УДК 621.38(62-52)

В.А. Слюсаренко, Т.А. Ліхоузова

ПРОГРАМНО-АПАРАТНИЙ КОМПЛЕКС ДЛЯ СИНТЕЗУ МОВИ ТА РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД

Анотація: Розглядається програмно-апаратний комплекс, призначений для підвищення якості життя людей з вадами мовлення та слуху. Основними перевагами комплексу є портативність, мобільність та невисока ціна.

Ключові слова: пристрій для людей з вадами мовлення, портативний синтезатор мови, розпізнавання голосових команд.

Вступ

За неофіційними даними в Україні більше 150–200 тисяч нечуючих, статистики про кількість людей, що з тих чи інших причин не можуть говорити взагалі немає. Проте, таких людей багато і держава про них майже не піклується. Найбільша складність для таких людей полягає у спілкуванні в повсякденному житті з людьми, яким невідома мова жестів. У світі існують пристрої для вирішення подібних проблем, але в нашій країні їх важко знайти і зазвичай не кожен може собі це дозволити.

Аналіз існуючих рішень

На сьогодні існує кілька програмних комплексів, які в тій чи іншій формі здійснюють синтез мови, в таблиці 1 наведено короткий опис їх характерних особливостей. Ще можна виділити цілу групу програмних засобів, що застосовуються в операційній системі Windows, але їх ми не розглядаємо, оскільки вони мають інші цілі і не є кросплатформними.

Табл. 1: Характеристика існуючих засобів синтезу мови

Назва	Опис
RSynth	Програма не дуже високої якості. Незрозумілий інтерфейс та відсутня адекватна документація по використанню. Російський текст можна озвучити лише за умови друку транслітерацією. На вихід користувач отримує аудіо файл маловідомого формату. Немає версії для КПК. Низька якість озвучування. Розвиток припинено.
Festival	Незрозумілий інтерфейс, хоча наявна документація і розібратися можливо. Потребує великої кількості додаткових модулів для зручної роботи та необхідної мови. Теоретично російська мова підтримується, якість озвучування задовільна. Розвиток припинено.

Назва	Опис
Epos	Оскільки даний продукт створювався в Чехії, звучання російської мови є добрим. Проблемними лишаються лише деякі звуки (ч, ш, щ). Є підтримка кількох мов, але їх кількість дуже обмежена. Інтерфейс зрозумілий і майже зручний. Розвиток припинено.
FLite	Зручний та зрозумілий інтерфейс. Повна направленість на англійську мову, хоча має аналоги для інших мов, серед яких поки що лише італійська. Може озвучувати російський текст при умові набору транслітерацією. Продовжує розвиватись.
eSpeak	Особливо зручний інтерфейс, підтримка багатьох мов, висока швидкість роботи. Озвучування далеке від ідеального, проте цілком розбірливе. Працює під багатьма операційними системами, тому може бути встановлена на КПК. Активно розвивається. Використовує наголос на перший склад, що не притаманно правилам російської та української мов. Для виправлення цього необхідно вручну вказувати, який саме склад необхідно наголосити.
translate. google.com	Основною перевагою цього засобу є майже ідеальна вимова будь-якою мовою. При цьому основним недоліком є те, що це веб-сервіс, що працює лише за наявності доступу до інтернету. Ця особливість спричинює низьку швидкодію, що напряму залежить від якості підключення.
пристрій, розроблений у 2012 році донецькими студентами, що озвучує мову жестів	Винахід має форму спеціальних рукавичок, що за допомогою веб-камери відслідковують рухи людини (розпізнають специфічні жести) та озвучують показану фразу. Недоліком цього підходу є те, що людина з вадами мови повинна весь час перебувати у зоні видимості веб-камери, тобто перед комп'ютером чи ноутбуком, які і виконують озвучування фраз.

Постановка задачі

Розроблена система синтезу людської мови призначена для підвищення якості життя осіб з вадами мовлення у частині виконання таких процесів:

- придбання продуктів у магазинах;
- спілкування з оточуючими, що не володіють мовою жестів;
- уточнення необхідної інформації;
- інформування оточуючих при необхідності поділитись думками;
- відслідковування звернень до користувача;

- повідомлення користувача у разі голосних вигуків чи звертань;
- привернення уваги користувача до джерела звуку.

Для реалізації поставлених цілей система має вирішувати такі задачі:

- ручне введення тексту з клавіатури;
- аналіз введеного тексту;
- складання і відтворення відповідного аудіоряду;
- забезпечення можливості завчасної підготовки фраз;
- забезпечення швидкого введення тексту;
- розпізнавання характерних команд (ім'я та найуживаніші звернення);
- привернення уваги користувача у разі впізнання команд.

