УДК 004.934.5:004.85.032.26

Зяноўка Яўгенія Сяргееўна
малодшы навуковы супрацоўнік
АІПІ НАН Беларусі
г. Мінск, Беларусь

Yauheniya Zianouka
Junior Researcher
UIIP of NASB
Minsk, Belarus
evgeniakacan@gmail.com

Супрунчук Мікіта Віктаравіч
кандыдат філалагічных навук, старшы выкладчык
кафедры тэарэтычнага і славянскага мовазнаўства
МДЛУ
г. Мінск, Беларусь

Mikita Suprunchuk
PhD in Philology,
Senior lecturer
of the Departments
of Theoretical and Slavic Linguistics
Minsk State Linguistic University
Minsk, Belarus
msuprunch@gmail.com

Латышэвіч Давід Іосіфавіч
малодшы навуковы супрацоўнік
АІПІ НАН Беларусі
г. Мінск, Беларусь

David Latyshevich
Junior Researcher
UIIP of NASB
Minsk, Belarus
david.latyshevich@gmail.com

Гецэвіч Юрась Станіслававіч
кандыдат тэхнічных навук,
загадчык лабараторыі распазнавання і сінтэзу маўлення
АІПІ НАН Беларусі
г. Мінск, Беларусь

Yuras Hetsevich
PhD in Technical Sciences,
Head of the Speech Synthesis and Recognition Laboratory
UIIP of NASB
Minsk, Belarus
yuras.hetsevich@gmail.com

СУЧАСНЫЯ ПАДЫХОДЫ ДА РАСПРАЦОЎКІ МУЛЬТЫГАЛАСАВЫХ СІНТЭЗАТАРАЎ МАЎЛЕННЯ НА АСНОВЕ ГЛЫБОКАГА МАШЫННАГА НАВУЧАННЯ

У артыкуле апісваюцца сістэмы сінтэзу маўлення па тэксце як інструмент пераўтварэння тэкставай інфармацыі ў галасавое паведамленне. Абгрунтавана мэтанакіраванасць пошуку новых метадаў і алгарытмаў іх рэалізацыі. Прыведзены сучасныя падыходы да

распрацоўкі мультыгаласавых сінтэзатараў маўлення. Прадстаўлены актуальныя метады распрацоўкі сінтэзу маўлення на аснове глыбокага машыннага навучання нейронных сетак.

К л ю ч а в ы я с л о в ы: *мультыгаласавыя сістэмы сінтэзу маўлення; галасавыя тэхналогіі; камп'ютарная апрацоўка натуральнай мовы; нейронныя сеткі; глыбокае машыннае навучанне.*

MODERN APPROACHES TO THE DEVELOPMENT OF MULTI-VOICE SPEECH SYNTHESIZERS BASED ON DEEP MACHINE LEARNING

The article describes text-to-speech synthesis systems as a tool for converting text information into a voice message. The focus of the search for new methods and algorithms for their implementation is presented. Modern approaches to the development of multi-voice speech synthesizers are depicted. The current methods of developing speech synthesizers based on deep machine learning of neural networks are presented.

K e y w o r d s: *multi-voice text-to-speech systems; voice technologies; natural language processing; neural networks; deep machine learning.*

Сістэмы сінтэзу маўлення па тэксце (*ССМТ, eng. Text-to-speech technology – TTS*) – гэта тэхналогія, якая пераўтварае пісьмовы тэкст у вуснае маўленне [1, p. 7]. У апошнія гады яна зрабіла велізарны крок наперад, стаўшы неад'емнай часткай паўсядзённага жыцця. Ад галасавых памочнікаў, такіх як *Siri, Alexa, Google Assistant, Cortana, Alica, Маруся* і інш., да электронных кніг і навігатараў – сінтэз маўлення дапамагае карыстальнікам узаемадзейнічаць з тэхналогіямі больш натуральна і інтуітыўна. ССМТ дэманструюць шырокі спектр прымянення ў розных галінах і сектарах эканомікі, адлюстроўваючы шматлікія перавагі і магчымасці для павышэння эфектыўнасці, даступнасці і інавацый. Падобныя сістэмы – гэта ўніверсальны інструмент, які прапануе каштоўныя функцыянальныя магчымасці для камунікацыі, забеспячэння даступнасці, адукацыі, абслугоўвання кліентаў, моўных служб, дапаможных тэхналогій і аўтамабільных дадаткаў. Выкарыстоўваючы магчымасці TTS, арганізацыі і прыватныя асобы могуць прымяняць іх для павышэння прадукцыйнасці, уцягнутасці, інклюзіўнасці і інавацый у самых розных кантэкстах і сцэнарыях. Гэта і ўказвае на іх запатрабаванасць у сучасным лічбавым асяроддзі.

