УДК 004.522, 004.031.6
DOI: 10.20535/SRIT.2308-8893.2016.4.06

ПЕРСПЕКТИВИ РЕАЛІЗАЦІЇ ВБУДОВАНИХ СИСТЕМ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ НА БАЗІ RISC-МІКРОКОНТРОЛЕРІВ

І.А. МАРТИНЮК

Анотація. Реалізація систем автоматичного розпізнавання мови як складової частини звукового інтерфейсу керування інформаційними інтелектуальними системами сприяє підвищенню ефективності взаємодії людини з такими системами. Особливо актуальними натепер є дослідження в галузі вбудованих систем автоматичного розпізнавання. Проаналізовано перспективи реалізації вбудованих систем автоматичного розпізнавання мови на базі високопродуктивних RISC-мікроконтролерів. Обґрунтовано переваги такої реалізації порівняно з іншими рішеннями в цій галузі. Здійснено порівняльну характеристику високопродуктивних серій мікроконтролерів. Досліджено перспективи реалізації кожного етапу задачі розпізнавання за допомогою мікроконтролерної системи.

Ключові слова: автоматичне розпізнавання мови, вбудовані системи, мікроконтролерні системи.

ВСТУП

На сучасному етапі розвитку комп'ютерних технологій постає проблема зручного та ефективного способу взаємодії людини з інформаційними інтелектуальними системами. Оскільки усне мовлення є природним способом спілкування для людини, то технології автоматичного розпізнавання мови дозволять їй найбільш ефективно взаємодіяти з такими системами.

Використання систем автоматичного розпізнавання мови (САРМ) відкриває широкий спектр застосувань: від додатків для автоматичного набору тексту і транскрибації аудіозаписів до керування бортовими пристроями автомобілів та автоматизації процесів систем масового обслуговування (наприклад, збирання показників лічильників для комунальних служб).

Типово САРМ реалізовуються у вигляді програмних додатків для персональних комп'ютерів чи серверів (хмарні обчислення), оскільки потребують значних обчислювальних потужностей та ресурсів для розпізнавання мови в реальному часі. Проте така реалізація накладає ряд обмежень, зокрема щодо портативності та автономності пристроїв, які використовують таку технологію.

Реалізація вбудованих систем розпізнавання мови у вигляді автономних портативних модулів дозволить уникнути указаних обмежень та ефективно використовувати ці системи у таких галузях, як автотранспорт, авіація, соціальна сфера, робототехніка тощо.

Сфери застосування САРМ показано на рис. 1.

Як приклади застосування таких систем можна навести: голосове керування функціоналом автомобіля (для якого помилка розпізнавання не при-

зведе до аварійної ситуації), тим самим розвантажити водія для сконцентрування його увагу на дорозі; реалізацію мовного інтерфейсу в інвалідних кріслах для людей з обмеженими можливостями.

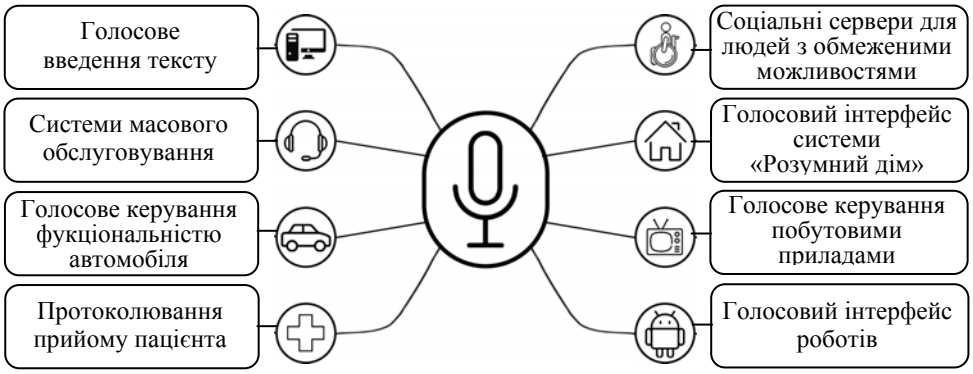

Рис. 1. Основні сфери застосувань САРМ

АНАЛІЗ ІСНУЮЧИХ ДОСЛІДЖЕНЬ

Ураховуючи основні проблеми [1] та аналізуючи останні дослідження в галузі розроблення вбудованих САРМ [2–12], варто відзначити, що реалізація технологій автоматичного розпізнавання мови для таких систем має обмеження, зокрема щодо економії обчислювальних ресурсів та пам'яті, які значно менші порівняно з персональними комп'ютерами. Також актуальною, особливо для систем з батарейним живленням, є проблема помірного споживання електроенергії.

Можна виділити такі основні напрями реалізації вбудованих САРМ за допомогою:

- програмованих вентильних матриць (FPGA);
- одноплатних комп'ютерів;
- цифрових сигнальних процесорів (DSP).

Ґрунтуючись на дослідженнях з побудови вбудованих САРМ на базі FPGA [2–6], зокрема враховуючи праці B. Schrauwen, S. J. Melnikoff та ін. [2, 3], варто зазначити, що такі рішення ефективні для процесу цифрового оброблення сигналів і мають низький рівень споживання електроенергії. Проте вони не достатньо ефективні для оброблення чисел з плаваючою точкою та вирішення інших завдань процесу розпізнавання. Матриці FPGA мають низький рівень інтеграції, тому такі системи потребують додаткових зовнішніх модулів, зокрема аналого-цифрового перетворювача (АЦП) та модуля доступу до зовнішньої постійної пам'яті для збереження бази слів. Це, у свою чергу, призводить до підвищення собівартості та ускладнює проектування і розроблення цих систем.

