КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1(21), 2023
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2023.21.6574
УДК 004.65

Дичка Іван Андрійович
Доктор технічних наук, професор, декан факультету прикладної математики Національного Технічного Університету України «Київський Політехнічний Інститут імені Ігоря Сікорського»
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна
ORCID ID 0000-0002-3446-3076
dychka@pzks.fpm.kpi.ua

Терейковський Ігор Анатолійович
Доктор технічних наук, професор, професор кафедри системного програмування і спеціалізованих комп'ютерних систем Національного Технічного Університету України «Київський Політехнічний Інститут імені Ігоря Сікорського»
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна
ORCID ID 0000-0003-4621-9668
terejkowski@ukr.net

Самофалов Андрій Вікторович
Аспірант кафедри системного програмування і спеціалізованих комп'ютерних систем Національного Технічного Університету України «Київський Політехнічний Інститут імені Ігоря Сікорського»
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна
ORCID ID 0009-0002-1205-5044
andrew.samofalov@gmail.com

Терейковська Людмила Олексіївна
Доктор технічних наук, доцент, доцент кафедри інформаційних технологій проєктування та прикладної математики Київського національного університету будівництва і архітектури
Київський національний університет будівництва і архітектури, Київ, Україна
ORCID ID 0000-0002-8830-0790
tereikovskal@ukr.net

Романкевич Віталій Олексійович
Доктор технічних наук, професор, завідувач кафедри системного програмування та спеціалізованих комп'ютерних систем Національного Технічного Університету України «Київський Політехнічний Інститут імені Ігоря Сікорського»
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна
ORCID ID 0000-0003-4696-5935
zavkaf@scs.kpi.ua

МНОЖИНА КРИТЕРІЇВ ЕФЕКТИВНОСТІ ФОРМУВАННЯ БАЗ ДАНИХ ЕМОЦІЙНО ЗАБАРВЛЕНИХ ГОЛОСОВИХ СИГНАЛІВ

Анотація. Значна кількість створених баз даних емоційного мовлення на різних мовах свідчить про великий інтерес дослідницької спільноти до питань синтезу емоційних голосових сигналів та розпізнавання емоцій у голосі людини. У наш час значного використання набувають пристрої, які використовують голосовий інтерфейс взаємодії з користувачем, що особливо виражено в певних роботехнічних системах.

В якості основи для створення комп'ютерних систем розпізнавання емоцій в голосі людини зазвичай використовують нейронні мережі, для навчання яких і потрібні достатньо великі за обсягом бази даних емоційно забарвлених голосових сигналів. Основним підходом, який застосовується при створенні таких баз даних є залучення акторів для відтворення заданого спектру емоцій в їх голосових висловлюваннях, та, відповідно, використання спеціалізованого обладнання для запису та аналізу отриманих аудіоданих. Однак цей підхід

вимагає значних часових та ресурсних затрат, що не дозволяє генерувати значні масштаби емоційних голосових висловлювань в осяжні проміжку часу.
Тому для оцінки ефективності формування баз даних емоційно забарвлених голосових сигналів наведено перелік критеріїв, за якими були оцінені засоби формування емоційних баз даних. Результати оцінювання дозволяють обґрунтовано стверджувати, що відомі засоби формування емоційно забарвлених баз даних голосових сигналів людини мають певний ряд недоліків. Для підвищення ефективності засобів формування баз даних емоційних голосових сигналів людини доцільно мати можливість формування баз даних без залучення професійних акторів, наявність спонтанних висловлювань, а не тільки попередньо визначених, наявність багатоголосих висловлювань, а саме діалогів, та наявність можливостей для підрахування часу та обчислювальних ресурсів, які необхідні для формування елементів бази даних.

Ключові слова: база даних; розпізнавання емоцій; голосовий сигнал; критерій ефективності.

