CHUQUR O‘QITISH MODELLARI VA ULARNING TTS TIZIMLARIDAGI O‘RNI

Raximov Nodir Odilovich[1], Primqulov Oybek Dilmurodovich[1], Erkinova Dilnoza Abdurazzoq qizi[1]
[1]Toshkent Axborot Texnologiyalari Universiteti
E-mail: primqulovoybek1@gmail.com
ORCID: 0000-0001-9118-5220

K E Y W O R D S

chuqur o‘qitish, sun’iy neyron tarmoqlar, Text-to-Speech, TTS, Transformer, diqqat mexanizmi, rekurrent neyron tarmoqlar, FastSpeech, VITS, prosodik modellashtirish, nutq sintezi, o‘zbek tili

A B S T R A C T

Ushbu maqolada chuqur o‘qitish modellarining nazariy asoslari hamda ularning matnni nutqqa aylantirish (Text-to-Speech, TTS) tizimlaridagi o‘rni tahlil qilingan. Sun’iy neyron tarmoqlarning matematik asoslari, TTS tizimlarining umumiy arxitekturasi, rekurrent neyron tarmoqlar, Transformer va diqqat mexanizmi asosidagi yondashuvlar ko‘rib chiqilgan. Shuningdek, autoregressiv, non-autoregressiv, oqimga asoslangan va generativ qarama-qarshi neyron tarmoqlar asosidagi zamonaviy TTS modellarining ishlash tamoyillari yoritilgan. Nutq sintezida prosodik modellashtirish va ko‘p komponentli yo‘qotish funksiyalarining ahamiyati tahlil qilingan. O‘zbek tilida TTS tizimlarini yaratishda uchraydigan lingvistik va prosodik muammolar hamda ularni hal etishning istiqbolli yo‘nalishlari ko‘rsatib berilgan.

Kirish

Sun’iy intellekt va tabiiy tilni qayta ishlash texnologiyalarining jadal rivojlanishi inson va kompyuter o‘rtasidagi o‘zaro muloqotni yanada tabiiy va qulay shakllarga olib chiqmoqda. Matnni nutqqa aylantirish (Text-to-Speech, TTS) tizimlari ovozli yordamchilar, avtomatik tarjima tizimlari, nogironligi bo‘lgan shaxslar uchun yordamchi texnologiyalar, ta’lim platformalari hamda multimedia ilovalarida keng qo‘llanilmoqda. Nutq sintezi texnologiyalarining rivojlanishi natijasida inson nutqiga yaqin bo‘lgan tabiiy va ravon audio signallarni hosil qilish imkoniyati sezilarli darajada kengaydi.

So‘nggi yillarda Tacotron, FastSpeech, VITS va boshqa zamonaviy neyron modellar TTS tizimlarining samaradorligini sezilarli darajada oshirdi. Biroq morfologik jihatdan murakkab va agglutinativ tuzilishga ega bo‘lgan o‘zbek tilida nutq sintezi tizimlarini yaratish hali ham dolzarb ilmiy muammolardan biri bo‘lib qolmoqda. Nutq korpuslarining cheklanganligi, prosodik xususiyatlarning o‘ziga xosligi hamda lingvistik murakkabliklar mavjud yondashuvlarni chuqur tahlil qilishni talab etadi.

Tadqiqotning maqsadi chuqur o‘qitish modellarining matematik asoslarini, zamonaviy TTS tizimlarining arxitekturasi va ishlash tamoyillarini tahlil qilish hamda nutq sintezida qo‘llaniladigan asosiy neyron yondashuvlarning imkoniyatlarini o‘rganishdan iborat. Shuningdek, o‘zbek tilida nutq sintezi tizimlarini yaratishda uchraydigan muammolar va ularni hal etishning istiqbolli yo‘nalishlari ko‘rib chiqiladi. Sun’iy intellekt sohasida chuqur o‘qitish modellarining rivojlanishi nutq texnologiyalari sifatini yangi bosqichga olib chiqdi. Avvalgi mashinaviy o‘rganish yondashuvlari ajratilgan xususiyatlar va statistik modellarga asoslangan bo‘lsa, chuqur o‘qitish modellarida ma’lumotlardan bevosita avtomatik xususiyat ajratish amalga oshiriladi. Bu esa nutq signali kabi murakkab va yuqori o‘lchamli ma’lumotlarni samarali modellashtirish imkonini beradi. Text to Speech matnni tabiiy eshitiladigan nutqqa aylantiruvchi texnologiya hisoblanadi. An’anaviy TTS tizimlari cheklangan tabiiylik va moslashuvchanlikka ega edi.