Системи автоматичного розпізнавання мови з використанням DSP-мікропроцесорів дозволяють реалізовувати лише окремі перетворення на етапі цифрового оброблення звукового сигналу. Тому для реалізації повного процесу розпізнавання потрібно використовувати додаткові зовнішні модулі. Цей напрям розвинено в дослідженні A. Aldahoud з побудови гібридних DSP-FPGA систем розпізнавання [7]. Такі реалізації більш ефективні, проте мають недоліки, притаманні системам на базі FPGA (крім оброблення чисел з плаваючою точкою).

Варто відзначити також дослідження в галузі САРМ [8, 9] на базі одноплатних комп'ютерів. Такі системи керуються в основному операційною системою Linux, що має значні переваги під час розроблення САРМ завдяки достатній кількості програмних інструментів та бібліотек. Але такі системи потребують наявності дисплея або термінала для розроблення та керування, мають значно більшу вартість, показники споживання електроенергії та менший рівень відмовостійкості порівняно з апаратною (низькорівневою) реалізацією.

Найперспективнішою, на мою думку, є реалізація вбудованих систем розпізнавання мови на базі сучасних високопродуктивних RISC-мікроконтролерів завдяки збільшеним обчислювальним ресурсам, високому рівню інтеграції, низькому рівню енергоспоживання та невисокій собівартості.

Переваги і недоліки апаратної бази для вбудованих САРМ наведено в табл. 1.

Таблиця 1. Порівняльна характеристика апаратної бази для вбудованих САРМ

Апаратна база	Переваги	Недоліки
FPGA	Низький рівень енергоспоживання	Низький рівень інтеграції (необхідність підключення зовнішньої пам'яті та інших модулів)
DSP-процесори	Невисока собівартість. Низький рівень енергоспоживання	Низький рівень інтеграції. Вузькоспеціалізовані (реалізують тільки основні алгоритми процесу розпізнавання)
Одноплатні комп'ютери	Значний запас обчислювальних ресурсів. Велика ємність пам'яті Зручність програмування	Висока собівартість. Високий рівень енергоспоживання
Високопродуктивні мікроконтролери	Достатня кількість обчислювальних ресурсів. Високий рівень інтеграції. Низький рівень енергоспоживання. Невисока собівартість	Обмежена ємність пам'яті (можна розширити)

Мета роботи — дослідження перспектив реалізації вбудованих СРАМ на базі передових RISC-мікроконтролерів шляхом оцінювання найбільш продуктивної технічної бази та визначення способів оптимального використання ресурсів системи в процесі розв'язання задачі розпізнавання.

ОСНОВНИЙ МАТЕРІАЛ ДОСЛІДЖЕННЯ

Порівняльна характеристики RISC-мікроконтролерів на базі мікропроцесорного ядра Cortex®-M7

Оскільки ресурси мікроконтролерної системи для вбудованих рішень обмежені, розглянемо найбільш прогресивну архітектуру для використання та оцінимо її можливості.

Завдяки науково-технічному прогресу з'являються дедалі більш продуктивні мікропроцесорні системи. Найефективнішим ядром для мікропроцесорних систем натепер є Cortex®-M7 [14], розроблене ARM Holdings. Це ядро порівняно з аналогами дозволяє значно підвищити продуктивність побудованих на його базі мікроконтролерних систем не тільки в обчисленнях, але й у цифровому обробленні сигналів. Компанія-розробник позиціонує Cortex®-M7 для вбудованих рішень, зокрема у сфері розпізнавання мови, і передбачає, що підвищена продуктивність даного ядра дозволить підвищити швидкість аналізу звукової інформації для завдань розпізнавання [13]. Ядро включає DSP-модуль, що дозволяє ефективно використовувати можливості алгоритмів цифрового оброблення сигналів як на рівні даних, так і на рівні операцій, зокрема для вейлвет-перетворень та швидкого перетворення Фур'є.

Оскільки ліцензії на використання Cortex-M7 у своїх розробках мають лише три компанії (STMicroelectronics, NXP та Atmel), то аналізуємо їх рішення щодо використання ядра, розглянувши найбільш перспективні серії мікроконтролерів цих виробників — STM32F7 (STMicroelectronics), KV5x (NXP), SAM V (Atmel) [15–17] — та оцінимо їх характеристики.

Продуктивність. Мікроконтролери серії SAMV мають найвищі показники продуктивності порівняно з аналогами. За даними синтетичних тестів Dhrystone продуктивність зазначених мікроконтролерів досягає 645 DMIPS за максимальної тактової частоти ядра 300 МГц. Найнижчі показники за цим параметром мають мікроконтролери серії STM32F7 – 462 DMIPS за тактової частоти 216 МГц.

Ємність постійної пам'яті. Ємність постійної пам'яті мікроконтролерів усіх трьох серій становить не менше одного мегабайта, що є достатнім для розроблення та програмування більшості складних програмних систем. Проте серія KV5x удвічі поступається аналогам.

