Объединенный институт проблем информатики
Национальной академии наук Беларуси

XXIII Международная
научно-техническая конференция

РАЗВИТИЕ ИНФОРМАТИЗАЦИИ И ГОСУДАРСТВЕННОЙ СИСТЕМЫ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

РИНТИ-2024

21 ноября 2024 г., Минск

Доклады

Минск
ОИПИ НАН Беларуси
2024

УДК 002; 004

Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2024) : доклады XXIII Международной научно-технической конференции, Минск, 21 ноября 2024 г. – Минск : ОИПИ НАН Беларуси, 2024. – 424 с. – ISBN 978-985-7198-18-4.

Представлены доклады XXIII Международной научно-технической конференции «Развитие информатизации и государственной системы научно-технической информации» (РИНТИ-2024), Минск, 21 ноября 2024 г., в которых рассмотрены порядок оценки эффективности мероприятий по развитию систем научно-технической информации, результаты научно-методического обеспечения развития информатизации в НАН Беларуси в 2023–2024 гг., назначение и структура «Офиса цифровизации» НАН Беларуси, вопросы правового регулирования в области искусственного интеллекта в Республике Беларусь, концептуальная схема киберфизической системы «умного» города, подходы к стратегическому планированию цифрового развития на 2026–2030 гг. и на перспективу до 2035 г., искусственный интеллект в образовании, эффективное управление цифровым развитием и др.

Рассмотрены вопросы научно-методического, информационного, технологического и правового обеспечения цифровой трансформации, проектирования и внедрения автоматизированных систем научно-технической информации, библиотечно-информационных систем и технологий, публикационной активности ученых, а также искусственного интеллекта и когнитивных технологий в информатизации.

Материалы конференции будут полезны специалистам в области информационно-коммуникационных технологий, занимающихся научно-методическим обеспечением информатизации и решением задач построения ИТ-страны, цифровой экономикой, разработкой и внедрением автоматизированных информационных систем управления, систем научно-технической информации, автоматизированных библиотечно-информационных систем и технологий, а также развитием информационной инфраструктуры Беларуси и других стран, реализацией проектов государственных и отраслевых программ в сфере информатизации.

Одобрены программным комитетом и печатаются по решению редакционной коллегии Объединенного института проблем информатики Национальной академии наук Беларуси в виде, представленном авторами.

Научные редакторы:

доктор военных наук, кандидат технических наук, доцент С. В. Кругликов,
кандидат технических наук, доцент Р. Б. Григянец,
кандидат технических наук, доцент В. Н. Венгеров

ISBN 978-985-7198-18-4

УДК 811.161.3’322.6:004.85

БЕЛАРУСКАМОЎНЫ СІНТЭЗ МАЎЛЕННЯ ПА ТЭКСЦЕ: ПАДЫХОД НА АСНОВЕ ГЛЫБОКАГА НАВУЧАННЯ

Я. С. Зяноўка, Д. А. Бяляўскі, Д. І. Латышэвіч, М. В. Супрунчук, Ю. С. Гецэвіч
Аб’яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск

Прадстаўлены прататып сістэмы сінтэзу маўлення на беларускай мове, заснаваны на глыбокім машынным навучанні нейронных сетак. Падрабязна апісана архітэктура мадэлі VITS, якая пакладзена ў аснову сінтэзатара маўлення. Разгледжаны бібліятэкі для навучання мадэлі і прыкладныя сістэмы, якія выкарыстоўваюць беларускамоўны сінтэзатар маўлення.

Уводзіны

Сістэмы сінтэзу маўлення (*Text-to-Speech, TTS*) – гэта тэхналогія, якая пераўтварае пісьмовы тэкст у вусную мову. Яна выкарыстоўвае алгарытмы для мадэлявання працэсу маўлення чалавека, дазваляючы ажыццяўляць працэс камунікацыі «чалавек-машина». Традыцыйныя метады сінтэзу маўлення, заснаваныя на правілах, часта гучаць ненатуральна і нягнутка. З хуткім развіццём штучнага інтэлекту пачаўся пошук новых метадаў і тэхналогій распрацоўкі падобных сістэм. Нейрасеткавыя мадэлі і іх глыбокае навучанне сталі дамінуючым метадам у TTS, дасягаючы якасці, блізкай да чалавечага маўлення. Гэта адбываецца па шэрагу прычын. Нейрасеткі генеруюць больш натуральны і рэалістычны голас, паўтараючы чалавечыя інтанацыі, націск і прасодыю.