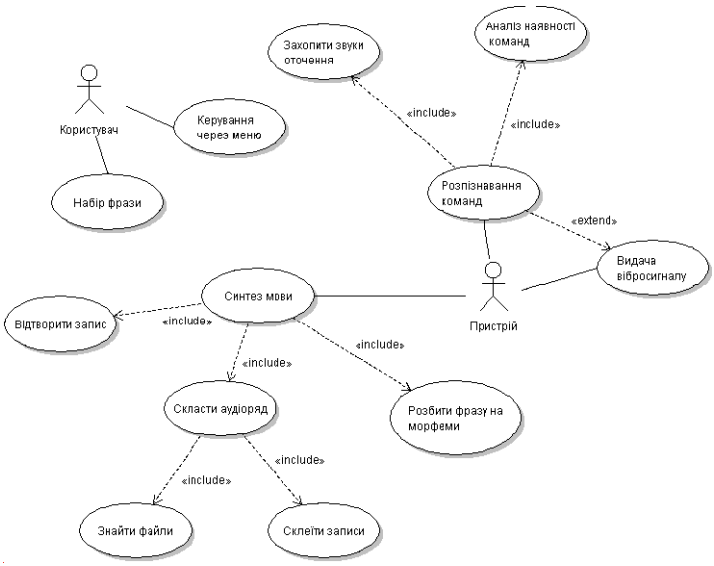

Рис. 1 – Структурна схема варіантів використання пристрою

Результати роботи

На рис. 2 представлена UML діаграма діяльності, що відповідає процесу синтезу мови. Користувач з усіх пунктів меню обирає той, що відповідає продукуванню мови, вводить з клавіатури текст для озвучування. Далі пристрій розділяє фрази на морфеми за заданими правилами розбору. Після цього відшукуються аудіо файли,

відповідні послідовності морфем. Вони почергово заносяться у буфер, з якого потім і відтворюються. Після цього користувач може або продовжити синтез мови, або обрати іншу функцію, або припинити роботу.

На рис.3 представлена UML діаграма діяльності, що відповідає процесу аналізу, для активації якого у пристрої є спеціальний режим – "контроль оточення". При увімкненні цього режиму відбувається постійний аналіз звуків оточуючого середовища. Відслідковуються характерні звуки, такі як ім'я користувача, гудок автомобіля, голосні тривожні звуки та попередження подається користувачу. Для сповіщення користувача використовується вібро-сигнал, що створюється вібро-мотором. Він привертає увагу та підказує, що слід озирнутися та бути насторожі. Виконання аналізу звуків оточення продовжується поки користувач не вимкне даний режим.

Рис. 2 – Синтезатор мови

Рис. 3. – Аналізатор мови

Вхідні дані для пристрою синтезу мови надходять з декількох джерел: від користувача (власника), розробників системи (інженерів) та оточення (найближчого оточуючого середовища). На момент придбання пристрою на картці пам'яті вже мають міститись записані у вигляді аудіо файлів морфеми, причому вони мають відповідати бажаній мові спілкування та статі голосу (чоловічий або жіночий). Ці дані можна назвати первісними. У процесі експлуатації до системи надходить ще дві групи даних: набрана з клавіатури фраза (певна послідовність слів, яку людина з вадами мови хоче озвучити) і голосова команда (вигук, оклик чи звертання, з яким може до власника звернутися стороння людина, чи звук, на який варто звернути увагу). Фраза для озвучування може бути абсолютно довільною, може містити всі букви алфавіту обраної мови та розділові знаки для досягнення бажаної інтонації висловлюван-

ня. Голосові команди включають ряд слів, на які користувач хоче звертати увагу (найбільш розповсюджені звертання, ім'я власника пристрою), а також звуки чи попереджувальні сигнали з високою гучністю (сигнал авто, звук гальмування, звук вибуху чи зіткнення, людський крик).

Вихідними даними є складений за уведеною користувачем фразою аудіо ряд, що відтворюється паралельно з його складанням. Тому можна віднести до вихідних даних і сам звуковий сигнал. При розпізнаванні довколишніх звуків вихідними даними є вібраційні сигнали, що надходять на руку користувача пристрою для сповіщення та привернення уваги.

Задача синтезу та розпізнавання мови може розглядатися як стандартна задача класифікації. Відомі методи розв'язання такого роду задач [1,2] можна умовно розділити на дві групи: ті, що потребують визначення ознак класифікації та ті, що в явному вигляді не потребують їх уточнення. Оскільки для визначення набору ознак потрібне використання додаткових ресурсів, яких у нашому випадку може не вистачити, було обрано методи без уточнення ознак. До них відносяться різного роду нейронні мережі, приховані послідовності Маркова (Hidden Markov Model), часові динамічні алгоритми (Dynamic Time Warping) та інші. Ми зупинились на використанні нейронних мереж, оскільки вони є добре вивченими і достатньо гнучкими для застосування на різних вхідних даних, крім того, можливість навчання — одна з головних переваг нейронних мереж перед традиційними алгоритмами для нашої задачі.