Швидкість доступу до постійної пам'яті. Крім ємності, важливим критерієм ефективності є швидкість зчитування даних з пам'яті, оскільки більшість типів flash-пам'яті не в змозі забезпечити безперервний доступ до даних за максимальної тактової частоти мікроконтролера.

Компанія STMicroelectronics використовує у своїй серії мікроконтролерів ART Accelerator власного виробництва, який забезпечує доступ до даних постійної пам'яті без затримок за максимальної тактової частоти, що дозволяє ефективно використовувати ресурси системи, запобігаючи простоям у роботі мікроконтролера.

У серії KV5x використовується 128-бітний інтерфейс, що мінімізує кількість станів очікування доступу до пам'яті під час виконання швидких контурів керування, проте повністю не запобігає таким ситуаціям.

Компанія Atmel у своїй серії контролерів забезпечує доступ з нульовим очікуванням лише для частини постійної пам'яті (для 384 кбайт з доступних 2 Мбайт).

Оперативна пам'ять. Мінімальну ємність оперативної пам'яті мають мікроконтролери серії KV5x, яка становить 256 кбайт, максимальна — мікроконтролери серії STM32F7 (512 кбайт).

Незважаючи на велику ємність вбудованої оперативної пам'яті (порівняно з більшістю мікроконтролерів), для задач розпізнавання такої ємності

може буди недостатньо. Тому важливо, щоб була можливість збільшувати ємність оперативної пам'яті за допомогою зовнішніх модулів. Виробники наведених серій мікроконтролерів забезпечили їх модулем доступу до зовнішньої пам'яті, який дозволяє розширювати оперативну пам'ять за допомогою окремих інтегральних схем.

У всіх цих серіях підтримуються модулі зовнішньої пам'яті типу SRAM. Серії STM32F7 і SAMV також підтримують модулі типу SDRAM, які значно дешевші і розраховані на більші ємності пам'яті порівняно з аналогами.

Зовнішня пам'ять. Оскільки для задачі розпізнавання необхідно зберігати великий обсяг даних, у тому числі бази слів, важливим показником розширюваності мікроконтролерної системи є можливість підключення зовнішньої постійної пам'яті. Мікроконтролери серій STM32F7 і SAMV мають спеціальні модулі, що підтримують SDIO-інтерфейс, який дозволяє підключати зовнішню пам'ять (SD і MMC) до мікроконтролера. За допомогою цих модулів і з використанням файлової системи FAT з'являється можливість зручного зберігання, записування та зчитування бази даних фонем і транскрипцій слів із зовнішньої карти пам'яті.

Мікроконтролер KV5x за цим параметром дещо поступається аналогам, оскільки не підтримує SDIO-інтерфейс. Зазвичай карти пам'яті підтримують доступ за допомогою SPI інтерфейсу, але така реалізація значно зменшить швидкість передавання даних і потребуватиме додаткових затрат обчислювальних ресурсів ядра мікропроцесорної системи.

Аналого-цифровий перетворювач. Цей модуль призначено для перетворення аналогового сигналу в дискретну форму, оскільки подальше оброблення даних у мікропроцесорній системі здійснюється в цифровій формі.

Мікроконтролери серій STM32F7 і SAMV мають у своєму складі АЦП модулі з максимальною розрядністю 12 біт. На відміну від них мікроконтролери серій KV5x мають також додатково модуль АЦП з розрядністю 16 біт.

Загалом для всіх серій максимальна частота вибірок за розрядності 12 біт становить понад 2 Мвибірок/с. Для 16-бітного режиму АЦП мікроконтролерів серії KV5x максимальна частота вибірок досягає 460 квибірок/с.

Рівень енергоспоживання. Цей параметр дуже залежить від тактової частоти системи та кількості запущених додаткових модулів. Найвищий рівень споживання мають мікроконтролери серії STM32F7 170 мА на частоті 216 МГц та з увімкненою всією периферією. Приблизно один рівень мають мікроконтролери серій KV5x і SAMV — їх рівень споживання становить 50 мА за тактової частоти 220 МГц з увімкненою периферією.

Ціна. Точно оцінити вартість різних серій мікроконтролерів важко, оскільки ціна конкретного екземпляра залежить від багатьох параметрів: ємності вбудованої пам'яті, периферійних пристроїв, кількості виводів тощо.

Ціна мікроконтролерів серії STM32F7 в середньому становить 6–9 дол. Вартість серій KV5x і SAMV оцінити не вдалось, оскільки на час написання роботи такі мікроконтролери не потрапили в продаж. Доступні тільки тестові макетні плати на їх основі.

Результати порівнянь розглянутих серій мікроконтролерів наведено в табл. 2.

Таблиця 2. Порівняльна характеристика серій мікроконтролерів на базі ядра Cortex-M7

Мікроконролери	STM32F7	KV5x	SAMV
Максимальна тактова частота ядра, МГц	216	220	300
Продуктивність, DMIPS	462	473	645
Максимальна ємність ПЗП, Мбіт	2	1	2
Максимальна ємність ОЗП, кбайт	512	256	384
Прискорення ПЗП	+	–	+ (частково)
Розрядність АЦП, біт	12	12/16	12
Підтримання SDIO	+	–	+
Підтримання зовнішньої оперативної пам’яті	+	+	+
Оцінний рівень енергоспоживання за робочої частоти 200 МГц, мА	155	45	45
Середня вартість, дол. США	7,5	–	–

Загальна схема процесу автоматичного розпізнавання мови

Для того щоб оцінити перспективи реалізації вбудованих САРМ за допомогою мікроконтролерної системи, необхідно розглянути загальну схему процесу розпізнавання (рис. 2), дослідити переваги підходів та визначити шляхи ефективної реалізації кожного етапу.

