre-training matnli vakillikni o'rganishga foydali bo'lishi mumkin, chunki ushbu o’qituvchisiz lingvistik birliklar fonema-gacha ekanligi isbotlangan [30]. Tac-VQ va T-Phone o'rtasidagi taqqoslashda, ko'pchilik baholovchilar imtiyoz ko'rsatmaydi, garchi baholovchilar T-Phone ni T-VQ ga nisbatan 20% ko'proq imtiyozga ega bo'lishadi.

3.3. Boshqa miqdordagi ma'lumotlar bo'yicha natijalar

Biz barcha model variantlarida turli miqdordagi ma'lumotlar bilan MCD obyektiv baholashni o'tkazamiz. Natijalar 3-rasmda keltirilgan. 3-rasmdagi har bir egri chiziq har bir model variantining turli miqdordagi juft ma'lumotlar bilan modelni o'qitish/fine-tuning qilish uchun erishilgan natija va yer haqiqati nutqi o'rtasidagi MCD ni ko'rsatadi. Egri chiziqlardan ko'rinib turibdiki, bazaviy Tacotron va boshqa model variantlari

o'rtasida 1 bo'lakda (ya'ni 24 daqiqa) katta farq mavjud. Yana bir aniq tendensiya shundaki, juft ma'lumotlar miqdori oshgani sayin MCD farqlari kamayadi, bu esa pre-training ning ta'sirini kamayishini bildiradi. Biroq, juft ma'lumotlar miqdoridan qat'i nazar, T-VQ va T-Phone har doim Tac va T-Dec ga qaraganda pastroq MCD ga ega.

3.4. Kam resursli tillar bo'yicha natijalar

Ushbu bo'limda biz taklif qilingan yondashuvning ikkita taxminiy past resursli tillar uchun samaradorligini tasdiqlaymiz. Ushbu bo'limda, biz ingliz va o’zbek tillarini ikkita past resursli tillar deb hisoblaymiz, bu tillarda katta miqyosda va omma e'tiboriga havola qilingan nutqni yig'ish qiyin bo'ladi. Shunday qilib, biz modelni boshqa tillardagi omma e'tiboriga havola qilingan nutq bilan pre-training qilishga murojaat qilamiz. Ushbu bo'limda biz asosan quyidagi ikkita savolga javob berishga qaratamiz:

1. Taklif qilingan usulimiz bu holda ma'lumot samaradorligini oshirishga foydalimi?

2. Taklif qilingan frameworkda qaysi pre-training tillar samaraliroq? Maqsadli til bilan akustik jihatdan yaqin tillar yoki akustik jihatdan o'xshamaydigan tillar?

Ushbu bo'limda, ingliz TTS uchun juft ma'lumotlar LJspeech dan, va o’zbek tili uchun esa audio kitobni olamiz. VQ-VAE ni o'qitish va Tacotron ni pre-training qilish uchun biz quyidagi beshta tilda ochiq manbali korpusdan foydalanamiz: koreys [34], yapon [35], ispan, frantsuz, nemis [36]. Yuqorida aytib o'tilganidek, biz VQ-VAE ni o'qitish va Tacotron ni pre-training qilish uchun faqat nutqdan foydalanamiz. VQ-VAE ni o'qitishda faqat bitta o'zgartirish kiritiladi: kodlar lug'ati hajmi 256 dan 512 ga o'zgartiriladi, chunki ushbu ssenariyada ko'p tilli ma'lumotlar ishlatiladi. Ingliz va o’zbek TTS modelini qurishda biz quyidagi uchta model variantlarini o'rganamiz:

- Tac: LJspeech yoki audio kitobdan olingan juft ma'lumotlar bilan o'qitilgan Tacotron;
- T-VQ-A: taklif qilingan rejimda, Osiyo tillaridagi (ya'ni koreys va yapon) tashqi nutq bilan pre-training qilingan, keyin LJspeech yoki audio kitobdan olingan juft ma'lumotlar bilan fine-tuning qilingan Tacotron;

- T-VQ-E: taklif qilingan rejimda, Yevropa tillaridagi (ya'ni ispan, frantsuz va nemis) tashqi nutq bilan pre-training qilingan, keyin LJspeech yoki Xiaomin dan olingan juft ma'lumotlar bilan fine-tuning qilingan Tacotron;

Baholovchilarni yukini yengillashtirish uchun, biz ushbu bo'limda faqat MCD obyektiv test natijalarini taqdim etamiz. Ingliz va o’zbek TTS ning MCD natijalari mos ravishda 3-jadval va 4-jadvalda keltirilgan. Bu shuni aniq ko'rsatadiki, taklif qilingan pre-training yondashuvi sintez qilingan nutqning sifatini yaxshilaydi, bu esa past resursli tillar uchun muhim, chunki juft ma'lumotlarni yig'ish ancha qiyin bo'ladi.

