УДК 004.65
DOI https://doi.org/10.32782/2663-5941/2024.5.1/39

Самофалов А.В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Терейковський І.А.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

КОНЦЕПТУАЛЬНА МОДЕЛЬ ФОРМУВАННЯ ВЕБ-ОРІЄНТОВАНИХ БАЗ ДАНИХ ЕМОЦІЙНО ЗАБАРВЛЕНИХ ГОЛОСОВИХ СИГНАЛІВ

Дана стаття розглядає концептуальну модель формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів. Наведена модель досліджує процес створення та надає перелік критеріїв оцінювання ефективності формування таких баз даних. З'ясовано, що розпізнавання емоцій людини в мовленні є одним із найважливіших завдань, яке виконується щодня при взаємодії людей одних з одними. Також на сьогоднішній день стрімко розвиваються робототехніка та застосунки на основі штучного інтелекту, які намагаються зробити взаємодію з користувачами наближеною до взаємодії між людьми по відчуттям. Для досягнення максимально комфортного користувацького досвіду, ці застосунки та робототехнічні засоби повинні бути натренованими на великих кількостях емоційних даних. Розпізнавання емоцій людини за допомогою автоматизованих комп'ютерних засобів є особливо цінним та актуальним для таких додатків, як розпізнавання емоцій в обслуговуванні клієнтів, моніторинг психічного здоров'я та взаємодія людини з комп'ютером, тощо. На початку даної статті розглянуто перелік основних типів баз даних емоційно забарвлених голосових сигналів та визначено виклики й завдання з якими стикаються дослідники під час їх формування. Після цього проведено аналіз деяких підходів для обробки та класифікації емоційних станів. Наостанок, представлена розроблена модель та наведені критерії ефективності формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів, що дозволяють спростити процес створення емоційних аудіоданих для навчання нейронних мереж, оскільки виявлено, що цей процес може контролюватися відповідно до попередньо заданих вимог, використовуючи наведені числові метрики. Завдяки цьому взаємодія між людиною і робототехнічними засобами та застосунками з штучним інтелектом може бути суттєво покращена.

Ключові слова: *концептуальна модель, бази даних, нейронні мережі, емоція, критерії ефективності.*

Постановка проблеми. Розпізнавання емоцій людини в мовленні є однією із найважливіших завдань з яким ми стикаємось кожен день коли взаємодіємо один з одним. Також у наш час стрімко розвивається робототехніка та застосунки на основі штучного інтелекту, які намагаються зробити взаємодію з користувачами максимально наближеною до взаємодії між людьми по відчуттям [1]. Для досягнення цієї мети ці застосунки та робототехнічні засоби мусять бути натренованими на великих кількостях емоційних даних. Однак створення та обробка великої кількості емоційно забарвлених голосових сигналів людини є операцією, яка вимагає великих затрат по часу та людських ресурсів. Тому формалізація процесу формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів є суттєвою задачею, вирішення якої дозволить створювати емоційні аудіодані для навчання нейронних мереж у більших кількостях, ніж вдається наразі.

Аналіз останніх досліджень і публікацій. Існують різні бази даних емоційних голосових сигналів, що збираються з використанням різних методів, щоб забезпечити широкий спектр емоцій та природність. При цьому використовуються такі поширені підходи:

1. Викликана мова – у цьому методі учасники поміщаються у контрольовані середовища та їм пропонується виконати завдання, призначені для виклику певних емоцій. Наприклад, вони можуть слухати історію, дивитися відео чи брати участь у спрямованих обговореннях, щоб викликати такі емоції, як щастя, сум, гнів чи подив [2].

2. Награна мова – для цього часто використовуються професійні актори, щоб імітувати різні емоційні стани. Їм дають сценарії та просять вимовляти репліки з певними емоційними інтона-

ціями. Цей метод використовується у таких базах даних, як CREMA-D, AESDD та інших [3].