Рис. 2. Послідовність ключових етапів процесу розпізнавання мови

Оцифрування звукового сигналу

Оскільки звукові коливання являють собою аналоговий сигнал, то для оброблення їх мікропроцесорною системою необхідно перевести його у цифровий вигляд. На цьому етапі потрібно визначити оптимальну частоту вибірок та глибину дискретизації звукового сигналу, урахувавши при цьому можливість використання вбудованого АЦП-модуля мікроконтролера, щоб спростити проектування системи та отримати відповідний економічний ефект.

Частота дискретизації (вибірок) звукового сигналу. Надто мала частота вибірок призведе до недостачі інформації про характер сигналу та втрати відмінностей між деякими частинами мови. Так, наприклад, загальноприй-

нята частота вибірок звукового сигналу для завдань розпізнавання становить 8 кГц [18], проте в результаті досліджень виявлено, що за такої частоти дещо втрачається інформація про деякі фонеми української мови. Підвищивши частоту дискретизації до 16 кГц, можна зберегти додаткову інформацію, яка допоможе краще розпізнати конкретну частину мови в невизначеній ситуації. Проте надвелика частота вибірок призведе до надлишковості інформації, що потребуватиме більше обчислювальних ресурсів та пам'яті.

Глибина дискретизації (розрядність). Оскільки в процесі розпізнавання важливі тільки характерні значення форми звукового сигналу, то 8-бітної розрядності цифрового сигналу достатньо для вирішення цього завдання. Таким чином, можна мінімізувати потребу в оперативній пам'яті, обмежившись одним байтом на вибірку.

Отже, для збереження однієї секунди запису за мінімальних значень параметрів цифрового сигналу необхідно $\frac{8 \cdot 8000}{8} = 8000$ байт оперативної пам'яті за максимальних — $\frac{8 \cdot 16000}{8} = 16000$ байт.

Для оптимізації потреб в оперативній пам'яті необхідно встановити значення частоти дискретизації 12 кГц, затративши при цьому $\frac{8 \cdot 12000}{8} = 12000$ байт пам'яті для однієї секунди запису, тим самим зберегти баланс між достатньою інформативністю звукового сигналу і потрібною ємністю пам'яті для його зберігання.

Максимальна розрядність вбудованого АЦП більшості мікроконтролерів становить 12 біт. Цей модуль спроможний забезпечити частоту вибірок понад 16 кГц. Тому можливостей вбудованого АЦП усіх серій мікроконтролерів достатньо для якісного оцифрування звукового сигналу.

Розбиття на фрейми

Оскільки форма мовного сигналу на коротких часових проміжках може повторюватись і кожен такий проміжок має певні характерні ознаки, то такий сигнал розбивають на фрейми. Ширина фрейму повинна охоплювати одну повторювану частину. Дослідним шляхом встановлено ширину 5 мс (рис. 3).

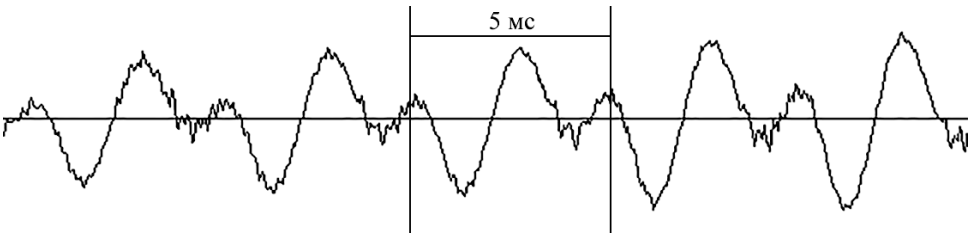

Рис. 3. Період звукової хвилі фонеми

Оскільки початок фрейму не завжди збігається з початком конкретної базової одиниці (наприклад, фонеми) з бази зразків, то фрейми накладають один на одного з деяким зсувом. Таку систему фреймів можна розглядати як «плаваюче вікно» відносно вхідного звукового сигналу.

Накладати фрейми один на одного необхідно з мінімальним «захопленням» — 50% від розміру фрейму. Чим більшим буде таке «захоплення», тим точніше можна визначити відповідність фрейму конкретній базовій одиниці. Максимальний розмір «захоплення» доцільно встановити в межах 90%, оскільки подальше уточнення не призводить до значного підвищення ефективності розпізнавання.

Проте збільшення кількості фреймів зумовить підвищення витрат обчислювальних ресурсів і пам'яті. Так, на збереження одного фрейму за оптимальних значень частоти і глибини вибірок сигналу (оцифрування звукового сигналу) потрібно затратити $\frac{8 \cdot 12000 \cdot 0{,}5}{8} = 6000$ байт оперативної пам'яті. Умовно визначимо середню довжину слова рівною 0,5 с. Тоді за мінімального накладання на слово припаде $\frac{500 \cdot 2}{5} = 200$ фреймів, що потребуватиме близько 11,7 кбайт оперативної пам'яті, за максимального — 58,6 кбайт.