ВСТУП

Важко переоцінити значення емоцій в житті кожної людини. Завдяки емоціям робиться можливою участь людини в соціальних аспектах життя. Без емоції люди будуть схожі в певній мірі на машини, у тому сенсі, що весь процес спілкування між людьми зведеться до обміну повідомленнями. Зазначимо, що в деяких мовах структури питальних речень збігаються зі структурами звичайних речень, що використовуються у нейтральному, радісному, саркастичному та інших тонах. Тому беземоційний обмін повідомленнями викликатиме значні труднощі у повсякденному соціальному житті людей.

Під час стрімкого розвитку голосових асистентів та застосунків, що використовують штучний інтелект, існує необхідність в застосуванні емоційного мовлення в синтезованих голосах. Це синтезовані емоційні голосові сигнали можуть використовуватися як в різних електронних портативних ґаджетах, так і в більш складних роботехнічних системах.

У той самий час, як роботи можуть використовувати емоційне мовлення під час своєї взаємодії з користувачем, вони все ж таки не можуть “зрозуміти” природу цих емоцій. Через це, використання емоцій, притаманних людині, в синтезованій мові робота є по суті покращеним інтерфейсом для взаємодії з користувачем, схожим на гарний графічний дизайн додатку. Як синтезовані емоції, так і вдалий графічний дизайн покликані зробити більш комфортною взаємодію користувача з програмою, чи власне роботом. Ця асоціація штучно згенерованих емоцій з інтерфейсом пов’язана з тим, що з точку зору роботи програми, що лежить в основі, немає різниці чи є графічний інтерфейс додатку максимально приємним, чи він складається з купи незрозумілих елементів, які не викликають у людини нічого, окрім бажання скоріше припинити взаємодію з цієї програмою та, за можливості, знайти іншу. Для того щоб зменшити виникнення в користувачів цих негативних емоцій, необхідно приділити увагу застосуванню якомога більшого діапазону емоцій при синтезі штучного мовлення. Саме тут і стають в нагоді бази даних емоційно забарвлених голосових сигналів. Ці бази даних можуть бути використані не тільки для генерації емоційних висловлювань, але і для розпізнавання емоцій у вже записаних голосових сигналах.

Якщо відокремити нейтральний тон голосу в окрему категорію, то усі бази емоційного мовлення умовно можна поділити на два типи: бази даних, які мають висловлювання тільки в нейтральному тоні, та бази даних, які містять висловлювання в інших емоціях, часто включаючи і власне нейтральний тон. Основною перепоною, яка не дозволяє дослідникам створювати велику кількість навчальних прикладів в базах

даних емоційних висловлювань, є необхідність залучення акторів для відтворення емоцій в голосових сигналах та використання спеціалізованого обладнання для запису та аналізу отриманих даних. Трудомісткість процесу та значні затрати часу при цьому і є власне чинниками, які роблять важким створення значних масштабів емоційних аудіоданих в осяжні проміжку часу.

Постановка проблеми. Вдосконалення методологічного забезпечення формування емоційно забарвлених корпусів людських висловлювань.

Аналіз останніх досліджень і публікацій. Аналіз літературних джерел показав, що зазвичай автори виділяють шість основних емоцій у людини: радість, страх, сум, здивування, злість та відраза [1]. Ці емоції можуть бути виражені і в голосі людини, тому їх можна взяти як основні емоції, які будуть розглядатися. Зазвичай, під час аналізу голосу людини ще вводять стан, або емоцію, яка відповідає нейтральному, або спокійному тону. Звісно, кількість емоцій людини є набагато більшою, аніж сім, але саме це звуження до невеликого числа дискретних емоційних станів є гарною основою для формування баз даних емоційних голосових сигналів людини. Для аналізу отриманих аудіофрагментів можна використовувати різні підходи. В роботі [2] наводяться такі показники як швидкість мовлення, фундаментальна частота та амплітуда голосу. У статті [3] вказується вплив синтаксису висловлювання на його вимову. Звичайно, це лише невеликий перелік критеріїв, які можуть бути використані для оцінювання отриманих голосових даних.