Chuqur o‘qitish modellarida sun’iy neyron tarmoqning o‘rni.

(cc) BY

Chuqur o‘qitish modellari paydo bo‘lishi bilan TTS sifati inson nutqiga juda yaqinlashishni boshladi [1].

Chuqur o‘qitish ko‘p qatlamli sun’iy neyron tarmoqlarga asoslanadi. Oddiy neyron quyidagi ko‘rinishda ifodalanadi:

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \qquad (1)$$

bunda, x_i-kirish qiymatlar, w_i-og’irlik koeffitsiyentlari, b-bias ma’lumotlari, f-nolinear aktivatsiya funksiya.

Ko‘p qatlamli tuzilma yordamida model yuqori darajadagi abstrakt xususiyatlarni o‘rganadi. Nutq sintezida bu fonetik, akustik va prosodik bog‘lanishlarni aniqlash imkonini beradi. Modelni o‘qitish jarayoni odatda yo‘qotish funksiyasini minimallashtirishga asoslanadi. Masalan, TTS tizimlarida keng qo‘llaniladigan o‘rtacha kvadratik xato (MSE) quyidagicha ifodalanadi: [2]

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i^{true} - y_i^{pred})^2 \qquad (2)$$

bunda, y_i^{true} – haqiqiy qiymat, y_i^{pred} – model bashorati, N – namunala soni.

Chuqur neyron tarmoq tizimlarining matematik asoslari

Chuqur o‘qitish modellari sun’iy neyron tarmoqlarga asoslanadi. Sun’iy neyron inson miyasidagi biologik neyronning soddalashtirilgan matematik modeli bo‘lib, kirish signallarini qabul qiladi, ularni ma’lum og‘irlik koeffitsiyentlari orqali qayta ishlaydi va nolinear transformatsiya yordamida chiqish qiymatini hosil qiladi. Neyronning asosiy vazifasi kirish ma’lumotlari orasidagi yashirin bog‘lanishlarni aniqlash va ularni umumlashtirilgan matematik ifodaga keltirishdan iborat.

Sun’iy neyron ikki bosqichda ishlaydi:

1. Kirish signallarining chiziqli kombinatsiyasi hisoblanadi;

2. Hosil bo‘lgan qiymat nolinear aktivatsiya funksiyasi orqali o‘tkaziladi.

Matematik jihatdan oddiy sun’iy neyron quyidagicha ifodalanadi:

$$y=f(\sum_{i=1}^{n} w_i x_i+b) \qquad (3)$$

bunda x_i- kirish qiymatlar (inputlar), w_i- har bir kirishga mos og‘irlik koeffitsiyentlari, f - nolinear aktivatsiya funksiyasi, y-neyron chiqish qiymati.

Agar aktivatsiya funksiyasi chiziqli bo‘lsa, model murakkab bog‘lanishlarni o‘rgana olmaydi. Shu sababli chuqur o‘qitish modellarida chiziqli bo’lmagan muhim ahamiyatga ega. Aynan aktivatsiya funksiyasi modelga murakkab lingvistik va akustik munosabatlarni o‘rganish imkonini beradi. TTS tizimlarida bu xususiyat fonetik, prosodik va semantik bog‘lanishlarni modellashtirishda hal qiluvchi rol o‘ynaydi [3].