Можна дійти висновку, що ресурсів мікроконтролерної системи достатньо для збереження навіть максимальної кількості фреймів, проте збільшення кількості фреймів спричинить також додаткові затрати обчислювальних ресурсів. Тому доцільно встановити накладання фреймів на рівні 70%, щоб зберегти баланс між якісним рівнем розпізнавання та затратами ресурсів системи.

Спектральне зображення сигналу

Інформації про амплітуду і форму сигналу не достатньо для виділення з мовного сигналу лексичних елементів. Залежно від різних обставин форма звукового сигналу може змінюватись у широких межах. Спектральне зображення звукового сигналу є одним з найбільш важливих інструментів аналізу й оброблення мовного сигналу, оскільки, крім важливої інформації у звуковому сигналі, наявні й інші елементи, які не є значущими для вирішення цього завдання і негативно впливають на процес розпізнавання.

Спектральне зображення сигналу ілюструє рис. 4.

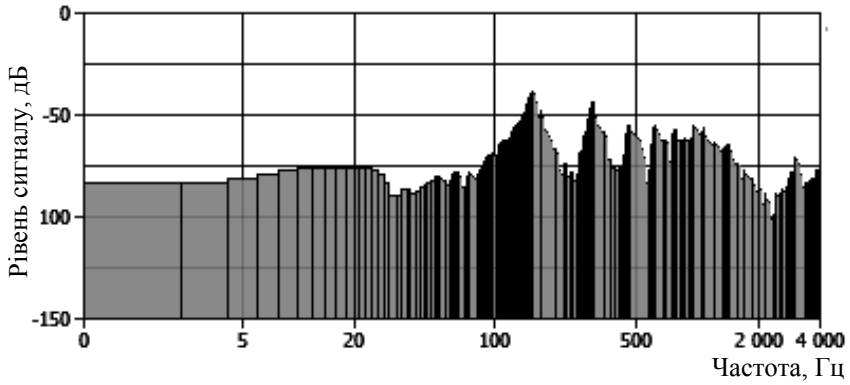

Рис. 4. Спектральне зображення сигналу

Для спектрального зображення сигналу використовують здебільшого *перетворення Фур'є* [19], яке полягає в розкладанні ряду на функції синусів

та косинусів різних частот для виявлення найбільш значущих із них. Для зменшення затрат процесорного часу розроблено алгоритми швидкого перетворення Фур'є (FFT), які дозволяють отримати амплітудний спектр та інформацію про фазу сигналу, «проріджуючи» дані за часом та частотою.

Перевагою такого методу є те, що його реалізація у вигляді FFT є досить простою і не потребує значних затрат ресурсів системи. Проте для аналізу звукової інформації перетворення Фур'є має низку недоліків, які спричиняють втрату інформації про часові характеристики оброблюваних сигналів, оскільки цей метод припускає використання штучних прийомів, за допомогою яких здійснюється частотно-часова локалізація.

Для спектрального зображення сигналу також використовують *вейвлет-аналіз,* за допомогою якого вхідний сигнал розкладається в базис функцій, що характеризують частоту і час. Таким чином, метод дає змогу проаналізувати властивості звукового сигналу одночасно у фізичному та частотному просторах.

Вейвлет-перетворення дозволяють уникнути недоліків, притаманних перетворенню Фур'є [19], зберігаючи при цьому всі його переваги. Недоліком вейвлет-аналізу є відносна складність перетворення, що накладає додаткові витрати ресурсів мікроконтролерної системи. Крім того, для отримання оптимальних алгоритмів перетворення розроблені певні критерії, які не можна вважати остаточними, бо вони не враховують зовнішніх критеріїв, пов'язаних із сигналами та цілями їх перетворення. Тому під час практичного використання вейвлет-перетворень належить приділяти достатню увагу перевірці їх працездатності та ефективності порівняно з іншими методами.

Оскільки DSP-модуль мікроконтролера зазвичай оптимізований для FFT перетворень, то використання перетворень Фур'є є найбільш оптимальним для розпізнавання за допомогою мікроконтролерної системи.

Виділення мовних одиниць

Виділення слів як базової одиниці. У найпростішому випадку за базову одиницю обирають слово або навіть цілу фразу. Для виділення такої одиниці потрібно визначити граничне значення, яке відділяє слово (команду) від тиші. Для цього необхідно визначити фіксоване граничне значення рівня сигналу або кластеризовані значення сигналу (для виділення множини значень, що відповідають тиші). Проте такі методи не досить точні. Для більш якісного результату потрібно розрахувати інформаційну двійкову ентропію за допомогою такої формули:

$$H(x) = -\sum_{i=1}^{n} p(i)\log_2 p(i),$$

де x — незалежні випадкові стани; i — конкретний стан системи, n — кількість розглянутих варіантів; p — апріорна ймовірність.

Тут ентропія буде вказувати на те, на скільки сильно змінюється сигнал в межах одного фрейму.

Перевагою цього методу є те, що за його допомогою можна досить просто реалізувати командні системи розпізнавання з обмеженим розміром словника (одиниці або десятки команд).

Серед недоліків методу варто відзначити такі:

– у разі «просідання» сигналу на середині слова неможливо правильно визначити його межі;

– за допомогою методу можна розпізнавати тільки роздільне мовлення (не придатний для систем розпізнавання суцільного мовлення);

– такі системи не зручні для користувача, оскільки потребують неприродного способу спілкування — з паузами між словами;

– метод придатний для розпізнавання лише невеликого набору команд, оскільки збільшення їх кількості зумовлює значні витрати обчислювальних ресурсів системи.