Існуюць разнастайныя падыходы да распрацоўкі TTS. Да асноўных метадаў іх генерацыі адносяцца *артыкуляцыйная, канкатэнатыўная (кампіляцыйная), параметрычная (фармантная) мадэлі і мадэль сінтэзу на аснове глыбокага навучання* [2, с. 54]. Першыя тры мадэлі апісваюць класічны падыход да распрацоўкі ССМТ і маюць шэраг недахопаў. Яны часта апрацоўваюць тэкст у маўленне манатонна і з адсутнасцю натуральнай інтанацыі. Звычайна падобныя сістэмы патрабуюць значных намаганняў для адаптацыі да розных галасоў, моў або стыляў маўлення. Акрамя таго, складанасць у кіраванні прасодыяй прыводзіць да цяжкасцей з кіраваннем прасадычнымі характарыстыкамі маўлення, такімі як націск, тэмп і інтанацыя. Класічныя падыходы звычайна патрабуюць складаных алгарытмаў і вылічальна інтэн-

сіўных аперацый, што можа быць рэсурсаёмістым. Яны могуць быць павольнымі ў апрацоўцы і генерацыі маўлення, асабліва на прыладах з абмежаванымі рэсурсамі.

Усе гэтыя пытанні вырашаюцца ў новым метадзе распрацоўкі ССМТ – *сінтэзе маўлення па тэксце на аснове глыбокага навучання нейрасетак* [3]. Ён з'яўляецца адным з самых перадавых метадаў у галіне штучнага інтэлекту і апрацоўкі натуральнай мовы. Нейронная сетка – гэта асобная матэматычная мадэль, якая мае мноства параметраў і спрабуе вырашыць пэўную задачу – класіфікацыйную або рэгрэсіўную. Адным з пераваг выкарыстання нейронных сетак для сінтэзу маўлення з'яўляецца навучанне генерацыі больш натуральнага маўлення з перадачай чалавечай інтанацыі і акцэнтаў. Акрамя таго, нейронныя сеткі можна навучаць на шырокім спектры ўваходных даных, уключаючы тэкст на некалькіх мовах і з рознымі акцэнтамі, што забяспечвае больш надзейны і гнуткі сінтэз маўлення.

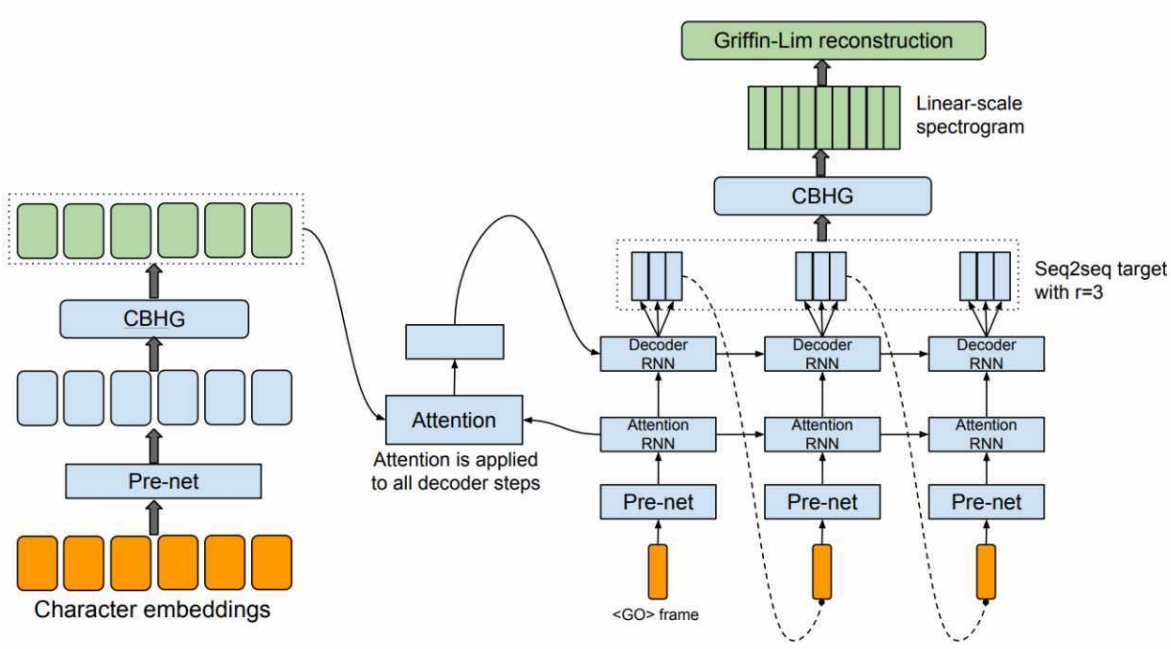
Сёння існуюць чатыры асноўныя віды нейрасетак: паўназвязныя, скруткавыя, рэкурэнтныя і трансформерныя. Кожны з відаў можа быць прыменены для распрацоўкі TTS. У паўназвязных сетках кожны нейрон у адным пласце злучаны з кожным нейронам у наступным пласце. Гэта дазваляе сетцы навучацца складаным нелінейным залежнасцям у даных. Такія сеткі простыя ў рэалізацыі і могуць навучацца складаным функцыям.