Шырокі спектр галасоў з рознымі акцэнтамі, тонамі і стылямі дазваляе ствараць мультыгаласавыя сістэмы без прывязкі да аднаго дыктара. Нейрасеткі лёгка адаптуюцца да новых галасоў і моў, што спрашчае стварэнне мованезалежных сінтэзатараў маўлення. Апрацоўка вялікіх аб’ёмаў даных спрыяе аўтаматызацыі працэса распрацоўкі TTS, скарачаючы час і намаганні. Усё гэта спрыяе рэалізацыі беларускамоўнага сінтэзатара маўлення новага пакалення, які распрацаваны ў лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі (https://ssrlab.by/).

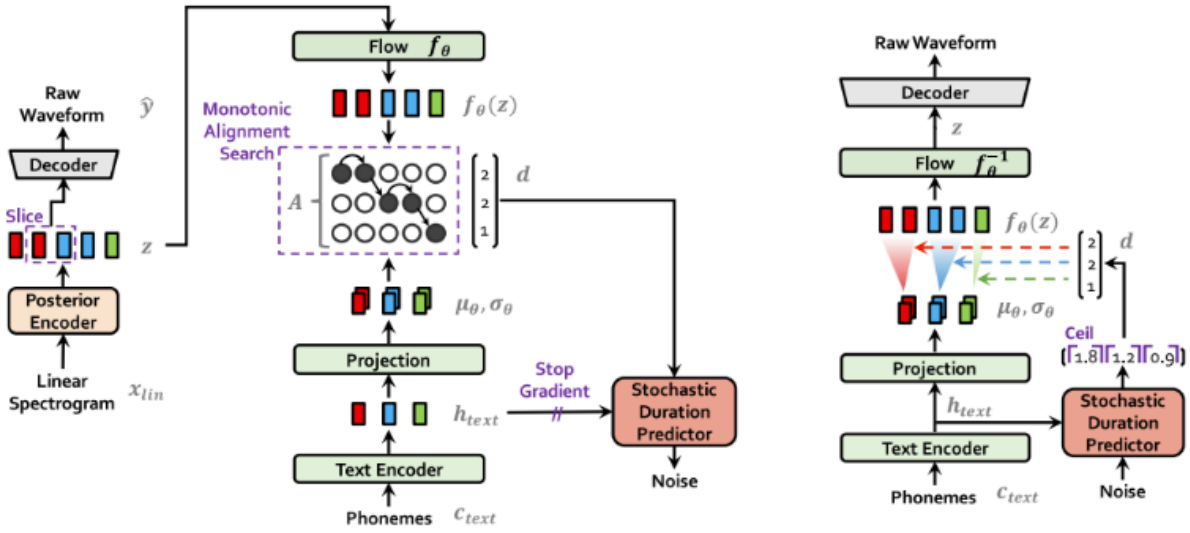
1. Тэхналогія TTS, заснаваная на мадэлі VITS

У якасці мадэлі для сінтэзу маўлення на беларускай мове была абрана *end-to-end* мадэль *VITS* (*Variational Inference with adversarial learning for Text-to-Speech*) – *варыятыўны вывад са спаборным навучаннем для пераўтварэння тэксту ў маўленне*. Мадэль уяўляе сабой аднаступеньчатую неаўтарэгрэсіўную мадэль, здольную генераваць больш натуральны гук у параўнанні з існуючымі двухступенчатымі мадэлямі, такімі як Tacotron 2, Transformer TTS ці нават Glow-TTS. Выкарыстоўваючы варыятыўную аснову, VITS мадэлюе латэнтную прастору характарыстык маўлення, адлюстроўваючы ўласцівую зменлівасць і нявызначанасць пры генераванні маўлення. Гэта дазваляе мадэлі вывучаць латэнтнае ўяўленне, якое эфектыўна кадуе асноўныя характарыстыкі ўваходнага тэксту, дазваляючы генераваць адпаведнае маўленне.