Формуванню критеріїв ефективності формування баз даних емоційно забарвлених голосових сигналів присвячено науково-практичну роботу [4]. Особливий інтерес в даній роботі представляє розділ 1.3 «Аналіз технологій розпізнавання емоцій та особи за голосом». В даному розділі наведені найбільш відомі засоби розпізнавання емоцій людини за її голосом, а також наведено параметри оцінки ефективності засобів розпізнавання емоційних станів людини за голосом. Деякі з цих параметрів наведені в «*Таблиці 1*». Варто зазначити, що в цьому розділі увага на оцінювання розроблених баз даних емоційного мовлення не була акцентована.

Таблиця 1

Критерії оцінки ефективності засобів розпізнавання емоційного стану

Номер критерію	Опис критерію
1	Адаптованість до віку диктора
2	Адаптованість до статі диктора
3	Адаптованість до мови диктора
4	Можливість розпізнавання в умовах багатоголосся
5	Дикторонезалежність процесу розпізнавання
6	Можливість фільтрації шуму
7	Можливість аналізу спонтанного мовлення
8	Можливість нормалізації гучності голосового сигналу

При цьому в проаналізованій літературі [1]–[10] не було знайдено переліку характеристик баз даних емоційного мовлення, за якими можна оцінювати створені бази даних.

Мета статті. Метою статті є визначення множини критеріїв ефективності засобів формування баз даних емоційно забарвлених голосових сигналів.

ВИЗНАЧЕННЯ КРИТЕРІЇВ ЕФЕКТИВНОСТІ

Для визначення множини критеріїв ефективності формування баз даних емоційних висловлювань було використано результати роботи [2], де обґрунтовано підхід, що передбачає проведення аналізу відомих рішень в області створення відповідних баз даних.

Спочатку були розглянуті бази даних, які складаються тільки з голосових фрагментів у нейтральному тоні. На це є дві причини. По-перше, судячи з власного досвіду, нейтральний тон є найбільш розповсюдженим у повсякденному житті людини, та, відповідно, найбільш розповсюдженою "емоцією". По-друге, існує велика кількість баз даних нейтрально забарвлених голосових висловлювань. Тому, якщо ми спочатку зможемо навчити нейронну мережу розпізнавати нейтральний тон, то всі подальші роботи по розпізнаванню емоцій можуть бути спрощені, оскільки всі фрагменти аудіофайлів, що не містять нейтрального тону, відповідно, будуть мати інші емоції у собі. Однак при цьому треба враховувати, що кожна людина може мати свій відтінок нейтрального тону.

Так як мета статті полягає в огляді баз даних емоційних висловлювань, тому розглянуто лише одну базу даних голосу людини у нейтральному тоні. В якості прикладу візьмемо базу даних CMU Arctic, представлену в роботі [5]. При виборі вихідного тексту для читання спікерами, автори зупинились на книжках з проекту Гутенберг. Причиною цьому є випуск бази даних під відкритою ліцензією, через це і книжки, що були використані, мають також бути у вільному доступі. Аудіодані записані чотирма спікерами в загальній американські англійській, канадському та шотландському акцентах. Кожна з баз даних Arctic складається з близько 1200 висловлювань.

Перейдемо саме до огляду існуючих баз даних емоційно забарвлених голосових сигналів. Їх особливістю є те, що аудіозаписи, що містять емоції, були записані дикторами в лабораторних умовах. У деяких дослідженнях, після запису голосового матеріалу, були проведені тести на певні властивості отриманих даних, наприклад, натуральність та силу отриманих емоцій. Треба зауважити, що при проведені тестів на сприйняття емоцій, потрібно враховувати, що не всі люди сприймають емоції однаково. На цей процес сприйняття та класифікації емоцій впливає життєвий досвід кожної людини та її особисте ставлення до навколишнього світу. Наприклад, актор може вимовити речення, на їх думку, у нейтральному тоні, а рецензент почує в цьому висловлюванні пасивну агресію, і тому класифікує цей аудіофрагмент, як той, що містить емоцію гніву, а не нейтральний тон. Це в свою чергу призведе до того, що дослідники отримають результати відмінні від очікуваних. Але з точки зору рецензенту, вони не зробили нічого неправильного, оскільки класифікували це висловлювання з позиції особистого ставлення до нього. Це можна трактувати як інакше трактування емоції в голосовому сигналі. Тому при навчанні нейронних мереж це може бути розглянуто не як помилка, а як інше сприйняття емоції в голосовому сигналі людиною.