Аднак яны патрабуюць шмат параметраў нягледзячы на функцыю пераднавучання на маленькіх наборах даных [4]. Скруткавыя сеткі (Convolutional Neural Networks, SNN) выкарыстоўваюць звышдакладныя аперацыі для здабывання прыкмет з даных. Яны асабліва эфектыўныя для працы з выявамі, відэа і аўдыя. Перавагай іх прымянення з'яўляецца інварыянтнасць да зруху, кручэння і маштабавання, зніжэнне колькасці параметраў. Рэкурэнтныя сеткі (Recurrent Neural Networks, SNN) апрацоўваюць паслядоўныя даныя, выкарыстоўваючы інфармацыю з папярэдніх крокаў. Гэта дазваляе ім запамінаць кантэкст і будаваць залежнасці паміж элементамі паслядоўнасці. Як правіла, падобныя сеткі прымяняюцца для апрацоўкі натуральнай мовы, тэхналогій машыннага перакладу, распазнавання маўлення. Трансформеры выкарыстоўваюць механізм увагі для вылучэння важных элементаў паслядоўнасці. Яны не абапіраюцца на рэкурсіўныя аперацыі і таму не пакутуюць ад праблемы знікаючага градыента. Такія нейрасеткі больш магутныя, чым RNN, могуць апрацоўваць доўгія паслядоўнасці, пры гэтым могуць быць рэсурсаёмістымі і больш складанымі ў рэалізацыі. Трансформеры – новы напрамак у распрацоўцы нейронных сетак для навучання мадэляў сінтэзу маўлення на вялікіх аб'ёмах даных.

Для стварэння TTS выкарыстоўваюцца розныя тыпы нейронных сетак. Некаторыя з найбольш вядомых відаў архітэктуры, якія ўжываюцца ў мультыгаласавых сістэмах сінтэзу маўлення, уключаюць у сябе *Tacotron, WaveNet, Deep Voice, Transformer TTS, VITS* [5].

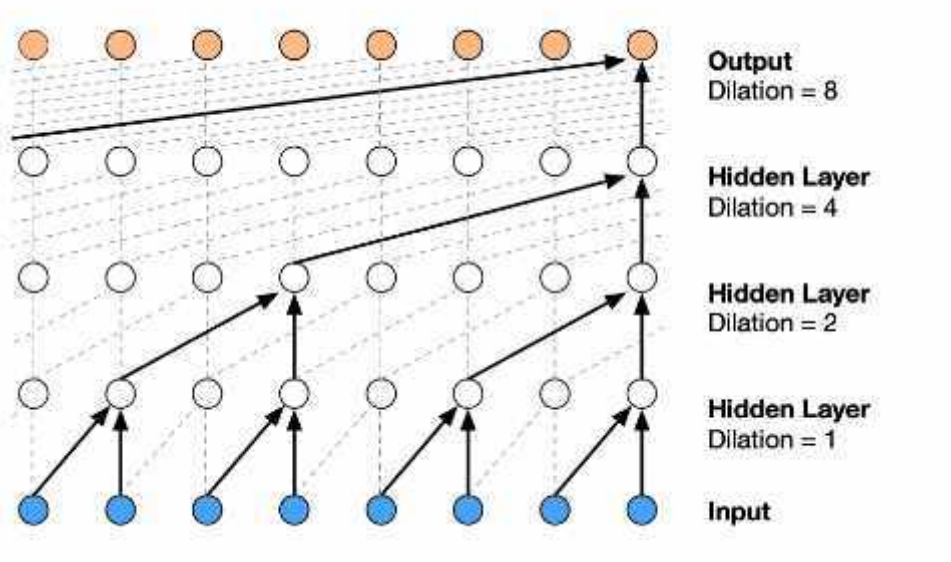
Tacotron – гэта скразная генерацыйная мадэль пераўтварэння тэксту ў маўленне, якая прымае паслядоўнасць сімвалаў у якасці ўваходных даных і выводзіць адпаведную спектраграму. Асновай Tacotron з'яўляецца мадэль *seq2seq*. Яна выкарыстоўвае тэкставы ўвод і непасрэдна генерыруе маўлен-