Ko‘p qatlamli neyron tarmoqda bir nechta yashirin qatlam mavjud bo‘lib, har bir qatlam quyidagicha ifodalanadi:

$$h^{(l)}=f\left(W^{(l)}h^{(l-1)}+b^{(l)}\right) \qquad (4)$$

Modelni o‘qitish jarayoni parametrlarni shunday tanlashga asoslanadiki, model chiqishi haqiqiy qiymatlarga maksimal darajada yaqin bo‘lishi kerak. Buni ta’minlash uchun yo‘qotish funksiyasi kiritiladi. Yo‘qotish funksiyasi model bashorati bilan haqiqiy qiymat o‘rtasidagi farqni miqdoriy baholaydi. Modelni o‘qitish jarayoni yo‘qotish funksiyasini minimallashtirishga asoslanadi [4]. TTS tizimlarida keng qo‘llaniladigan MSE yo‘qotish funksiyasi quyidagicha ifodalanadi:

$$L=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \qquad (5)$$

bunda, L - umumiy yo‘qotish qiymati, y_i - haqiqiy qiymat, $\hat{y}_i$ - model bashorat qilgan qiymat.

TTS tizimining umumiy arxitekturasi.

Zamonaviy neyron TTS (Text to Speech) tizimlari odatda modulli, ammo yagona optimallashtiriladigan arxitekturaga asoslanadi. Ularning asosiy vazifasi matnli kirishni akustik signalga avtomatik va tabiiy tarzda aylantirishdan iborat. Matematik jihatdan ushbu jarayon quyidagicha ifodalanishi mumkin:

$$X_{text}\rightarrow Encoder\rightarrow S_{mel}\rightarrow Vocoder\rightarrow Y_{audio} \qquad (6)$$

bunda, X_{text} - matn kiritish ketma-ketligi, $Encoder$ - lingvistik kodlovchi modul, S_{mel} - mel-spektrogramma (akustik oraliq ifoda),

(cc) BY

$Vocoder$-audio signal generatsiya qiluvchi modul, Y_{audio}-yakuniy nutq signali [5].

Mazkur arxitektura uch asosiy bosqichdan iborat bo‘lib, har bir bosqichning vazifasi va matematik mohiyati alohida ahamiyatga ega.

1. Encoder bosqichi - lingvistik modellashtirish.

Encoder moduli TTS tizimining birinchi va eng muhim bosqichi hisoblanadi. Uning asosiy vazifasi kirish matnini model tomonidan qayta ishlanishi mumkin bo‘lgan raqamli lingvistik ifodaga o‘tkazishdan iborat. Bu jarayon bir nechta ketma-ket bosqichlardan iborat bo‘lib, har biri matematik va lingvistik jihatdan muhim ahamiyatga ega.

1. Matnni tokenizatsiya qilish (harf, bo‘g‘in, fonema yoki subword birliklarga ajratish);

2. Embedding qatlam orqali har bir tokenni vektor ko‘rinishga o‘tkazish;

3. Ketma-ketlikni kontekstual modellashtirish (RNN, LSTM yoki Transformer yordamida). [6]

Agar Transformer asosida qurilgan bo‘lsa, encoder quyidagicha ishlaydi:

$$H=Encoder(X_{text}) \qquad (7)$$

bunda, H - kontekstual lingvistik ifoda.

Encoder matn fonetik tuzulishini, sintaktik bog’lanishlarini, semantik kontekstini, urg’u joylashuvini ichki fazoda kodlaydi. O‘zbek tilining agglutinativ va morfologik jihatdan murakkab tuzilishi sababli subword yoki fonema darajasidagi modellashtirish yanada samarali yondashuv hisoblanadi.

2. Akustik model - mel-spektrogramma generatsiyasi [7].

Encoder chiqishi akustik modelga uzatiladi. Bu modul lingvistik ifodani vaqt bo‘yicha akustik xususiyatlarga aylantiradi. Mel-spektrogramma - bu nutq signalining vaqt-chastota domenidagi energiya taqsimotini ifodalovchi matritsa bo‘lib, u inson eshitish tizimiga mos ravishda mel-shkalada hisoblanadi.