Першою з баз даних емоційних висловлювань розглянемо Emotional Speech Database (ESD), описану в статті [6]. Особливістю цієї статті є те, що в ній представлено не тільки саму ESD, але і наведено останні досягнення та виклики в сфері емоційного перетворення голосу. Задекларовано, що серед відомих баз даних емоційних висловлювань, ця база даних дозволяє найбільш повно задовольнити зростаючим потребам дослідників в їх роботах у напрямку синтезу емоційно забарвлених голосових сигналів. У створенні ESD взяли участь двадцять акторів - по п'ять жінок та чоловіків, які вимовили 350 висловлювань у різних емоціях на англійській та китайських мовах. Вік спікерів складає від 25 до 35 років. В якості емоцій, які відтворювали актори, були обрані радість, злість, здивування, сум та нейтральний тон. Запис висловлювань у

лабораторний умовах та за допомогою спеціалізованого обладнання надає гарну якість отриманих голосових записів.

Наступною базою даних розглянемо EmoDB [7]. Ця база даних емоційних висловлювань на німецькій мові. Загальна кількість висловлювань складає близько 800 у семи різних емоційних станах: нейтральний тон, радість, страх, злість, відраза, сум та нудьга. Актори, що прийняли участь у записі голосів для бази даних, були відібрані під час конкурсу за допомогою трьох експертів. Експертне журі відібрало десять спікерів - п’ять жінок та п’ять чоловіків. Важливим позитивним аспектом даного дослідження є реалізація тесту на якість отриманих емоцій та їх природність, у якому взяли участь двадцять учасників.

Перейдемо до бази RAVDESS, якій присвячена стаття [8]. Мовою аудіозаписів в даній базі даних є північноамериканська англійська. Відмінною особливістю цієї бази даних є наявність не тільки емоційних висловлювань, але і записів емоційного співу. Автори вказують точну кількість отриманих записів, що складає вражаючих 7356. В записі аудіоданих прийняли участь двадцять чотири професійних актори (12 жінок та 12 чоловіків) у віці від 21 до 33 років. В якості відтворюваних емоцій визначили наступні вісім станів: радість, злість, страх, здивування, відраза та сум. Два стани, що залишились, представлені нейтральним та спокійним тонами. Цікаво зазначити, що автори спеціально зробили сепарацію нейтрального тону на два різних стани, щоб уникнути різного ставлення до нейтральної емоції в акторів. Задля цього спокійний тон зазначений як нейтральний, але з невеликим позитивним відтінком. Отримані записи були ретельно проаналізовані та протестовані за допомогою оцінювачів. Наостанок зазначимо, що всі записи емоціональних висловлювань доступні не тільки у лише голосовому вигляді, але і в форматах лише-обличчя та обличчя-та-голос. Це є чудовою особливістю цієї бази даних, що дозволяє проводити подальші дослідження спираючись не лише на голосові дані, але і на візуальні також.

Далі зазначимо базу даних JL-Corpus, наведену в статті [9]. З самого початку автори вказують важливістю емоційного мовлення при побудові взаємодії між людьми та роботами. Автори зазначають про важливість вторинних емоцій при побудові цієї взаємодії. Задля цього, розроблена база даних включає не тільки п’ять основних емоцій (нейтральний тон, радість, злість, сум та хвилювання), але і п’ять вторинних (задумливість, вибачливість, збентеження, тривога та ентузіазм). Загальна кількість речень становить 2400 (1200 речень в основних емоціях та 1200 речень у вторинних емоціях). Всі речення були записані за участю чотирьох професійних спікерів. Варто зазначити, що хоча мовою бази даних є англійська, все ж таки авторами поставлено акцент на представлення чотирьох довгих голосних, що зустрічаються в новозеландській англійській мові.