3. Природна мова – деякі бази даних спрямовані на захоплення спонтанного емоційного мовлення в реальних умовах. Це може включати запис розмов, інтерв'ю чи інших взаємодій, у яких емоції виникають природним чином. Потім ці записи анотуються для емоційного змісту.

4. Краудсорсингові дані – платформи, такі як Amazon Mechanical Turk, або інші краудсорсингові сервіси використовуються для збору даних про емоційне мовлення у різноманітної групи учасників. Цей метод допомагає швидко збирати великий обсяг даних із різних демографічних групп.

5. Мультимодальний збір даних – деякі бази даних, такі як CMU-MOSEI, включають не лише аудіо, а й відео та текстові дані. Цей мультимодальний підхід забезпечує більш багатий контекст для розуміння емоцій, оскільки він фіксує вирази обличчя, мову тіла та вимовлені слова [1].

Наведені підходи дозволяють зробити бази даних всеосяжними та корисними для різних дослідницьких додатків, включаючи перетворення емоційного голосу, аналіз настроїв та розпізнавання емоцій.

Розглянемо далі підхід природного мовлення, що фокусується на фіксації спонтанних емоційних виразів у реальних умовах. У цьому методі дослідники записують розмови, інтерв'ю, публічні виступи чи інші взаємодії, у яких емоції виникають природним чином. Однак збір даних природного мовлення включає кілька етичних міркувань [4]. По-перше, учасники повинні бути поінформовані про мету запису та надати свою усвідомлена згоду. По-друге, дослідники повинні запровадити конфіденційність, тобто забезпечити захист особистої інформації та відповідально використати отримані записи. Та по-третє, увесь процес збору даних мусить бути максимально прозорим, а саме учасники повинні знати, як використовуватимуться їх дані, та мати можливість відкликати свою згоду у будь-який час.

Після того, як перелічені етичні міркування враховані, а записи зібрані, вони проходять докладний процес анотування. Найважливіший крок, це маркування емоцій: навчені анотатори слухають записи та маркують сегменти відповідними емоціями (наприклад, щастя, смуток, гнів). На додачу до цього, анотатори можуть також відзначати контекстні фактори, які можуть впливати на емоційне вираження, такі як тема розмови чи стосунки між тими, хто розмовляє [5].

Для аналізу емоцій у природному мовленні з використанням автоматизованих систем існують певні переліки характеристик для визначення емоційного стану розмовляючого. Ці характеристики допомагають ідентифікувати такі емоції, як, наприклад, щастя, смуток, гнів та страх. Системи розпізнавання емоцій аналізують різні голосові характеристики, зокрема просодія (ритм, наголос та інтонація мови), висота голосу, тон (якість чи характер голосу), темп (швидкість, з якою людина розмовляє) та гучність мовлення.

Розширені моделі машинного навчання також використовуються для обробки та класифікації емоційних станів [6]. Одними з поширених підходів є:

– Згорткові нейронні мережі (CNN): ефективні для отримання ознак з аудіоспектрограм.

– Рекурентні нейронні мережі (RNN): корисні для фіксації тимчасових залежностей у мовленні.

– Моделі Transformer: стають все більш популярними завдяки своїй здатності обробляти довгострокові залежності та складні шаблони.

Звичайно, як і всі інші методи, підхід розпізнавання емоцій у природному мовленні має свої переваги та деякі проблеми. Почнемо з переваг, а саме:

1. Справжність: дані природного мовлення є більш репрезентативними для реальних емоційних виразів у порівнянні з постановочною або викликаною мовою.

2. Багатий контекст: фіксація спонтанних взаємодій забезпечує більш багатий контекст для розуміння емоцій, включаючи невербальні сигнали та ситуативні фактори.

Далі наведемо деякі проблеми при використанні цього підходу:

1. Мінливість: природна мова може бути дуже мінливою, що ускладнює її стандартизацію та аналіз [4] [7].