Matematik jihatdan quyidagicha ifodalanadi:

$$S_{mel}=AcousticModel(H) \qquad (8)$$

Akustik model fonema davomiyligi (duration), asosiy chastota (pitch), energiya darajasi, intonatsiya konturi kabi parametrlarni o‘rganadi. Bu bosqich nutq tabiiyligini belgilovchi asosiy qism hisoblanadi.

3. Vokoder mel-spektogrammani to‘lqin shakliga aylantirib berish vazifasini bajaradi.

Vokoderning matematik modeli quyidagicha ifodalanadi:

$$Y_{audio} = Vocoder(S_{mel}) \qquad (9)$$

Vokoder tabiiy tembr, shovqinsiz signal, silliq o‘tishlar, realistik tovush tembri kabi xususiyatlarni ta’minlaydi. Agar vokoder sifati past bo‘lsa, hatto mukammal akustik model ham robotga o‘xshash nutq generatsiya qiladi.

Rekurrent neyron tarmoqlar (RNN, LSTM)

Nutq sintezi jarayoni vaqt bo‘yicha ketma-ketlikni modellashtirishni talab qiladi, chunki nutq signali dinamik va vaqtga bog‘liq jarayon hisoblanadi. Har bir tovush yoki fonema oldingi va keyingi birliklarga kontekstual jihatdan bog‘liq bo‘ladi. Shu sababli ketma-ketlik ma’lumotlarini qayta ishlashga mo1ljallangan rekurrent neyron tarmoqlar (Recurrent Neural Networks - RNN) TTS tizimlarining dastlabki neyron arxitekturalarida muhim rol o‘ynagan[8].

Rekurrent neyron tarmoqlar har bir vaqt bosqichidagi yashirin holatni quyidagicha hisoblaydi:

$$h_t=f(Wx_t+Uh_{t-1}+b) \qquad (10)$$

Biroq oddiy RNN uzoq bog‘lanishlarni saqlay olmaydi. Shu sababli LSTM modeli ishlab chiqilgan bo‘lib, u maxsus mexanizmlar (input, forget, output gates) orqali uzoq muddatli bog‘lanishlarni saqlaydi. TTS tizimlarida LSTM modellar prosodik bog‘lanishlarni barqaror o‘rganishda muhim rol o‘ynaydi [9, 10].

Transformer va diqqat mexanizmi

Ushbu model rekurrent va konvolyutsion yondashuvlardan farqli ravishda ketma-ketlik elementlari o‘rtasidagi bog‘lanishlarni to‘liq diqqat mexanizmi orqali modellashtiradi. Natijada uzoq masofali bog‘lanishlarni samarali o‘rganish va hisoblash jarayonini parallel bajarish imkoniyati paydo bo‘ldi. Diqqat mexanizmi uzoq masofali bog’lanishlarni aniqlaydi, parallel hisoblash imkonini beradi, prosodik va semantik

(cc) BY

kontekstni samarali o‘rganish xususiyati namoyon qiladi va matematik modeli quyidagicha ifodalanadi: [11]

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

TTS tizimlarida Transformer encoder matnni chuqur lingvistik fazoga o‘tkazadi.

Transformer arxitekturasi va diqqat mexanizmi ketma-ketlik modellashtirishni yangi bosqichga olib chiqdi. U global kontekstni samarali hisobga olishi, parallel ishlashi va murakkab lingvistik boglanishlarni aniqlay olishi bilan TTS tizimlarida muhim o‘rin egallaydi. Ayniqsa morfologik jihatdan boy va sintaktik jihatdan murakkab tillarda, jumladan o‘zbek tilida, diqqat mexanizmi prosodik va semantik bog1lanishlarni chuqur modellashtirish imkonini beradi [12,13].