Наостанок, проаналізуємо базу даних EMOVO Corpus [10]. В якості базових емоцій в цій базі даних також були обрані шість основних емоцій (радість, страх, гнів, відраза, сум та здивування) плюс нейтральний тон. Для запису аудіоданих були залучені шість професійних акторів, а саме три жінки та три чоловіки. В якості лінгвістичного матеріалу були обрані чотирнадцять фраз, не всі з яких були семантично нейтральними. На записаних аудіоданих був проведений оцінний тест, який показав гарні загальні результати розпізнавання емоцій у вісімдесят відсотків.

Переходячи до власне оцінки баз даних емоційних висловлювань, був розроблений перелік критеріїв, наведений у «*Таблиці 2*».

Таблиця 2

Критерії оцінки баз даних емоційних висловлювань

Назва критерію	Опис критерію
C_1	Можливість формування баз даних без залучення професійних акторів
C_2	Наявність різних за статтю акторів
C_3	Наявність різних за віком акторів
C_4	Запис у лабораторних умовах
C_5	Фільтрація шуму
C_6	Були проведені тести на отриманих даних
C_7	Наявність спонтанних висловлювань
C_8	Одна основна мова голосових висловлювань
C_9	Наявність багатоголосих висловлювань
C_{10}	Більш ніж один дискретний емоційний стан
C_{11}	Наявність візуальних даних акторів, на додачу до голосових висловлювань
C_{12}	Наявність можливості автоматичного маркування елементів бази даних
C_{13}	Наявність можливостей для підрахування часу, необхідного для формування елементів бази даних
C_{14}	Наявність можливості для підрахування необхідних обчислювальних потужностей в залежності від наявного часу, які необхідні для формування елементів бази даних

Перелік параметрів, наведених у «*Таблиці 2*», обґрунтований тим, що він містить не тільки критерії присутні в розглянутих базах даних голосових сигналів людини, але і ті, які могли б підвищити ефективність пристосування баз даних емоційного мовлення до апаратного забезпечення комп'ютерних засобів.

Оцінка ефективності засобів формування баз даних голосових сигналів людини, розглянутих у статті, за критеріями із «*Таблиці 2*», наведена у «*Таблиці 3*». Критерії були оцінені по бінарній шкалі, де "+" вказує на наявність відповідного критерію, а "—" означає його відсутність.

Таблиця 3

Оцінка критеріїв ефективності засобів формування баз даних емоційних голосових сигналів людини

Засіб формування бази даних	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}
CMU Arctic [5]	—	+	+	+	+	—	—	+	—	—	—	+	—	—
ESD [6]	—	+	+	+	+	+	—	—	—	+	—	—	—	—
EmoDB [7]	—	+	+	+	+	+	—	+	—	+	—	+	—	—
RAVDESS [8]	—	+	+	+	+	+	—	+	—	+	+	—	—	—
JL-Corpus [9]	—	+	+	+	+	+	—	+	—	+	—	+	—	—
EMOVO Corpus [10]	—	+	+	+	+	+	—	+	—	+	—	—	—	—

Отже, результати оцінювання дозволяють обґрунтовано стверджувати, що відомі засоби (методи, моделі, підходи) формування емоційно забарвлених баз даних голосових сигналів людини мають ряд недоліків, що в першу чергу пов'язані з критеріями C_1, C_7 , C_9, C_{13} та C_{14}.

Очевидно, що вдосконалення засобів формування баз даних емоційних голосових сигналів людини повинно бути пов'язане з ліквідацією означених недоліків. Таким чином, перспективні шляхи подальших досліджень доцільно співвіднести із забезпеченням засобів формування баз даних емоційних висловлювань можливостями

формування баз даних без залучення професійних акторів, наявністю спонтанних висловлювань, а не тільки попередньо визначених, наявністю багатоголосих висловлювань, а саме діалогів, та наявністю можливостей для підрахування часу та обчислювальних ресурсів, які необхідні для формування елементів бази даних.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Проведений аналіз відомих засобів формування баз даних емоційно забарвлених голосових сигналів дозволяє стверджувати, що одним із основних недоліків є необхідність залучення акторів та професійного обладнання до створення цих баз даних, що не дозволяє генерувати значні масштаби емоційних аудіоданих в осяжні проміжку часу. Також залучення певної кількості акторів не робить можливою значну сепарацію дикторів за віком, що може сказатися на якості розпізнавання та синтезі емоційних голосових висловлювань.