Уключэнне спаборнага навучання ў VITS яшчэ больш удасканальвае працэс сінтэзу. Спаборніцкае навучанне ўключае навучанне сеткі дыскрымінатара для адрознення рэальнай і сінтэзаванай гаворкі, а сетка генератара імкнецца генераваць маўленне, якое паспяхова падманвае дыскрымінатара. Такая спаборнасць узаемадзеяння дапамагае палепшыць агульную якасць і рэалістычнасць сінтэзаваных узораў маўлення. VITS служыць аўтаномным рашэннем для сінтэзу тэксту ў маўленне, паколькі

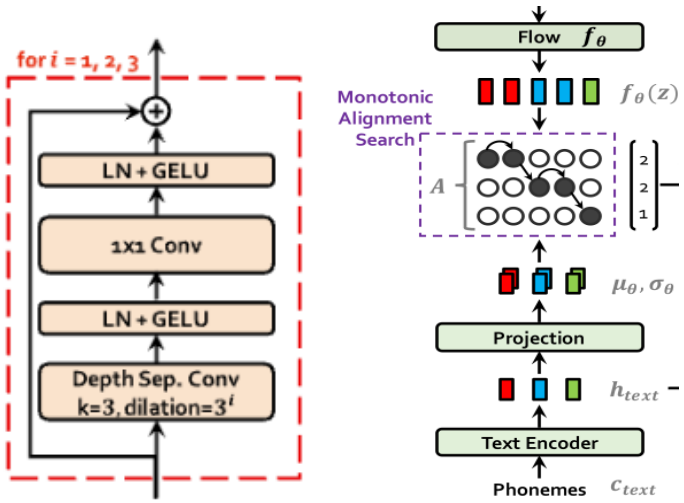
не патрабуе асобнага вакодэра. Здольнасць мадэлі спалучаць метады варыяцыйнага вываду і спаборнасці навучання спрыяе атрыманню высокаякаснага і выразнага сінтэзаванага маўлення.

Агульная архітэктура VITS прадстаўлена на мал. 1. Мадэль VITS складаецца з кодэра Posterior, кодэра Prior, дэкодэра Decoder і стахастычнага прадказальніка працягласці. Модулі Posterior Encoder і Decoder Discriminator выкарыстоўваюцца толькі падчас навучання. Для Posterior Encoder выкарыстоўваецца 16 рэшткавых блокаў WaveNet, якія складаюцца з слаёў пашыраных скрутак з блокам актывацыі і пропускам сувязі. Задні энкодэр прымае спектраграмы лагарыфмічнай велічыні ў лінейным маштабе xlin у якасці ўваходных даных і вырабляе латэнтныя зменныя z з 192 каналамі.



Мал. 1. Агульная архітэктура VITS

Архітэктура Posterior Encoder адлюстравана на мал. 2. Ідэя Posterior Encoder заключаецца ў перакладзе аўдыяданых з прасторы mel-спектраграм у прастору нармальнага размеркавання. Менавіта таму ў даследаванні выкарыстоўваецца лінейны пласт па-над Posterior Encoder для атрымання сярэдняга значэння і дысперсіі нармальнага апастэрыёрнага размеркавання.



Мал. 2. Архітэктуры VITS Posterior Encoder (злева) і Prior Encoder (справа)

Prior Encoder складаецца з розных кампанентаў, такіх як Text Encoder, Projection Layer, Normalizing Flow, і выкарыстоўвае Monotonic Alignment Search (MSA) (мал. 2). Як і Posterior Encoder, Pro Encoder накіраваны на адлюстраванне тэкставых даных з прасторы фанем у прастору нармальнага размеркавання.

Дэкодэр па сутнасці з'яўляецца генератарам HiFi-GAN V1. Архітэктура HiFi-GAN характарызуецца высокай якасцю гуку (генеруе аўдыя з высокай дакладнасцю і рэалістычнасцю, практычна неадрознае ад рэальнага), стабільнай трэніроўкай, мінімізуючы рызыку ўзнікнення праблем, звязаных з навучаннем GAN. Акрамя таго, HiFi-GAN выкарыстоўвае рэкурсіўныя блокі і шматэтапную апрацоўку для аптымізацыі вылічальных рэсурсаў.