2. Шум: записи реального світу часто містить фоновий шум, що може ускладнити аналіз [8].

3. Складність анотації: точне маркування емоцій у природній мові потребує кваліфікованих анотаторів і може займати багато часу [9].

4. Культурні відмінності: емоції можуть виражатися по-різному в різних культурах, що потребує адаптивності моделей. Розпізнавання емоцій з використанням природного мовлення – це область, що швидко розвивається, зі значним потенціалом для поліпшення взаємодії людини з комп'ютером і різних інших додатків [6].

Розпізнавання емоцій людини за допомогою автоматизованих комп’ютерних засобів є особливо цінним для таких додатків, як розпізнавання

емоцій в обслуговуванні клієнтів (покращення взаємодії з клієнтами шляхом виявлення розчарування чи задоволення) [4], моніторинг психічного здоров'я (моніторинг емоційного благополуччя та виявлення ознак депресії чи тривоги) та взаємодія людини з комп'ютером (покращення чуйності та емпатії віртуальних помічників та роботів), тощо. У цих сферах розуміння природних емоційних виразів людини має вирішальне значення для покращення досвіду людини від взаємодії з цими системами, та, власне, покращення самих систем на основі зворотного зв'язку.

Постановка завдання. Метою статті є розробка концептуальної моделі формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів, яка формалізує процес створення таких баз даних та оцінювання ефективності їх формування.

Виклад основного матеріалу. Перед тим як навести концептуальну модель формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів, спочатку дамо визначення деяким поняттям, оскільки їх визначення варіюються між авторами та джерелами.

– Термін «голос» може мати кілька значень, залежно від контексту, але ми візьмемо фізіологічний – звук, що виробляється людьми та іншими хребетними за допомогою легких та голосових зв'язок у гортані.

– Голосовий сигнал відноситься до електричного представлення звукових хвиль, що виробляються людським голосом. До деяких ключових моментів про голосові сигнали відносяться висота тону, тон і гучність.

– Мовний сигнал зазвичай відноситься до подання усного або письмового мовлення у формі, яку можуть обробляти машини. Мовні сигнали можна аналізувати за різними ознаками, включаючи фонетичні, просодичні, синтаксичні та семантичні ознаки.

– Набір даних (dataset) – це набір даних, зазвичай організований у структурованому форматі, який використовується для аналізу, дослідження чи навчання моделей машинного навчання.

– База даних (database) – це організований набір даних, які зберігаються та доступні в електронному вигляді. Набір даних та база даних – це пов'язані поняття, але вони служать різним цілям і мають різні характеристики. Набори даних зазвичай використовуються для певних аналізів або завдань, в той час як бази даних використовуються для постійного зберігання та керування даними.

– Емоція – це складний психологічний стан, який включає три окремих компоненти: суб'єктивний досвід, фізіологічну реакцію і поведінкову або експресивну реакцію.

– Емоційний стан відноситься до поточного стану емоцій людини у певний момент. Він охоплює почуття та настрій, які відчуваються, та на які можуть впливати різні фактори, такі як оточення, думки, фізичний стан та взаємодія з іншими людьми.

– Емоції можна загалом розділити на базові та складні. Ось перелік базових емоцій [2], які універсальні і відчуваються людьми у різних культурах: щастя, гнів, відраза, страх, сум і здивування.

Тепер перейдемо безпосередньо до концептуальної моделі формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів. У самому загальному вигляді ідея формування таких баз даних полягає в наступному: необхідно подати відеофайл певного типу на вхід програми з розпізнавання емоцій по обличчях людини. Після цього, певні кадри будуть помічені як ті, що містять певні наперед визначенні емоції. Потім, необхідно виділити аудіофрагменти у ці моменти часу, які відповідають класифікованим емоціям.

Модель формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів наведена на рис. 1.