Zamonaviy neyron TTS modellar

Chuqur o‘qitish metodlarining rivojlanishi natijasida TTS tizimlari an’anaviy qoidaviy va HMM asosidagi statistik yondashuvlardan to‘liq neyron arxitekturalarga o‘tdi. Zamonaviy neyron TTS modellarining asosiy xususiyati end to end o‘qitilish, ya’ni matndan to‘g‘ridan-to‘g‘ri akustik ifodaga, so‘ngra audio signalga o‘tish jarayonining yagona optimallashtiriladigan tizim sifatida qurilishi hisoblanadi. Zamonaviy neyron TTS modellarini uch asosiy turga ajratish mumkin:

1. Autoregressiv akustik modellar
2. Non-autoregressiv (parallel) modellar
3. Flow-based va GAN asosidagi modellar

1. Autoregressiv TTS modellar.

Autoregressiv modellar chiqishni ketma-ket generatsiya qiladi. Ya’ni har bir vaqt bosqichidagi akustik frame oldingi frame ga bog‘liq bo‘ladi.

Matematik jihatdan quyidagicha ifodalanadi:

$$p(S)=\prod_{t=1}^{T} p\,(s_t|s_{<t},X) \quad (12)$$

bunda, S - mel-spektogramma, X - matn ketma-ketligi.

2. Non-autoregressiv modellar.

Autoregressiv modellar yuqori sifatli nutq sintezini ta’minlasa-da, ularning asosiy kamchiligi generatsiya jarayonining ketma-ket amalga oshirilishidir. Bunda har bir akustik frame oldingi frame asosida hisoblanadi. Natijada model barcha vaqt bosqichlarini ketma-ket generatsiya qilishi kerak bo‘ladi, bu esa hisoblash jarayonini sezilarli darajada sekinlashtiradi. Ayniqsa real vaqt tizimlari yoki katta hajmdagi matnlarni ovozga aylantirishda bu muammo yanada yaqqol namoyon bo‘ladi.

Ushbu cheklovni bartaraf etish maqsadida non-autoregressiv TTS modellar ishlab chiqilgan. Bu modellar chiqish ketma-ketligini parallel tarzda generatsiya qiladi, ya’ni har bir akustik frame boshqa framelarga bog‘liq bo‘lmagan holda bir vaqtning o‘zida hisoblanadi. Shuning uchun non-autoregressiv modellar hisoblash tezligi jihatidan ancha samarali hisoblanadi.

Matematik jihatdan non-autoregressiv model quyidagicha ifodalanadi:

$$S=f(H) \quad (13)$$

bunda, H - encoder tomonidan hosil qilingan lingvistik yashirin ifoda, S - generatsiya qilingan mel-spektrogramma, f - neyron transformatsiya funksiyasi [14].

Bunday yondashuvda model butun mel-spektrogrammani bir vaqtning o‘zida hisoblaydi. Natijada generatsiya tezligi bir necha barobar oshadi. Non-autoregressiv modellarni samarali ishlashi uchun matn va nutq uzunligi o‘rtasidagi moslikni aniqlash muhim hisoblanadi. Chunki matn tokenlari soni va akustik frame lar soni odatda bir-biriga teng bo‘lmaydi. Shu sababli zamonaviy non-autoregressiv TTS arxitekturalarida duration predictor moduli qo‘llaniladi. Ushbu modul har bir fonema yoki token uchun nutq davomiyligini bashorat qiladi [15].

Matematik jihatdan bu quyidagicha ifodalanadi:

$$d_i=f_{dur}(h_i) \quad `(14)$$

bunda, d_i- i-chi fonemaning davomiyligi, h_i- encoder chiqishidagi yashirin vektor, f_{dur} - davomiylikni bashorat qiluvchi neyron model.

Duration predictor yordamida lingvistik ifoda vaqt bo‘yicha kengaytiriladi va akustik

(cc) BY

model uchun mos ketma-ketlik hosil qilinadi. Bu esa parallel generatsiyani ta'minlaydi.

3. Oqimga asoslangan modellar

Oqimga asoslangan modellar ehtimollik nazariyasiga asoslangan generativ modellar turiga kiradi. Ushbu modellar murakkab ma'lumot taqsimotini oddiy ehtimollik taqsimotidan boshlab ketma-ket invertible transformatsiyalar yordamida o'rganadi. Ya'ni model latent fazodagi oddiy tasodifiy o'zgaruvchini akustik fazodagi murakkab nutq xususiyatlariga mos ravishda transformatsiya qiladi.