Водночас, як розмежовування емоцій в обмежену кількість дискретних станів значно спрощує процес створення таких баз даних, все ж таки людські емоції є більш складними за своєю природою. Наприклад, дуже часто не можна перейти зі стану суму в стан радості за один крок. Для цього переходу потрібні проміжні емоції, як-от нейтральний стан, здивування, інтерес. З іншого боку, збільшення спектру емоції, з якими може оперувати програма, може зменшити якість розпізнавання чи відтворення штучних емоцій, оскільки між великою кількістю емоцій лежить дуже тонка межа, яку тільки за наявним голосовим сигналом і не кожна людина може відрізнити.

Тому для підвищення ефективності засобів формування баз даних емоційних голосових сигналів людини повторить доцільно мати можливість формування баз даних без залучення професійних акторів, наявність спонтанних висловлювань, а не тільки попередньо визначених, наявність багатоголосих висловлювань, а саме діалогів, та наявність можливостей для підрахування часу та обчислювальних ресурсів, які необхідні для формування елементів бази даних.

Таким чином, існує необхідність в розробках більших за обсягом і гендерною та віковою диференціацією баз даних емоційного мовлення. Для цього є важливим опанування нових шляхів до формування цих баз даних. На додачу, підходи до визначення певних емоційних ознак в голосі людини спочатку можна спробувати опанувати на одній певній мові, а потім застосувати найкращі рішення вже на інших. Відомо, що кожна мова має свій набір фонетичних ознак та принципів побудов висловлювань, але сам інструментарій для визначення та виокремлення цих ознак може збігатися у багатьох мовах.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Ekman, P. (2005). Basic Emotions. У Handbook of Cognition and Emotion (с. 45–60). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch3

2 Bachorowski, J.-A., & Owren, M. J. (1995). Vocal Expression of Emotion: Acoustic Properties of Speech Are Associated With Emotional Intensity and Context. Psychological Science, 6(4), 219–224. https://doi.org/10.1111/j.1467-9280.1995.tb00596.x

3 Hirschberg, J. (2006). Pragmatics and Intonation. У The Handbook of Pragmatics (eds L.R. Horn and G. Ward). https://doi.org/10.1002/9780470756959.ch23

4 Терейковська, Л. О. (2023). Методологія автоматизованого розпізнавання емоційного стану слухачів системи дистанційного навчання [Дис. д. т. н., Київський національний університет будівництва і архітектури]. Інституційний репозитарій Національного транспортного університету. http://www.ntu.edu.ua/nauka/oprilyudnennya-disertacij/

5 Kominek, J., & Black, A. (2004). The CMU Arctic speech databases. SSW5-2004. https://www.lti.cs.cmu.edu/sites/default/files/CMU-LTI-03-177-T.pdf (дата звернення: 01.06.2023)

6 Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1–18. https://doi.org/10.1016/j.specom.2021.11.006

7 Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. У Interspeech 2005. ISCA. https://doi.org/10.21437/interspeech.2005-446

8 Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), Стаття e0196391. https://doi.org/10.1371/journal.pone.0196391

9 James, J., Tian, L., & Inez Watson, C. (2018). An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. У Interspeech 2018. ISCA. https://doi.org/10.21437/interspeech.2018-1349

10 Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: an Italian Emotional Speech Database. У Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3501–3504, Reykjavik, Iceland. European Language Resources Association (ELRA).