Рис. 1. Модель формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів

Розглянемо кожен етап цієї модель більш детально. По-перше, як вже було зазначено, нам необхідно підібрати відеофайли особливого типу, а саме ті, в яких будуть присутні обличчя людини в таких положеннях, які дозволять програмі по визначенню емоцій розпізнавати ці певні емоції. Ці відеофайли також повинні містить голосові сигнали в достатній кількості, щоб можна було отримати емоційно забарвлені аудіофрагменти. Що не менш важливо, голосові сигнали людей у цих відеофайлах не повинні накладатися один на одного, тобто у кожен момент часу говорити повинна тільки одна особа. Після цього варто впевнитися, що дані відеофайли гарної якості та

не мають у собі перешкод, як візуального, так і аудіо характеру.

По-друге, необхідно підготувати необхідний інструментарій із програмного та апаратного забезпечення. Під апаратним забезпеченням мається на увазі комп'ютер або ноутбук, який може права та можливості запускати програми. Під програмний інструментарій підпадають програми для розпізнавання зазначеного переліку емоцій людини по її виразу обличчя та програма яка зможе виділити аудіофрагменти із відеофайлу по визначених часових позиціях.

Наступним кроком є власне запуск програми по розпізнаванню емоцій людини по її обличчю для заданого відеофайлу та отримання результатів виконання цієї програми. Далі, базуючись на цих результатах, наступна програма виокремлює відповідні аудіофрагменти з емоційними ознаками із початкового відеофайлу та зберігає їх в попередньо визначеній директорії.

Після цього, настає етап обробки отриманих емоційно забарвлених аудіофрагментів. По-перше, видаляються всі аудіофрагменти, тривалість яких є нижчою, а ніж попередньо встановлене мінімальне значення тривалості. В якості цього значення можуть виступати різні величини, наприклад, півсекунди, або секунда. Наступним кроком є аналіз отриманих аудіофрагментів на тривалість тиші в них, оскільки можуть бути моменти, коли обличчя людини в кадрі вказує на якусь емоційну ознаку, але при цьому в аудіопотоці панує тиша. Таким чином, слід виміряти тривалість цих беззвучних моментів та відняти її від загальної тривалості цього аудіофрагменту і якщо отримане значення виявляється меншим ніж попередньо встановлене мінімальне значення тривалості, то, відповідно, цей аудіофрагмент видаляється.

Нарешті, з отриманих емоційно забарвлених аудіофрагментів формується база даних, яка потім може бути використана для навчання нейромережевих моделей.

Перейдемо до критеріїв оцінювання ефективності формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів. Деякі з критеріїв взяті з попередньої статті авторів [10], які також доповнені новими і наведені у таблиці 1.

Таблиця 1

Критерії оцінювання ефективності формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів

Назва критерію	Опис критерію
C_1	Можливість формування баз даних без залучення професійних акторів
C_2	Наявність спонтанних висловлювань
C_3	Наявність візуальних даних акторів, на додачу до голосових висловлювань
C_4	Наявність можливості автоматичного маркування елементів бази даних
C_5	Більш ніж один дискретний емоційний стан
C_6	Можливість автоматичного формування елементів бази даних після запуску програми та відповідного відеофайлу
C_7	Відповідність сформованих аудіофрагментів до попередньо визначених вимог

Завдяки цим критеріям ефективності можна оцінювати ефективності формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів. Через це увесь процес можна буде формалізувати та контролювати його відповідність до попередньо заданих вимог завдяки наявним числовим метрикам.

Висновки. У даній статті розроблено концептуальну модель формування веб-орієнтованих баз даних емоційно забарвлених голосових сигналів, яка формалізує процес створення таких баз даних. Розроблена модель та наведені критерії ефективності формування таких баз даних дозволяють спростити процес створення емоційних аудіоданих для навчання нейронних мереж, оскільки цей процес може контролюватися відповідно до попередньо заданих вимог, використовуючи наведені числові метрики. Завдяки цьому, взаємодія між людиною та робототехнічними засобами та застосунками з штучним інтелектом може бути суттєво покращена.