Oqimga asoslangan modellarining asosiy g'oyasi shundan iboratki, ma'lumotlar taqsimoti invertible funksiyalar zanjiri orqali modellashtiriladi. Agar latent o'zgaruvchi quyidagi normal taqsimotdan olinadigan bo'lsa:

$$z \sim N(0,I) \qquad (15)$$

unda model quyidagi transformatsiya orqali akustik ifodani hosil qiladi:

$$S = f_\theta(z) \qquad (16)$$

bunda, f_θ - latent fazodagi tasodifiy vektor, S - generatsiya qilingan mel-spektrogramma, z - parametrlar orqali o'rganiladigan invertible transformatsiya [16].

Oqimga asoslangan modellar ehtimollik zichligini aniq hisoblash imkoniyatiga ega bo'lgani sababli o'qitish jarayoni barqaror bo'ladi. Bundan tashqari, ular parallel generatsiyani qo'llab-quvvatlaydi, bu esa nutq sintezini tezlashtiradi.TTS tizimlarida oqimga asoslangan yondashuv alignment muammosini kamaytirishga yordam beradi. Alignment - matn tokenlari bilan akustik frame lar o'rtasidagi moslikni aniqlash jarayonidir. Oqimga asoslangan modellar ushbu moslikni latent fazoda o'rganishi sababli generatsiya jarayoni barqarorroq bo'ladi.Zamonaviy TTS tizimlarida qo'llaniladigan flow-based modellar qatoriga oqim TTS va Flowtron kabi arxitekturalarni kiritish mumkin. Ushbu modellar mel-spektrogrammani yuqori aniqlik bilan generatsiya qilish imkonini beradi hamda prosodik xususiyatlarni samarali modellashtiradi [17, 18].

4. Generativ qarama -qarshi neyron tarmoq asosidagi modellar.

Generativ qarama-qarshi neyron tarmoq asosidagi modellar generativ neyron tarmoqlar oilasiga kiradi. Ushbu yondashuv ikki asosiy komponentdan tashkil topgan: generator va diskriminator. Generator real nutq signaliga o'xshash audio yaratishga harakat qiladi, diskriminator esa generatsiya qilingan signal bilan haqiqiy signalni farqlashga intiladi. Ushbu ikki model o'rtasidagi raqobat jarayoni natijasida yuqori sifatli audio signal hosil qilinadi. Generativ qarama-qarshi neyron tarmoq modellarining o'qitish jarayoni adversarial optimallashtirishga asoslanadi. TTS tizimlarida umumiy yo'qotish funksiyasi quyidagicha ifodalanadi:

$$L = L_{mel} + \lambda L_{adv} \qquad (17)$$

bunda, L_{mel} -mel-spektrogramma bo'yicha rekonstruksiya xatosi, L_{adv} - mel-spektrogramma bo'yicha rekonstruksiya xatosi, λ - vazn koeffitsiyenti.

Raqobatga asoslangan yo'qotish modeli generatsiya qilingan audio signalning perceptual sifatini oshirishga yordam beradi. Bu esa sintez qilingan nutqning inson nutqiga yanada yaqin bo'lishini ta'minlaydi. Generativ qarama-qarshi neyron tarmoq asosidagi vokoderlar nutq sintezida muhim rol o'ynaydi. Ular mel-spektrogrammadan yuqori sifatli audio signal generatsiya qilish imkonini beradi. [19]

Yo'qotish funksiyalarining kombinatsiyasi.