3-jadval: Ingliz TTS da turli miqdordagi juft ma'lumotlar bo'yicha model variantlarining MCD natijalari (shardlarda). Yaxshi natijalar qalin qilib belgilangan.

Model	**0.5 shard**	**1 shard**	**1.5 shard**	**2 shard**	**2.5 shard**	**3 shard**
Tac	28.72	22.24	21.10	20.39	19.10	18.90
T-VQ-A	25.25	20.77	19.64	18.92	18.62	18.50
T-VQ-E	**24.2**	**20.14**	**18.73**	**18.54**	**18.56**	**18.45**
T-VQ	-	**19.06**	-	**18.33**	-	**18.09**

4-jadval: O’zbek TTS da turli miqdordagi juft ma'lumotlar bo'yicha model variantlarining MCD natijalari (shardlarda). Yaxshi natijalar qalin qilib belgilangan.

Model	**0.5 shard**	**1 shard**	**1.5 shard**	**2 shard**	**2.5 shard**	**3 shard**
Tac	24.18	23.55	22.59	21.67	20.13	19.73
T-VQ-A	**23.48**	**18.44**	16.91	16.31	**15.81**	**15.49**
T-VQ-E	23.69	18.63	**16.81**	**16.29**	16.02	15.93

Bundan tashqari, ingliz TTS da T-VQ-E T-VQ-A ga nisbatan samaraliroq va o’zbek tili tajribasida T-VQ-A ko'p hollarda T-VQ-E ga nisbatan biroz samaraliroq. Bu natija maqsadli til bilan akustik jihatdan yaqin tillar bilan pre-training qilish akustik jihatdan o'xshamaydigan nutq bilan pre-training qilishga qaraganda samaraliroq ekanligini ko'rsatadi. Shuningdek, oldingi bo'limdagi eng yaxshi model variantini (ya'ni

T-VQ) ushbu bo'limdagi eng yaxshi model varianti bilan taqqoslaganda (T-VQ-E), biz hali ham maqsadli tilning nutqi bilan pre-training qilish va akustik jihatdan yaqin til bilan pre-training qilish o'rtasida sezilarli farq borligini topdik (3-jadvaldagi qalin satrlarga qarang), bu esa yana qo'shimcha tadqiqotni talab qiladi.

4. Xulosa

Ushbu maqolada, biz past resursli tillar uchun ketma-ketlikdan ketma-ketlikka TTS da ma'lumot samaradorligini oshirish uchun o’qituvchsiz o'rganishni taklif qilamiz. Bizning usulimiz Tacotron ga katta miqyosda transkripsiya qilinmagan nutqdan foydalanib matnli va akustik ma'lumotlarni tashqi tarzda taqdim etadi. Biz taklif qilingan yondashuvning ketma-ketlikdan ketma-ketlikka TTS frameworkda ishlashini ko'rsatdik. Aniqrog'i, taklif qilingan pre-training usuli bilan, Tacotron kamroq juft o'quv ma'lumotlari bilan intellektual nutq yaratishi mumkin. Garchi biz o'z tajribalarimizda Tacotron arxitekturasidan foydalansak ham, bizning framework boshqa ketma-ketlikdan ketma-ketlikka TTS modellarida ham ishlashini ishonamiz. Biz taxminiy past resursli tillar uchun usulning samaradorligini tasdiqlaymiz. Bu shuni umidvor qiladi ki, hatto maqsadli bo'lmagan transkripsiya qilinmagan nutq bilan ham, taklif qilingan yondashuvimiz sezilarli natija yaxshilanishini ta'minlay oladi. Garchi biz taxminiy past resursli tillardan foydalansak ham, biz usulimiz haqiqiy past resursli tillarga umumlashishi mumkinligini ishonamiz. Garchi umidvor natijalar keltirilgan bo'lsa-da, ko'p narsalarni o'rganish kerak. Misol uchun, ko'plab boshqa o’qituvchisiz modellarni o'rganish kerak. Bundan tashqari, taklif qilingan framework ning samaradorligini tasdiqlashga e'tibor qaratganimiz sababli, biz Griffin-Lim dan spektrogrammani to'lqin shakliga aylantirish algoritmi sifatida foydalanamiz. Kichik juft-ma'lumotli ketma-ketlikdan ketma-ketlikka TTS ni to'liq amalga oshirish uchun, biz neural vocoderlar ni kichik ma'lumotlar bilan moslashtirishni o'rganishimiz kerak.

Manbalar

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio va boshqalar, "Tacotron: Towards end-to-end speech synthesis," Proc. Interspeech 2017, pp. 4006–4010, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan va boshqalar, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4779–4783.

[3] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman va J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in International Conference on Learning Representations, 2018.