Виділення частини мови як базової одиниці. Оскільки командні системи дозволяють розпізнавати лише роздільне мовлення, то в системах розпізнавання суцільного мовлення доцільно використовувати мінімальні частини мови (фонеми, алофони) як базової мовної одиниці, що дозволяє розробляти гнучкі та дикторонезалежні САРМ із середнім та великим розмірами словника. У простому випадку достатньо зробити базу фонем із записів голосу різних дикторів та базу слів.

Проте для реалізації таких систем потрібно використовувати складні алгоритми визначення ймовірних послідовностей мовних одиниць. Крім того, існують ситуації, коли неможливо правильно визначити конкретну мовну одиницю у звуковому потоці, у результаті чого багато років ведуться дослідження в цьому напрямі.

Основні методи розпізнавання мови

Алгоритм динамічної трансформації часової шкали (*DTW*). Цей метод був одним із перших, що використовувався для автоматичного розпізнавання мови [20]. Суть його полягає в розтягуванні або звужуванні часових рамок звукового сигналу, щоб зробити його схожим за характеристиками із зразком з бази даних (рис. 5).

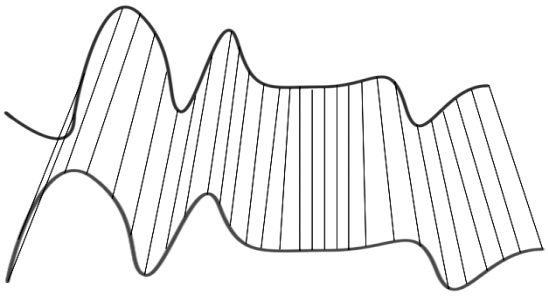

Рис. 5. Зіставлення сигналу за допомогою DTW

Шкала DTW добре підходить для задач розпізнавання ізольованих слів з малим розміром словника (наприклад, голосового набору на мобільних телефонах). Цей алгоритм найменш ресурсомісткий, тому придатний для вбудованих систем з обмеженими обчислювальними ресурсами.

Серед недоліків DTW-алгоритму слід відзначити, що в деяких випадках він може видавати неправильні результати (може спробувати визначити непостійність осі y за допомогою осі x). Це може зумовити вирівнювання, за якого одній точці першої послідовності відповідає велика підгрупа точок іншої послідовності. Інший недолік методу — алгоритм може не знайти очевидного вирівнювання двох рядів унаслідок того, що особлива точка од-

ного ряду міститься дещо вище або нижче відносно відповідної їй точки іншого ряду.

Приховані марковські моделі (ПММ). Метод прихованих марковських моделей (рис. 6.) застосується здебільшого для розпізнавання безперервного мовлення, коли модель стану системи невідома [21].

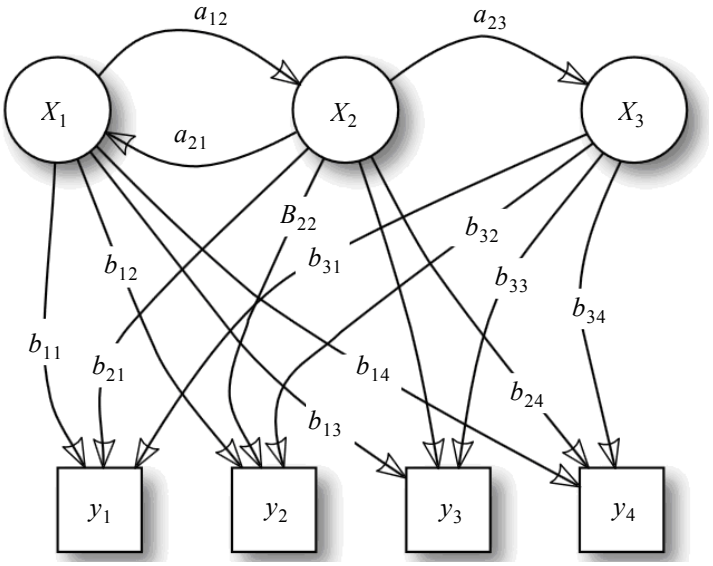

Рис. 6. Імовірнісні параметри прихованої марковської моделі: X — стани; y — можливі спостереження; a — імовірності переходів станів; b — імовірності виходів

У задачах розпізнавання окремі ПММ асоціюються з базовою частиною мови (фонемою, алофоном тощо). Далі виконуються обчислення для кожної моделі, щоб визначити, яка послідовність ПММ найбільш точно відповідає вхідному мовному сигналу. Оскільки ПММ еквівалентні частинам мови, які необхідно визначити, то процес розпізнавання зводиться до визначення послідовності ПММ.

Переваги методу:

– ефективне моделювання часових та спектральних варіацій мовного сигналу;

– легкість використання фонологічних та синтаксичних правил;

– розроблено навчальні алгоритми, що забезпечують ефективне навчання у разі великих розмірів словника;

– існують дикторонезалежні алгоритми розпізнавання як ізольованих слів, так і безперервного мовлення.

Серед недоліків слід відзначити такі:

– марковська модель покладається на модель першого порядку, тобто її стан у певний момент часу залежить тільки від попереднього стану системи;

– відсутність ефективних моделей тривалості станів та їх реалізації в межах ПММ;

– навчання лінгвістичної моделі відбувається окремо від акустичних моделей;

– слабкі дискримінантні можливості;

– частково-сталий характер моделі — кожний стан має стаціонарну статистику.