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (21), 2023
ISSN 2663 - 4023

Ivan Dychka
Doctor of Science, professor, dean of the Faculty of Applied Mathematics
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID 0000-0002-3446-3076
dychka@pzks.fpm.kpi.ua

Ihor Tereikovskyi
Doctor of Science, professor, professor of the Department of System Programming and Specialized Computer Systems
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID 0000-0003-4621-9668
terejkowski@ukr.net

Andrii Samofalov
Ph.D. student of the Department of System Programming and Specialized Computer Systems
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID 0009-0002-1205-5044
andrew.samofalov@gmail.com

Lyudmila Tereykovska
Doctor of Science, associate professor, associate professor of the the Faculty of Automation and Information Technologies
Kyiv National University of Construction and Architecture, Kyiv, Ukraine
ORCID ID 0000-0002-8830-0790
tereikovskal@ukr.net

Vitaliy Romankevich
Doctor of Science, professor, head of the Department of System Programming and Specialized Computer Systems
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID 0000-0003-4696-5935
zavkaf@scs.kpi.ua

MULTIPLE EFFECTIVENESS CRITERIA OF FORMING DATABASES OF EMOTIONAL VOICE SIGNALS

Abstract. A significant number of created databases of emotional speech in different languages testifies to the great interest of the research community in the problems of synthesis of emotional voice signals and recognition of emotions in the human voice. In our time, devices that use a voice interface for user interaction are gaining significant use, which is especially important in certain robotic systems.

As a basis for the creation of computer systems for recognizing emotions in a person's voice, neural networks are usually used, and for their training, sufficiently large databases of emotional voice signals are required. The main approach used in the creation of such databases is the involvement of actors to reproduce a given range of emotions in their utterances, and, accordingly, the use of specialized equipment for recording and analyzing the received audio data. However, this approach requires significant time and resource costs, which does not allow generating significant volumes of emotional voice expressions in a reasonable period of time.

Therefore, in order to evaluate the effectiveness of the formation of databases of emotional voice signals, a list of criteria is given, according to which the means of forming emotional databases were evaluated. The results of the evaluation allow us to reasonably claim that the known means of forming emotional databases of human voice signals have a certain number of shortcomings. In order to increase the efficiency of the means of forming databases of emotional voice signals, it is advisable to provide the possibility of forming databases without the involvement of professional actors, the presence of spontaneous expressions, not only predetermined ones, the presence of polyphonic expressions, namely dialogues, and the presence of opportunities for evaluating time and computing resources, which are necessary for the formation of database elements.

Keywords: database; emotion recognition; voice signal; efficiency criterion.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1 Ekman, P. (2005). Basic Emotions. In Handbook of Cognition and Emotion (p. 45–60). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch3

2 Bachorowski, J.-A., & Owren, M. J. (1995). Vocal Expression of Emotion: Acoustic Properties of Speech Are Associated With Emotional Intensity and Context. Psychological Science, 6(4), 219–224. https://doi.org/10.1111/j.1467-9280.1995.tb00596.x

3 Hirschberg, J. (2006). Pragmatics and Intonation. In The Handbook of Pragmatics (eds L.R. Horn and G. Ward). https://doi.org/10.1002/9780470756959.ch23

4 Tereykovska, L. (2023). Methodology of automated recognition of the emotional state of listeners of the distance learning system [Dissertation, Kyiv National University of Construction and Architecture]. Institutional repository of National transport university. http://www.ntu.edu.ua/nauka/oprilyudnennya-disertacij/

5 Kominek, J., & Black, A. (2004). The CMU Arctic speech databases. SSW5-2004. https://www.lti.cs.cmu.edu/sites/default/files/CMU-LTI-03-177-T.pdf (date of access: 01.06.2023)

6 Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1–18. https://doi.org/10.1016/j.specom.2021.11.006

7 Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In Interspeech 2005. ISCA. https://doi.org/10.21437/interspeech.2005-446

8 Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), Стаття e0196391. https://doi.org/10.1371/journal.pone.0196391

9 James, J., Tian, L., & Inez Watson, C. (2018). An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. In Interspeech 2018. ISCA. https://doi.org/10.21437/interspeech.2018-1349

10 10) Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: an Italian Emotional Speech Database. У Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3501–3504, Reykjavik, Iceland. European Language Resources Association (ELRA).