Chuqur o'qitish asosidagi TTS tizimlarida modelni samarali o'qitish uchun bir nechta yo'qotish funksiyalarining kombinatsiyasidan foydalaniladi. Bitta yo'qotish funksiyasi model sifatini to'liq ifodalay olmaydi, chunki nutq sintezi jarayoni bir vaqtning o'zida bir nechta akustik va prosodik xususiyatlarni o'z ichiga oladi. Shu sababli zamonaviy TTS modellarida turli xil yo'qotish funksiyalari birgalikda optimallashtiriladi. [27]

Umumiy yo'qotish funksiyasi quyidagi ko'rinishda ifodalanadi:

$$L_{total} = \lambda_1 L_{mel} + \lambda_2 L_{dur} + \lambda_3 L_{pitch} + \lambda_4 L_{adv} \qquad (18)$$

bunda, L_{mel} - mel-spektrogramma bo'yicha rekonstruksiya xatosi, L_{dur} - fonema yoki token davomiyligini bashorat qilish xatosi, L_{pitch} -

asosiy chastota (pitch) regressiyasi xatosi, L_{adv} - adversarial yo‘qotish funksiyasi, λ_i - har bir yo‘qotish funksiyasining vazn koeffitsiyentlari [20].

Mel-spektrogramma yo‘qotish funksiyasi model generatsiya qilgan akustik ifoda bilan haqiqiy nutq spektrogrammasi o‘rtasidagi farqni minimallashtirishga xizmat qiladi. Davomiylik yo‘qotish funksiyasi esa fonemalarning vaqt bo‘yicha to‘g‘ri joylashishini ta’minlaydi. Pitch regressiyasi nutqning intonatsion tuzilishini modellashtirishda muhim ahamiyatga ega. Adversarial yo‘qotish esa generatsiya qilingan nutq signalining perceptual sifatini oshiradi. Bir nechta yo‘qotish funksiyalarini birlashtirish modelga bir vaqtning o‘zida akustik aniqlik, fonema davomiyligining mosligi, intonatsiyaning tabiiyligi, eshitish sifatining yaxshilanishi kabi xususiyatlarni optimallashtirish imkonini beradi. Shu sababli zamonaviy TTS modellarida ko‘p komponentli yo‘qotish funksiyalari keng qo‘llaniladi.

Prosodik modellashtirish.

Prosodiya nutqning muhim lingvistik va akustik xususiyatlaridan biri bo‘lib, u nutqning ritmi, urg‘usi, intonatsiyasi va energiya taqsimotini belgilaydi. Nutq sintezi tizimlarida prosodik xususiyatlarni to‘g‘ri modellashtirish generatsiya qilingan nutqning tabiiyligini ta’minlovchi asosiy omillardan hisoblanadi. Prosodiya odatda asosiy chastota (pitch), fonema davomiyligi (duration), signal energiyasi (energy), intonatsiya konturi singari akustik parametrlar orqali ifodalanadi. [21]

Zamonaviy TTS modellarida ushbu parametrlar alohida regressiya modullari orqali modellashtiriladi. Matematik jihatdan prosodik parametrlarni quyidagicha ifodalash mumkin: [22]

$$P = f_{pitch}(H) \qquad (19)$$

$$D = f_{dur}(H) \quad (20)$$

$$E = f_{energy}(H) \qquad (21)$$

bunda, H - encoder tomonidan hosil qilingan lingvistik yashirin ifoda, P - pitch qiymatlari, D - fonema davomiyligi, E - signal energiyasi.

Prosodik modellashtirish jarayonida model lingvistik kontekstni hisobga olgan holda nutq ohangi va ritmini shakllantiradi. Bu jarayon ayniqsa gap darajasidagi intonatsiya va urg‘u strukturasini generatsiya qilishda muhim ahamiyatga ega. O‘zbek tilida prosodik xususiyatlar fonetik va morfologik omillar bilan chambarchas bog‘liq. Masalan, o‘zbek tilida urg‘u ko‘pincha so‘zning oxirgi bo‘g‘iniga tushadi, biroq ayrim grammatik shakllarda urg‘u joylashuvi o‘zgarishi mumkin. Bundan tashqari, qo‘shimchalarning qo‘shilishi fonema davomiyligi va intonatsiya konturiga ta’sir qiladi [23].