Ці недоліки значно обмежують можливості методу ПММ.

Штучні нейронні мережі (ШНМ). Нейронні мережі використовуються в задачах розпізнавання мови схожим із ПММ способом [22]. За основу процесу розпізнавання береться спектральне зображення фрейму звукового сигналу у вигляді набору чисел. Кількість вхідних і вихідних нейронів невідома. Кожен із вхідних нейронів відповідає одному набору чисел. На вихідному рівні існує тільки один нейрон, вихід якого відповідає бажаному значенню слова, що розпізнається (рис. 7). Нейронна мережа, яка має максимальне вихідне значення, є відповідно мережею розпізнавання цього слова. А таке слово вважається результатом розпізнавання.

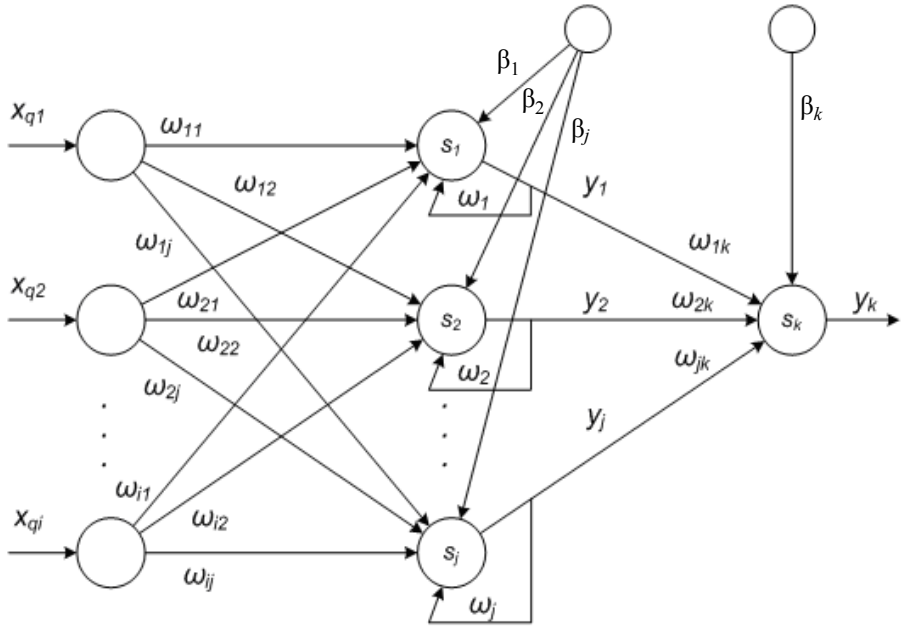

Рис. 7. Структура нейронної мережі зі зворотним зв'язком: x_{qt} — *i*-е вхідне значення *q*-го набору чисел; y_j — вихід *j*-го нейронного шару; ω_{ij} — ваговий коефіцієнт зв'язку, що об'єднує *i*-й та *j*-й нейрони; ω_j — ваговий коефіцієнт зворотного зв'язку *j*-го нейрона; β_j — зміщення *j*-го нейронного шару

Основною перевагою використання ШНМ є можливість їх навчання. Штучні нейронні мережі стійкі до шумів, які трапляються у вхідному звуковому сигналі, і легко адаптуються до змін навколишнього середовища. Слід відзначити потенційно високу відмовостійкість апаратної реалізації ШНМ.

Недоліки використання ШНМ для задач розпізнавання мови:

– не мають механізмів, які б адекватно відображали часову варіативність і послідовну природу мовного сигналу;

– не існує теоретичних основ для цілого ряду параметрів, що визначають динаміку і топологію ШНМ;

– більшість підходів до проектування ШНМ є евристичними і часто призводять до неоднозначних розв'язків;

– процес навчання є досить ресурсомістким процесом.

Хоча нейронні мережі були джерелом великої кількості досліджень, метод ПММ витісняє їх у задачах розпізнавання мови для вбудованих систем, передусім через потребу у великій кількості ресурсів для реалізації та навчання ШНМ. Так, наприклад, словник з кількістю 60 тисяч слів потребує оперативної пам'яті ємністю сотні мегабайтів.

ВИСНОВКИ

Проаналізовано найбільш продуктивну і технологічну базу для розроблення вбудованих САРМ та шляхів ефективного розв'язання задачі автоматичного розпізнавання мови в умовах обмежених ресурсів.

У процесі дослідження виявлено, що можливостей сучасних RISC-мікроконтролерів достатньо для оптимального розв'язання більшості задач розпізнавання.

Найпростішою та найменш ресурсомісткою реалізацією вбудованих САРМ є командні системи, побудовані на базі DWT-методу зі словом як базовою мовною одиницею. Проте такі системи залежні від диктора і не досить зручні для кінцевого користувача, оскільки не можуть розпізнавати суцільне мовлення.

Найбільш складною та ресурсомісткою для вбудованих систем є реалізація САРМ за допомогою ШНМ.

Оптимальним для розпізнавання суцільного мовлення у вбудованих дикторонезалежних системах є ПММ метод. Проте недоліки цього методу спонукають до проведення додаткових досліджень для пошуку шляхів підвищення якості розпізнавання в умовах обмежених обчислювальних ресурсів.