чыя сігналы, адхіляючы неабходнасць у прамежкавых даных, такіх як фанемы або лінгвістычныя прыкметы. На мал. 1 прадстаўлена схема працы Tacotron, якая ўключае кадзіравальнік, дэкодэр і сетку постапрацоўкі. На высокім узроўні мадэль прымае сімвалы ў якасці ўваходных даных і стварае кадры спектраграмы, якія затым пераўтвараюцца ў сігналы.



Мал. 1. Структура генерацыйнай мадэлі Tacotron

Мадэль *WaveNet* – гэта архітэктура глыбокай нейроннай сеткі, распрацаваная DeepMind, кампаніяй Alphabet Inc. Упершыню яна была прадстаўлена ў 2016 годзе для генерацыі рэалістычных маўленчых сігналаў. WaveNet вядомая сваёй здольнасцю прайграваць высакаякасны гук з натуральным гучаннем, што робіць яе прыдатнай для сінтэзу маўлення па тэксце (мал. 2).



Мал. 2. Структура мадэлі WaveNet

Ключавыя асаблівасці мадэлі WaveNet:

1. Структура аўтарэгрэсіі: WaveNet генерыруе выходныя даныя па адным сэмпле за раз на аснове папярэдніх сэмплаў. Гэта дазваляе фіксаваць доўгатэрміновыя залежнасці ў аўдыяданых.

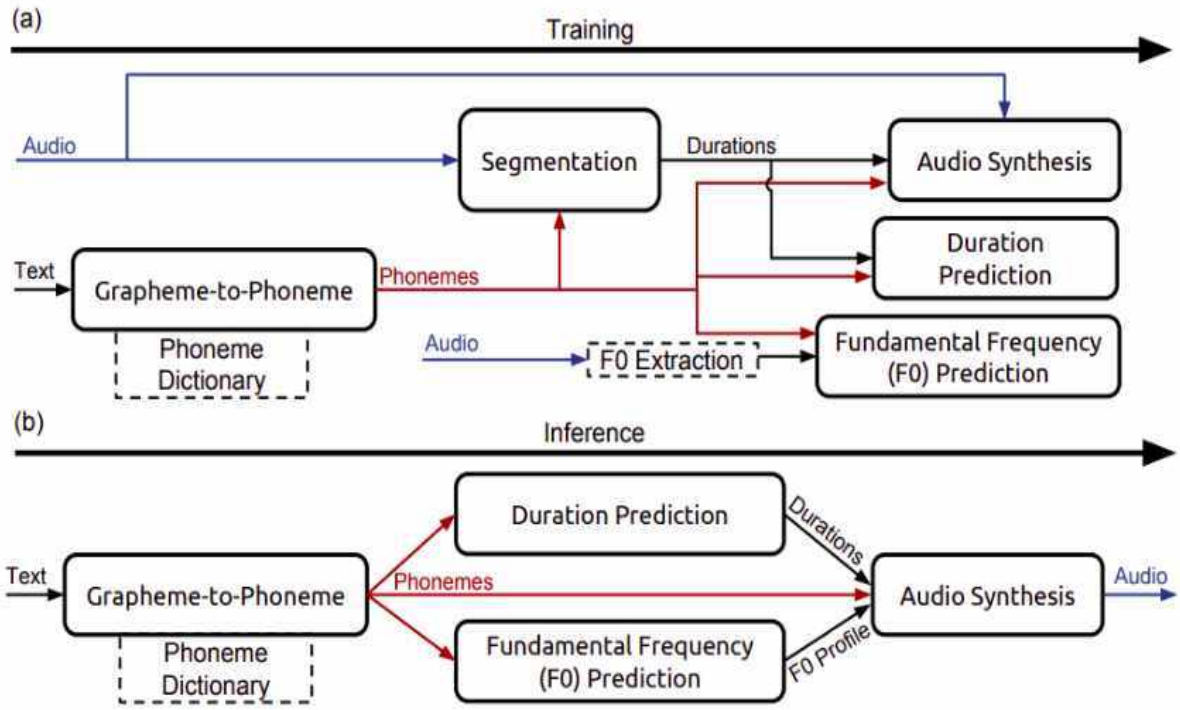
2. Пашыраныя прычынна-выніковыя сувязі, успрымальнасць якіх экспанентна ўзрастае з глыбінёй. Гэта дазваляе мадэлі атрымліваць кантэкстную інфармацыю ў шырокім дыяпазоне і генерыраваць больш рэалістычныя і падрабязныя гукавыя сігналы.