Список літератури:

1. Chamishka S., Madhavi I., Nawaratne R., Alahakoon D., De Silva D., Chilamkurti N., Nanayakkara V. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*. 2022. https://doi.org/10.1007/s11042-022-13363-4

2. Ekman P. Basic Emotions. *Handbook of Cognition and Emotion*. Chichester, UK, 2005. Pp. 45–60. https://doi.org/10.1002/0470013494.ch3.

3. Zhou K., Sisman B., Liu R., Li H. Emotional voice conversion: Theory, databases and ESD. *Speech communication*, *137*. 2022. Pp. 1–18. https://doi.org/10.1016/j.specom.2021.11.006

4. Deschamps-Berger T., Lamel L., Devillers L. End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings. У *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2021. https://doi.org/10.1109/acii52823.2021.9597419

5. Tits N., El Haddad K., Dutoit T. Emotional Speech Datasets for English Speech Synthesis Purpose: A Review. У *Advances in Intelligent Systems and Computing*. Springer International Publishing. 2019. Pp. 61–66. https://doi.org/10.1007/978-3-030-29516-5_6

6. Christy A., Vaithyasubramanian S., Jesudoss A., Praveen, M. D. A. Multimodal speech emotion recognition and classification using convolutional neural network techniques. *International Journal of Speech Technology*. 2020. №23(2). Pp. 381–388. https://doi.org/10.1007/s10772-020-09713-y

7. Larradet F., Niewiadomski R., Barresi G., Caldwell D. G., Mattos, L. S. Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection. *Frontiers in Psychology*, *11*. 2020. https://doi.org/10.3389/fpsyg.2020.01111

8. Niu M., Jaiswal M., Mower Provost E. From Text to Emotion: Unveiling the Emotion Annotation Capabilities of LLMs. У *Interspeech 2024*. ISCA. 2024. Pp. 2650–2654.https://doi.org/10.21437/interspeech.2024-2282

9. Brauwers G., Frasincar F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1. 2021. https://doi.org/10.1109/tkde.2021.3126456

10. Дичка І.А., Терейковський І.А., Самофалов А.В., Терейковська Л.О., Романкевич В.О. Множина критеріїв ефективності формування баз даних емоційно забарвлених голосових сигналів. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка».* 2023. №1(21). С. 65–74. https://doi.org/10.28925/2663-4023.2023.21.6574

Samofalov A.V., Tereikovskyi I.A. CONCEPTUAL MODEL OF THE FORMATION OF WEB-ORIENTED DATABASES OF EMOTIONALLY COLORED VOICE SIGNALS

This article considers the conceptual model of the formation of web-oriented databases of emotionally colored voice signals. The given model examines the creation process and provides a list of criteria for evaluating the effectiveness of the formation of such databases. It was found that recognition of human emotions in speech is one of the most important tasks that is performed every day when people interact with each other. Nowadays, robotics and applications based on artificial intelligence are rapidly developing, which try to make the interaction with users closer to the interaction between people by feeling. To achieve the most comfortable user experience, these applications and robotics must be trained on large amounts of emotional data. Human emotion recognition using automated computer tools is particularly valuable and relevant for such applications as emotion recognition in customer service, mental health monitoring, and human-computer interaction, etc. At the beginning of this article, a list of the main types of databases of emotionally colored voice signals is considered and the challenges and tasks faced by researchers during their formation are defined. After that, an analysis of some approaches to the processing and classification of emotional states is carried out. Finally, the developed model is presented and the criteria for the effectiveness of the formation of web-oriented databases of emotionally colored voice signals are presented, which allow to simplify the process of creating emotional audio data for training neural networks, since it is found that this process can be controlled according to the predefined requirements, using the given numerical metrics . Due to this, the interaction between humans and robotics and applications with artificial intelligence can be significantly improved.

Key words: *conceptual model, databases, neural networks, emotion, performance criteria.*