Падчас навучання функцыі агульных страт VITS можа быць выказана як камбінацыя пяці розных функцый страт, як адлюстравана ў наступнай формуле:

$$L_{vits} = L_{recon} + L_{kl} + L_{dur} + L_{adv}(G) + L_{fm}(G).$$

Reconstruction Loss L_{recon} – гэта страта $L1$ паміж прадказанай mel-спектраграмай і мэтавай mel-спектраграмай:

$$L_{Adv}(D;G) = \mathrm{E}_{(x,s)}\left[(D(x)-1)^2 + (D\big(G(s)\big))^2\right].$$

KL-Divergence Loss L_{kl}, увогуле, KL – гэта дывергенцыя, якая вымярае, наколькі супадаюць два розных размеркаванні:

$$L_{kl} = \log q_{\phi}(z|x_{lin}) - \log\ p_{\theta}\,(z|[c_{text,}\,A]).$$

Duration Loss L_{dur} – адмоўная варыяцыйная ніжняя мяжа, дзе вектар – прадказальнік працягласці, ў варыяцыйна квантаваны на два вектары аднолькавай памернасці u і v:

$$L_{dur} = -\mathrm{E}_{q_{\phi}(u,v|d,\,c_{text})}\left[\log \frac{p_{\theta}(d-u,v|c_{text})}{q_{\phi}(u,v|d,\,c_{text})}\right].$$

Adversarial Loss $L_{adv}(G)$ – гэта страта найменшых квадратаў паміж выходнай формай сігналу, генераванай дэкодэрам *G* і праўдзівай формай сігналу *y*:

$$L_{adv}(G) = \mathrm{E}_{z}\left[(D\big(G(z)\big) - 1)^2\right].$$

Feature Matching Loss $L_{fm}(G)$ – гэта сярэдняя страта рэканструкцыі дыскрымінатарам *D* схаваных прыкмет праўдзівай формы сігналу *y* і згенераванай дэкодэрам *D* для кожнага пласта l з *T* усіх слаёў, ведаючы, што N_l – гэта агульная колькасць прыкмет у пласце l:

$$L_{fm}(G) = \mathrm{E}_{(y,z)}\left[\sum_{l=1}^{T}\frac{1}{N_l}\left\|D^l(y) - D^l\big(G(z)\big)\right\|_1\right] L_{adv}(G) = \mathrm{E}_{z}\left[(D\big(G(z)\big) - 1^2\right].$$

2. Навучанне мадэлі TTS

Для распрацоўкі TTS на беларускай мове істотным крокам з'яўляецца збор даных для навучання сістэмы. Улічваючы абмежаваную даступнасць спецыялізаваных датасэтаў для беларускай мовы, у якасці асноўнай крыніцы быў абраны датасэт Mozilla CommonVoice. Ён уяўляе сабой вялікую калекцыю галасавых запісаў, сабраных ад добраахвотных удзельнікаў, якія начытваюць сказы на розных мовах, уключаючы беларускую. Датасэт даступны для свабоднага выкарыстання і распаўсюджваецца з адкрытай ліцэнзіяй, што робіць яго каштоўным рэсурсам для беларускамоўнага сінтэзу маўлення.

Недахопам датасэта ад CommonVoice з'яўляецца той факт, што ён прызначаны для задач распазнання маўлення. У сувязі з гэтым, запісы ў датасэце могуць быць непрафесійнымі і ўтрымліваць розныя артэфакты, шумы або недахопы, якія негатыўна ўплываюць на якасць сінтэзу маўлення. У працэсе збору даных з CommonVoice было праведзена папярэдняе фільтраванне і адбор запісаў дастатковай працягласці і разнастайнасці розных маўленчых фанетычных адзінак, такіх як фанемы, словы і фразы. На наступным кроку праведзены аналіз абраных даных для атрымання статыстычнай інфармацыі і разумення асаблівасцяў беларускай мовы ў кантэксце сінтэзу маўлення. Аналіз уключаў ацэнку размеркавання фанетычных адзінак, працягласці фраз і іншых характарыстык, якія могуць аказаць уплыў на якасць і натуральнасць сінтэзаванага маўлення.