3. Набор слаёў пашыраных віткоў, часта размешчаных у іерархічным парадку. Кожны пласт вучыцца мадэляваць розныя ўзроўні абстракцыі ў аўдыясігнале.

4. Функцыі закрытай актывацыі, аналагічныя тым, якія выкарыстоўваюцца ў сетках з доўгачасовай і кароткачасовай памяццю (LSTM) для кіравання патокам інфармацыі і ліквідацыі праблемы са знікаючым градыентам.

WaveNet паспяхова ўжываецца для вырашэння розных задач генерацыі гуку, уключаючы TTS, стварэнне музыкі і паляпшэнне якасці маўлення. Яе здольнасць генерыраваць высакаякасны і выразны гук зрабіла яе канкурэнтаздольнай сярод астатніх нейрасетак у галіне глыбокага навучання для апрацоўкі гуку.

Deep Voice ад Baidu заклаў аснову для актуальных дасягненняў у галіне скразнога сінтэзу маўлення. Ён складаецца з 4 розных нейронных сетак, якія разам утвараюць скразны канвеер, менавіта (мал. 3):



Мал. 3. Структура мадэлі Deep Voice

1. Мадэль сегментацыі, якая вызначае межы паміж фанемамі. Гэта гібрыд CNN і RNN-сетак, які навучаны прадказваць адпаведнасць паміж галасавымі гукамі і мэтавымі фанемамі, выкарыстоўваючы страты.

2. Мадэль пераўтварэння графемы ў фанемы. Для гэтай задачы была абраная шматслаёвая мадэль кодэра-дэкодэра з GRU (Gated Recurrent Unit), які распрацаваны для апрацоўкі паслядоўных даных, такіх як тэкст, маўленне, часовыя шэрагі і інш.

3. Мадэль для прагназавання працягласці фанем і асноўных частот. Два цалкам падлучаныя пласты, за якімі ідуць два аднанакіраваныя пласты GRU і яшчэ адзін падлучаны пласт (выкананне абедзвюх задач адначасова).

4. Мадэль для сінтэзу канчатковага гуку. WaveNet складаецца з сеткі кандыцыяніравання, якая павышае дыскрэтызацыю лінгвістычных характарыстык да пажаданай частаты, і сеткі аўтарэгрэсіі, якая генерыруе размеркаванне верагоднасці па дыскрэтызіраваных аўдыясэмплах.

Аўтарам таксама ўдалося ажыццявіць вывад даных у рэжыме рэальнага часу, стварыўшы высокааптымізаваныя ядры CPU і GPU для паскарэння вываду. У амерыканскай англійскай ён атрымаў MOS 2,67.

Transformer TTS заснавана на архітэктуры трансформера, першапачаткова прадстаўленай камандай Google Brain у 2017 г. для машыннага перакладу. Мадэлі Transformer TTS прадэманстравалі ўражлівую прадукцыйнасць у стварэнні маўлення з натуральным гучаннем на аснове тэкставага ўводу. У кантэксце TTS мадэль Transformer TTS выкарыстоўвае архітэктуру для пераўтварэння ўваходнага тэксту ў адпаведныя маўленчыя сігналы. Некаторыя ключавыя аспекты мадэлі Transformer TTS прадстаўлены ніжэй:

1. Механізм самакантролю: мадэль Transformer ў значнай ступені абапіраецца на механізмы самакантролю, які дазваляе эфектыўна шукаць доўгатэрміновыя залежнасці ва ўваходным тэксце. Гэты механізм дазваляе мадэлі ацэньваць значнасць кожнага слова ва ўваходнай паслядоўнасці пры стварэнні адпаведных маўленчых характарыстык.