ЛІТЕРАТУРА

1. *Мартинюк І.А*. Актуальність та основні проблеми реалізації технологій автоматичного розпізнавання мови для вбудованих систем / І.А. Мартинюк, В.А. Лахно // Інформаційна безпека та комп'ютерні технології: зб. тез доп. міжнар. наук.-практ. конф., 24–25 берез. 2016 р. — Кіровоград: КНТУ, 2016. — С. 112–113.
2. *Compact* hardware liquid state machines on FPGA for real-time speech recognition [Text] / [B. Schrauwen, M. D'Haene, D. Verstraeten et al.] // Neural Networks. — 2008. — **21**, № 2–3. — P. 511–523.
3. *Speech* recognition on an FPGA using discrete and continuous Hidden Markov Models [Text] / S.J. Melnikoff, S.F. Quigley, M.J. Russell // Lecture Notes in Computer Science. — 2002. — № 2438. — P. 202–211.
4. *Pan Shing-Tai*. The implementation of speech recognition systems on FPGA-based embedded systems with SOC architecture [Text] / Shing-Tai Pan, Chih-Chin Lai, Bo-Yu Tsai // International journal of innovative computing, information and control. — 2011. — **7**, № 11. — P. 6161–6175.
5. *Hu X*. Isolated word speech recognition system based on FPGA [Text] / X. Hu, H. Zhang, L. Zhan et al // Journ. of Computers. — 2013. — **8**, № 12. — P. 3216–3222.
6. *Li J*. Embedded speaker recognition system design and implementation based on FPGA [Text] / J. Li, D. An, L. Lang et al. // Procedia Engineering. — 2012. — **29**. — P. 2633–2637.

7. *Aldahoud A*. Robust automatic speech recognition system implemented in a hybrid design DSP-FPGA [Text] / A. Aldahoud, H. Atoui, M. Fezari // International Journal of signal processing, image processing and pattern recognition. — 2013. — **6**, № 5. — P. 333–342.
8. *Bourke P.J.* A Low-Power Hardware Architecture for Speech Recognition Search : thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy / P.J. Bourke. – Pittsburgh, PA : Carnegie Mellon University, 2011. — 166 p.
9. *Wei Z.* Embedded system for speech recognition and image processing [Text] / Z. Wei, J. Liang // Journal of electrical and electronic engineering. — 2015. — **2**, № 6. — P. 89–93.
10. *Varshney N*. Embedded speech recognition system [Text] / N. Varshney, S. Singh // International journal of advanced research in electrical, electronics and instrumentation energy. — 2014. — **3**, № 4. — P. 9218–9227.
11. *Moving* Speech Recognition from Software to Silicon: the In Silico VoxProject : research report (final) : 888.012 / C.L. Edward, K. Yu, R.A. Rutenbar, C. Tsuhan. — Pittsburgh: Carnegie Mellon University, 2006. — 4 p.
12. *Suryawanshi U.J*. Hardware implementation of speech recognition using mfcc and euclidean distance [Text] / U.J. Suryawanshi, prof. dr. S.R. Ganorkar // International journal of advanced research in electrical, electronics and instrumentation engineering. — 2014. — **3**, № 8. — P. 11248–11254.
13. *Lippman R.P*. An introduction to computing with neural nets [Text] / R.P. Lippman // IEEE Acoustics, speech and signal processing magazine. — 1987. — **4,** № 2 — P. 4–22.
14. *Cortex-M* Series [Electronic resource]. — Access mode: http://www.arm.com/ products/ processors/cortex-m/
15. *STMicroelectronics* STM32F7 Series [Electronic resource]. — Access mode: http://www.st.com/web/en/catalog/mmc/FM141/SC1169/SS1858/
16. *NXP* Kinetis V Series [Electronic resource]. — Access mode: http://www.nxp.com/products/microcontrollers-and-processors/arm-processors/kinetis-cortex-m-mcus/v-series
17. *Atmel* SAM V Series [Electronic resource]. — Access mode: http://www.atmel.com/ru/ru/products/microcontrollers/arm/sam-v-mcus.aspx
18. *Пресняков И.Н*. Автоматическое распознавание речи в каналах передачи [Текст] / И.Н. Пресняков, А.В. Омельченко, С.В. Омельченко // Радиоэлектроника и информатика. — 2002. — № 1. — С. 26–31.
19. *Зубаков А.П*. Фурье и вейвлет-преобразования в проблеме распознавания речи [Текст] / А.П. Зубаков // Вестн. Тамб. ун-та. Серия: Естественные и технические науки. — 2010. — **15**, № 6. — С. 1893–1899.
20. *Мещеряков Р.В*. Структура систем синтеза и распознавания речи / Р.В. Мещеряков // Известия Томск. политехн. ун-та. — 2009. — № 5. — С. 121–126.
21. *Rabiner L.R*. A tutorial on hidden Markov models and selected applications in speech recognition [Text] / L.R. Rabiner // Proceedings of the IEEE. — 1989. — **77**, № 2. — P. 257–286.
22. *Алимурадов А.К*. Обзор и классификация методов обработки речевых сигналов в системах распознавания речи [Текст] / А.К. Алимурадов, П.П. Чураков // Измерение. Мониторинг. Управление. Контроль. — 2015. — № 2. — С. 27–35.

Надійшла 31.05.2016