[4] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville va Y. Bengio, "Char2wav: End-to-end speech synthesis," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=B1VWyySKx

[5] W. Ping, K. Peng va J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in International Conference on Learning Representations, 2018.

[6] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie va boshqalar, "Sample efficient adaptive text-to-speech," in International Conference on Learning Representations, 2018.

[7] H. B. Moss, V. Aggarwal, N. Prateek, J. González, va R. Barra-Chicote, "BOFFIN TTS: Few-shot speaker adaptation by bayesian optimization," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7639–7643.

[8] J. Park, K. Zhao, K. Peng va W. Ping, "Multi-speaker end-to-end speech synthesis," arXiv preprint arXiv:1907.04462, 2019.

[9] E. Nachmani, A. Polyak, Y. Taigman, va L. Wolf, "Fitting new speakers based on a short untranscribed sample," in International Conference on Machine Learning, 2018, pp. 3683–3691.

[10] Y. Deng, L. He, va F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," arXiv preprint arXiv:1812.05253, 2018.

[11] S. Arik, J. Chen, K. Peng, W. Ping, va Y. Zhou, "Neural voice cloning with a few samples," in Advances in Neural Information Processing Systems, 2018, pp. 10 019–10 029.

[12] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu va boshqalar, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Advances in neural information processing systems, 2018, pp. 4480–4490.

[13] O. S. Watts, "Unsupervised learning for text-to-speech synthesis," Ph.D. dissertation, The University of Edinburgh, 2012.

[14] O. Watts, Z. Wu, va S. King, "Sentence-level control vectors for deep neural network speech synthesis," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[15] P. Wang, Y. Qian, F. K. Soong, L. He, va H. Zhao, "Word embedding for recurrent neural network based tts synthesis," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 4879–4883.

[16] E. Cooper va X. Wang, "Utterance selection for optimizing intelligibility of TTS voices trained on asr data," Interspeech 2017, vol. 1, 2017.

[17] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, va I. Ouyang, "Data selection for improving naturalness of TTS voices trained on small found corpuses," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 319–324.

[18] F. Kuo, I. Ouyang, S. Aryal, va P. Lanchantin, “Selection and training schemes for improving TTS voice built on found data,” Proc. Interspeech 2019, pp. 1516–1520, 2019.

[19] J. Fong, P. O. Gallegos, Z. Hodari, va S. King, “Investigating the robustness of sequence-to-sequence text-to-speech models to imperfectly-transcribed training data,” in Proc. Interspeech, 2019.

[20] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, va R. Skerry-Ryan, “Semi-supervised training for improving data efficiency in end-to-end speech synthesis,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6940–6944.

[21] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, va L. Cai, “Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5545–5549.

[22] A. Gutkin, “Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages,” Proc. Interspeech 2017, pp. 2183–2187, 2017.

[23] Y.-J. Chen, T. Tu, C.-c. Yeh, va H.-Y. Lee, “End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning,” Proc. Interspeech 2019, pp. 2075–2079, 2019.

[24] N. Perraudin, P. Balazs, va P. L. Søndergaard, “A fast Griffin-Lim algorithm,” in 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013, pp. 1–4.

[25] H. Lee, P. Pham, Y. Largman, va A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in Advances in neural information processing systems, 2009, pp. 1096–1104.

[26] J. Glass, “Towards unsupervised speech processing,” in 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA). IEEE, 2012, pp. 1–4.

[27] W.-N. Hsu, Y. Zhang, va J. Glass, “Unsupervised learning of disentangled and interpretable representations from sequential data,” in Advances in neural information processing systems, 2017, pp. 1878–1889.

[28] A. Van Den Oord, O. Vinyals va boshqalar, “Neural discrete representation learning,” in Advances in Neural Information Processing Systems, 2017, pp. 6306–6315.

[29] Y.-A. Chung, W.-N. Hsu, H. Tang, va J. Glass, “An unsupervised autoregressive model for speech representation learning,” Proc. Interspeech 2019, pp. 146–150, 2019.

[30] J. Chorowski, R. J. Weiss, S. Bengio, va A. van den Oord, “Unsupervised speech representation learning using WaveNet autoencoders,” IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 12, pp. 2041–2053, 2019.

[31] K. Ito, “The LJ speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[32] C. Veaux, J. Yamagishi, va K. MacDonald, “Superseded - CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017. [Online]. Available: http://datashare.is.ed.ac.uk/handle/10283/2651

[33] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1. IEEE, 1993, pp. 125–128.

[34] “Zeroth-Korean, http://www.openslr.org/40.”

[35] R. Sonobe, S. Takamichi, va H. Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017.

[36] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, va G. Weber, “Common Voice: A massively-multilingual speech corpus,” in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 4218–4222.