Мадэль VITS была навучана з дапамогай бібліятэкі Coqui TTS, папулярнага набору інструментаў з адкрытым зыходным кодам для сінтэзу тэксту ў маўленне. Coqui TTS прадастаўляе поўны набор інструментаў і ўтыліт для навучання і разгортвання мадэляў TTS. Галоўныя асаблівасці бібліятэкі – генерацыя рэалістычнага і высока-якаснага маўлення, магчымасць налады галасоў для дасягнення пажаданага тэмбру, танальнасці і іншых параметраў, падтрымка розных фарматаў аўдыявыхаду, такія як WAV, MP3 і OGG, лёгкая інтэграцыя ў розныя прыкладанні і сістэмы. У працэсе навучання мадэль VITS выкарыстала рэгістратар Weights and Biases. Гэта платформа для адсочвання і візуалізацыі эксперыментаў машыннага навучання. Яна дазваляе даследчыкам і распрацоўшчыкам рэгістраваць і адсочваць ход навучання, метрыкі і прадукцыйнасць мадэлі ў рэжыме рэальнага часу.

Выкарыстанне Coqui TTS і рэгістратара Weights and Biases спрыяла эфектыўнаму правядзенню эксперыментаў, аптымізацыі мадэлі і маніторынгу прадукцыйнасці на працягу ўсяго працэсу навучання. Для навучання быў выкарыстаны сервер з відэакартай Nvidia RTX4090. Аптымізацыя параметраў адбывалася з дапамогай алгарытма AdamW – выпраўленай версіі папулярнага алгарытма аптымізацыі Adam. Памер батча для навучання і ацэнкі якасці мадэлі быў роўны 74, час навучання мадэлі склаў 72 гадзіны.

Заключэнне

Артыкул прысвечаны распрацоўцы беларускамоўнай сістэмы сінтэзу маўлення новага пакалення. Сістэма заснавана на мадэлі нейронных сетак VITS з прымяненнем глыбокага машыннага навучання. Для навучання мадэлі абрана база даных, скампіляваная з Mozilla CommonVoice (https://commonvoice.mozilla.org/be). Гэта вялікая бібліятэка аўдыязапісаў на розных мовах. Для беларускай мовы на сённяшні дзень платформа налічвае 1815 гадзін запісу з 8400 рознымі галасамі. Мадэль навучана з дапамогай Coqui TTS і рэгістратара Weights and Biases. Ацэнка якасці мадэлі роўная 74, час навучання мадэлі склаў 72 гадзіны.

Рэалізацыя мадэлі прадстаўлена ў выглядзе асобнага сэрвісу «Сінтэзатар беларускага маўлення па тэксце РУ», які знаходзіцца ў адкрытым доступе на платформе для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by (https://corpus.by/TextToSpeechBelarusianPy/?lang=be). Прататып характарызуецца простым і зразумелым інтэрфейсам, агучваннем тэкстаў без абмежавання па колькасці слоў і магчымасцю спампавання канвертаванага аўдыяфайла. Галоўнай перавагай сістэмы з'яўляецца якаснае, зразумелае штучнае маўленне з выразнай інтанацыяй і захаваннем прасадычных асаблівасцей беларускай літаратурнай мовы.

Сінтэзатар маўлення таксама ўбудаваны ў галасавыя беларускамоўныя AI-асістэнты на платформе з пытальна-адказнымі сістэмамі, з якімі можна паразмаўляць голасам і тэкстам (https://asistent.io/). Кожная пытальна-адказная сістэма пабудавана з выкарыстаннем тэхналогій сінтэзу і распазнавання маўлення, машыннага перакладу і дыялогавых сістэм. Ажыццяўленне хуткага і якаснага адказу асістэнтамі абумоўлена прымяненнем апісанай TTS. На бягучы момант галасавыя асістэнты даступны ў фарматах Telegram-ботаў, Web-версіі, мабільных прыкладанняў для сістэм iOS і Android. На наступным этапе даследавання плануецца распрацоўка мультыгаласавога сінтэзатара маўлення з агучваннем тэкставай інфармацыі жаночымі і мужчынскімі галасамі.

Даследаванне выконваецца ў рамках праекта БРФФД «Распрацоўка і аптымізацыя мультыгаласавой сістэмы сінтэзу маўлення для беларускай мовы» па дамове №Ф24-061 ад 2 мая 2024 г.