2. Архітэктура кодэр-дэкодэра, у якой кодэр апрацоўвае ўваходныя тэкставыя даныя, а дэкодэр генерыруе маўленчы сігнал на аснове ўяўленняў кодэра.

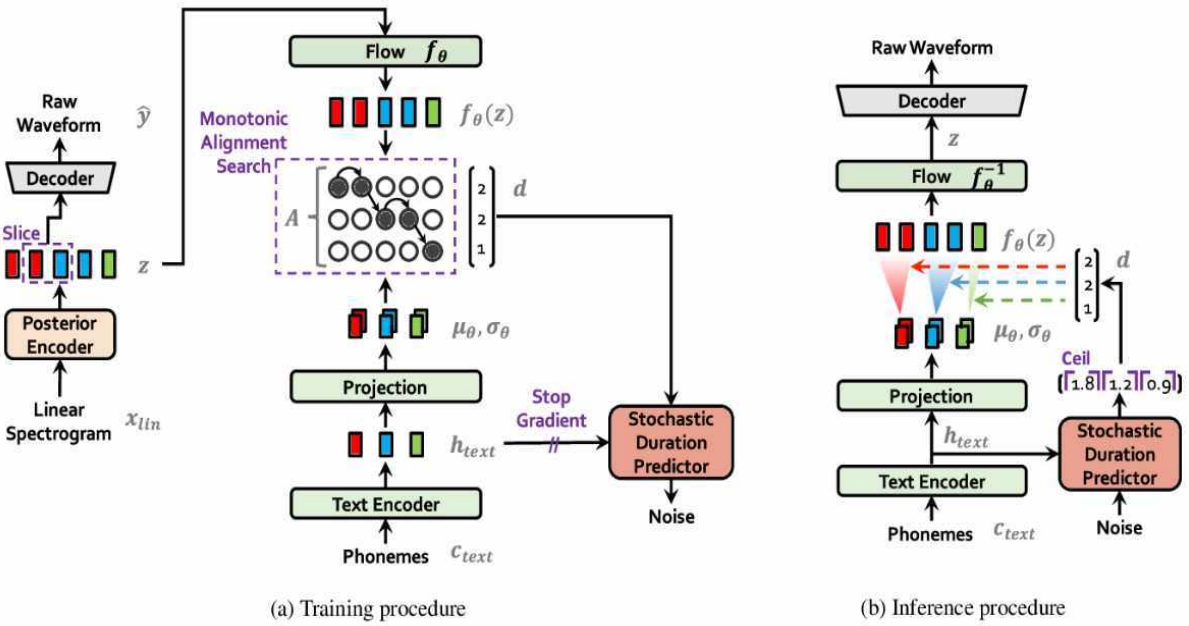
3. Паралельная апрацоўка: у адрозненне ад рэкурэнтных нейронных сетак, якія апрацоўваюць уваходныя даныя паслядоўна, мадэль Transformer можа працаваць паралельна дзякуючы механізму самарэгулявання. Гэта паскарае навучанне і час вываду.

4. Мадэль выкарыстоўвае ўвагу з некалькімі галоўкамі, што дазваляе ёй адначасова апрацоўваць розныя часткі паслядоўнасці ўводу. Гэта функцыя дапамагае мадэлі адсочваць розныя заканамернасці ва ўваходных даных.

Мадэль *VITS* (*Variational Inference with Adversarial Learning for Text-to-Speech*) уяўляе сабой аднаступеньчатую няаўтарэгрэсійную мадэль пераўтварэння тэксту ў маўленне, здольную генерыраваць больш натуральны гук, чым існуючыя двухступеньчатыя мадэлі, такія як Tacotron 2, Transformer TTS ці нават Glow-TTS. Выкарыстоўваючы варыяцыйную аснову, VITS мадэлюе

латэнтную прастору характарыстык маўлення, адлюстроўваючы ўласцівую зменлівасць і нявызначанасць пры генерыраванні маўлення (мал. 4). Наяўнасць спаборнасці навучання ў VITS яшчэ больш удасканальвае працэс сінтэзу. Спаборнае навучанне ўключае ў сябе навучанне сеткі дыскрымінатара для адрознення рэальнай і сінтэзаванай гаворкі, а сетка генератара імкнецца генерыраваць маўленне, якое паспяхова падманвае дыскрымінатара.