Shu sababli o‘zbek tilida TTS tizimini yaratishda prosodik modellashtirish alohida e’tibor talab qiladi. Prosodik parametrlarni aniq modellashtirish nutq sintezining tabiiyligini oshiradi va generatsiya qilingan nutqni inson nutqiga yanada yaqinlashtiradi.

Xulosa

Ushbu tadqiqotda tabiiy tilni qayta ishlash texnologiyalari va nutq sintezi tizimlarining nazariy asoslari hamda zamonaviy yondashuvlari tahlil qilindi. Chuqur o‘qitish modellarining TTS tizimlaridagi o‘rni, shuningdek, Tacotron, FastSpeech va VITS kabi zamonaviy arxitekturalarning ishlash tamoyillari va afzalliklari yoritildi. Tadqiqot natijalari chuqur neyron tarmoqlar yuqori sifatli va tabiiy nutq hosil qilishda muhim ahamiyatga ega ekanligini ko‘rsatdi. Shuningdek, o‘zbek tilida nutq sintezi tizimlarini yaratishda mavjud nutq korpuslarining cheklanganligi, lingvistik murakkablik va prosodik xususiyatlar bilan bog‘liq muammolar mavjudligi aniqlandi. Mazkur natijalar o‘zbek tili uchun samarali va tabiiy TTS tizimlarini ishlab chiqishda zamonaviy chuqur o‘qitish modellaridan foydalanish istiqbolli yo‘nalishlardan biri ekanligini tasdiqlaydi.

ADABIYOTLAR

1. Popov V., Vovk I., Gogoryan V. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech // ICML. – 2022.
2. Vaswani A., Shazeer N., Parmar N., et al. *Attention Is All You Need*. – NeurIPS, 2017.
3. van den Oord A., Dieleman S., Zen H., et al. WaveNet: A Generative Model for Raw Audio. – DeepMind, 2016.

(cc) BY

4. Wang Y., Skerry-Ryan R., Stanton D., et al. Tacotron: Towards End-to-End Speech Synthesis. – Google, 2017.
5. Shen J., Pang R., Weiss R.J., et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. – 2018.
6. Ren Y., Ruan Y., Tan X., et al. FastSpeech: Fast, Robust and Controllable Text to Speech. – NeurIPS, 2019.
7. Kim J., Kong J., Son J. *VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End TTS*. – ICML, 2021.
8. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models. – 2020.
9. Popov V., Vovk I., Gogoryan V., et al. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. – 2021.
10. Jurafsky D., Martin J.H. *Speech and Language Processing*. – Pearson, 3rd edition.
11. Goodfellow I., Bengio Y., Courville A. *Deep Learning*. – MIT Press, 2016.
12. Bishop C.M. *Pattern Recognition and Machine Learning*. – Springer, 2006.
13. Kamilov M.M., Musayev M.M. *Tabiiy tilni qayta ishlash asoslari*. – Toshkent, 2018.
14. Aripov M.M., Ignatyev N.A. *Sun'iy intellekt va ekspert tizimlari*. – Toshkent, 2015.
15. Xamdamov O'.R., Mamatov N.S. *Nutq texnologiyalari va ularning qo'llanilishi*. – Toshkent, 2020.
16. Raximov N.O. *Axborot texnologiyalari va dasturiy ta'minot*. – Toshkent, 2022.
17. Baevski A., Zhou H., Mohamed A., Auli M. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. – 2020.
18. Prenger R., Valle R., Catanzaro B. *WaveGlow: A Flow-based Generative Network for Speech Synthesis*. – 2019.
19. Ping W., Peng K., Chen J., et al. *Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning*. – 2018.
20. OpenAI. *GPT-based language models documentation*. – 2023.
21. HuggingFace. *Transformers Documentation*. – https://huggingface.co
22. Vaswani A., Shazeer N., Parmar N. etal. Attention Is All You Need // Neural Information Processing Systems. – 2021.
23. Wang Y., Skerry-Ryan R., Stanton D. Tacotron: Towards End-to-End Speech Synthesis // Interspeech. – 2022.

(cc) BY