Для навчання нейронної мережі обрано метод навчання з учителем, що полягає в тому, що під час навчання на вхідний шар подаються аудіо записи команд для розпізнавання, сказані різними людьми і різний час, а також задаються очікувані правильні відповіді, що мали б бути отримані на виходах нейронної мережі. Таким чином спочатку вхідні дані аналізуються мережею, потім отримані результати зіставляються з бажаними та в разі необхідності вносяться певні корективи у ваги ребер мережі. При цьому варто зазначити, що на входи такої нейронної мережі повинні подаватись дані, що можуть бути однозначно інтерпретовані для покращення якості аналізу (так само працює людський мозок). Для цього вирішено застосувати швидке перетворення Фур'є [3] для сигналу однієї тривалості (1 секунда), оскільки частота гармонік перетворення Фур'є залежить від розмірності вхідного масиву. Проте аналіз всього вихідного масиву даних отриманого після застосування перетворення Фур'є займає досить тривалий час. Рішенням даної проблеми є вибір найінформативніших гармонік спектру. Таким чином піддаватимемо аналізу полосу частот від 300 Гц до 3400 Гц. При такому підході складність алгоритму зменшується, що дає змогу перенести програму на низькопродуктивну платформу – мі-

кроконтролер.

Базою першого прототипу розроблюваного пристрою є макетна плата STM32VLDiscovery. Цей вибір зроблено через те, що така плата вже містить крім мікроконтролера основний комплекс пристроїв (програматор, підсилювачі, аналогово-цифрові перетворювачі, резистори, конденсатори, світло діоди, входи для підключення зовнішніх пристроїв та ін.), при цьому вартість такої плати помірна. Для збереження аудіо файлів використовується картка пам'яті SD на 1 Гб (формат карт пам'яті, розроблений для використання в основному в портативних пристроях). Також у першому прототипі використано екран 4 на 20 символів; у наступних версіях передбачено менший екран та спеціально створену основну плату, що не міститиме зайвих елементів для досягнення максимальної портативності пристроїв.

Для оцінки ефективності розробленого програмно-апаратного комплексу виконано серію його випробувань згідно з вимогами документів [4]. До випробувань ввійшли: перевірка записів морфем, перевірка складання слів та фраз, перевірка правильності роботи меню, перевірка правильності розпізнавання команд та інші. За результатами випробувань визначено оптимальний склад та довжину морфем, а також перелік найбільш необхідних для розпізнавання звуків.

Висновки

Розроблено програмно-апаратний комплекс для синтезу мови. Алгоритм спроектовано з умовою мінімізації часу, що витрачається на синтез. Це досягається завдяки поєднанню елементів не довше двох літер. Якість синтезу забезпечує синтез із одиниць, записаних у гарній якості на зовнішній пам'яті. Розроблений алгоритм синтезу мови добре підходить для платформ з низькою продуктивністю. Достатній об'єм пам'яті забезпечує достатньо дешевий тип зовнішньої пам'яті - SD карта. Використання мови С також гарно впливає на швидкодію алгоритму.

Також розроблено програмне забезпечення для розпізнавання голосових команд. В основі лежить алгоритм розпізнавання на базі нейромережі, доповнений фільтром гармонік звуку на основі швидкого перетворення Фур'є. Основною перевагою розробленого ПЗ є можливість його переносу на низькопродуктивні платформи. Це досягається завдяки розробленому швидкому алгоритму. Основний приріст швидкості надає вибір невеликої кількості гармонік перетворення Фур'є, адже це значно зменшує розмірність нейронної мережі. Успішно розроблено перший прототип пристрою.

Біблиографічний список

1. *Журавлев Ю.И.* Распознавание. Математические методы. Программная система. Практические применения /

Ю.И.Журавлев, В.В.Рязанов, О.В.Сенько — М.: Фазис, 2006. ISBN 5-7036-0108-8.

2. *Mitchell T.* Machine Learning / T.Mitchell — McGraw-Hill Science/Engineering/ Math, 1997. ISBN 0-07-042807-7
3. *Лобанов Б.М.* Комп'ютерний синтез і клонування мови / Б.М.Лобанов, Л.І.Цирульник – Мінськ: Білоруська Наука, 2008 - 316с.
4. ГОСТ 34.603-92. Інформаційна технологія. Види випробувань автоматизованих систем.

Отримано 13.02.2013