Такое спаборнае ўзаемадзеянне дапамагае палепшыць агульную якасць і рэалістычнасць сінтэзаваных узораў гаворкі. VITS служыць аўтаномным рашэннем для сінтэзу тэксту ў маўленне, паколькі не патрабуе асобнага вакодэра. Агульная архітэктура VITS адлюстравана на мал. 4. Яна складаецца з кодэра Posterior, кодэра Prior, дэкодэра Decoder і стахастычнага прадказальніка працягласці. Модулі Posterior Encoder і Decoder Discriminator выкарыстоўваюцца толькі падчас навучання, а не для вываду маўлення. 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. Ідэя Posterior Encoder заключаецца ў перакладзе аўдыяданых з прасторы mel-спектраграм у прастору нармальнага размеркавання. Менавіта таму ў мадэлі выкарыстоўваецца лінейны пласт па-над Posterior Encoder для атрымання сярэдняй дысперсіі нармальнага апастэрыёрнага размеркавання. Prior Encoder складаецца з Text Encoder, Projection Layer, Normalizing Flow і выкарыстоўвае Monotonic Alignment Search (MSA). Як і Posterior Encoder, Pro Encoder накіраваны на адлюстраванне тэкставых даных з прасторы фанем у прастору нармальнага размеркавання.



(a) Training procedure (b) Inference procedure

Мал. 4. Структура мадэлі VITS

Такім чынам, распрацаваныя на аснове нейронных сетак сістэмы сінтэзу маўлення па тэксце характарызуюцца большымі перавагамі ў параўнанні з класічнымі падыходамі. Нейронныя сеткі могуць навучацца на вялікім аб'ёме даных, што дазваляе ім генерыраваць больш натуральнае, выразнае і плаўнае маўленне, блізкае да чалавечага [6]. Нейронныя сеткі здольныя адаптавацца да розных моў, дыялектаў і акцэнтаў, што робіць іх больш універсальнымі і маштабаванымі, асабліва для моў з нізкім спажываннем рэсурсаў. Гэта спрыяе хуткаму навучанню мадэляў на новых мовах або галасах. У цэлым сістэмы сінтэзу маўлення, архітэктура якіх заснавана на нейронных сетках, дазваляюць ствараць больш якасны і рэалістычны аўдыякантэнт, што робіць іх прывабнымі для шырокай сферы прыкладання і задач.

Удасканаленыя алгарытмы сінтэзу маўлення, палепшаная якасць перадачы голасу і пашыраныя магчымасці лінгвістычнага аналізу спрыяюць стварэнню больш рэалістычнага і натуральнага штучнага маўлення, а таксама высокатэхналагічных прадуктаў для асобных моў. Акрамя таго, інтэграцыя метадаў штучнага інтэлекту і машыннага навучання павышаюць магчымасці распрацоўкі мультыгаласавых сістэм, павелічэння іх прадукцыйнасці і адаптыўнасці для розных моў.

Апісанае даследаванне падрыхтавана ў межах праекта па гранце БРФФД, дагавор № Ф24-061 ад 2 мая 2024 г.

ЛІТАРАТУРА

1. Taylor P. Text-to-Speech Synthesis. N. Y. : Cambridge University Press, 2009. 626 p.

2. Лобанов Б. М., Цирульник Л. И Компьютерный синтез и клонирование речи. Минск : Белорусская наука, 2008. 344 с.

3. A Survey on Neural Speech Synthesis [Electronic resource] / Xu Tan, Tao Qin, Frank Soong, Tie-Yan Liu // arXiv preprint arXiv:2106.15561. 2021. URL: https://arxiv.org/abs/2106.15561 (accessed: 12.02.24).

4. Fastspeech: Fast, robust and controllable text to speech / Y. Ren [et al.] //Advances in neural information processing systems. 2019.

5. Transfer learning from speaker verification to multispeaker text-to-speech synthesis / Y. Jia [et al.] //Advances in neural information processing systems. 2018.

6. Hayes B., Saitis C., Fazekas G. Neural waveshaping synthesis [Electronic resource] // arXiv preprint arXiv: 2107.05050. 2021. URL: https://benhayes.net/projects/nws/ (accessed: 12.02.24).