УДК 534.781:004.934.1

Г. О. Добрушкін, асп.;
В. Я. Данилов, д. т. н., проф.

ОСНОВНІ ПІДХОДИ ДО РОЗПІЗНАВАННЯ МОВЛЕННЄВОЇ ІНФОРМАЦІЇ (ЧАСТИНА 1)

Розглянуто етапи первинної обробки мовленнєвих сигналів. Класифіковано системи розпізнавання мови. Проаналізовано різноманітні підходи до обробки і розпізнавання мовленнєвої інформації. Обговорено методи видалення шумів з мовленнєвого сигналу, його сегментації та параметризації. Порівняно методи еталонного розпізнавання, фонемно-орієнтовний метод, застосування експертних систем, прихованих Марківських та змішаних Гаусівських моделей.

Вступ

Мова є найбільш природною формою людського спілкування і тому реалізація інтерфейсу на основі аналізу мовленнєвої інформації є перспективним напрямком розвитку інтелектуальних систем управління. Однією з актуальних невирішених проблем у галузі інформаційно-вимірювальних систем є побудова систем автоматичного розпізнавання мовленнєвих сигналів, інваріантних до диктора. Її вирішення дало б змогу розширити коло користувачів таких систем і значно підвищити ефективність обміну інформацією в людино-машинних системах.

Реалізація мовного інтерфейсу є дуже складною технічною задачею, розв'язання якої знаходиться на стику багатьох галузей науки. Так при сприйнятті мови людина використовує механізми асоціативного аналізу, не просто розбираючи і порівнюючи почуті звуки, але збираючи фонеми в словесні образи, підбираючи найбільш відповідні слова не тільки по звуковій подібності, але і по інтонації, емоційному забарвленню, контексту слова, фрази, речення і навіть всього тексу. Тому, людина здатна розпізнавати мову навіть при великому браку несучої інформації. Наприклад, людина набагато вимогливіша до якості звуку при прослуховуванні тексту на чужій мові яку вона погано знає, ніж при сприйнятті рідної мови.

Ще не настав час безпосереднього впровадження мовного інтерфейсу в повсякденне життя кінцевого користувача, однак наявний на даний час прогрес важко переоцінити. Програми і системи, що володіють засобами мовного введення інформації, одержують усе більше поширення, але, з огляду на всі їх недоліки, варто розглядати перспективи розвитку вузькоспеціалізованих систем, що мають чітке застосування:

— системи контролю присутності людини;
— аутентифікація та контроль доступу;
— телефонний банкінг;
— голосовий ввід інформації, що замінює текстовий набір;
— автоматизовані системи заповнення анкет та шаблонних інформаційних листів;
— біометрична реєстрація в різноманітних і різнонаправлених телефонних системах;
— розробка інформаційно-довідкових служб різного призначення, в яких клієнт запитує довідки, данні, що його цікавлять й одержує інформацію в мовній або іншій формі; телефоні лінії підтримки клієнтів, електронна комерція;
— управління системами життєзабезпечення для людей з обмеженими фізичними можливостями та побудови систем інтелектуалізації житла, так звані «розумні дома»;
— управління освітленням, водопостачанням, опалюванням, кондиціонуванням повітря тощо;
— створення індивідуальних автоматичних систем перекладу з одної мови на іншу, що працює у реальному часі;
— судові експертизи, зокрема системи відтворення спотворених та зашумлених мовних повідомлень;
— в майбутньому — пошук в інформаційних мережах мовної інформації за заданими ключовими словами або проблематикою.

Метою статті є систематизація наявних на сьогодні підходів до первинної обробки мовленнєвого сигналу, розв'язку задачі розпізнавання мовленнєвої інформації та представлення алгоритмів, що дозволяють підвищити швидкість та надійність розпізнавання мовленнєвих сигналів. Для досягнення цієї мети в цій частині статті розв'язуються такі задачі:

— відокремлення один від одного, систематизація і конкретизація етапів первинної обробки вхідного мовленнєвого сигналу;

— аналіз відомих методів реалізації кожного етапу первинної обробки сигналу;

— формулювання чітких властивостей, якими мають бути наділені нові алгоритми первинної обробки мовленнєвого сигналу;

— систематизування методів розпізнавання обробленого мовленнєвого сигналу.

В другій частині статті будуть описані нові методи первинної обробки і розпізнавання мовленнєвого сигналу, що ґрунтуються на вейвлет-перетворенні, перетворенні Фур'є, штучних нейронних мережах і новітнього апарату машинного навчання — штучних імунних системах.

Постановка задачі

Задача розпізнавання мови полягає в точному і ефективному, в контексті алгоритму класифікації, відтворенні вимовленого мовленнєвого сигналу. В підходах, що використовуються сьогодні, її рішення полягає в послідовному порівнянні з еталонами, що задані словником системи розпізнавання мови. Звісно словником можуть виступати різноманітні фонеми природної мови, що робить можливим побудову системи розпізнавання мови навіть без словника у прямому розумінні цього слова. Словник може лише допомагати виправити помилки розпізнавання.

З точки зору інформаційної ентропії, задача побудови ефективної дикторонезалежної стратегії розпізнавання мови може бути сформульована як задача пошуку оптимального за загальносистемним критерієм дерева рішень, в якому на кожному кроці класифікації з апріорного алфавіту вибирають підмножину ознак, що максимально зменшує на досягнутому кроці ентропію про образ і збільшує швидкість класифікації. Така стратегія передбачає використання множинного описання слів у термінах різних фонетичних класів, що відповідають різним рівням дерева класифікації, а також вибору інформативних дикторонезалежних ознак для виділення фонетичних класів на кожному рівні.

Коли говорять про розпізнавання мовленнєвої інформації, об'єктом дослідження вважають власне процес оброблення інформації про мовленнєві сигнали, а предметом дослідження вважається технологія аналізування, оброблення і параметризації мовленнєвої інформації. Процеси видалення шуму, сегментації і виявлення вокалізованих ділянок є власне обробкою мовленнєвих сигналів, тобто також є предметом дослідження.

Класифікація систем розпізнавання мови

Автоматизовані системи розпізнавання мови можуть бути класифіковані за багатьма ознаками: за типом мови, за множиною дикторів, за об'ємом та за повнотою словника, що необхідно розпізнавати.

За типами мову поділяють на дискретну і неперервну. Дискретною називають таку мову в якій паузи між словами значно більші за природні паузи всередині слів, наприклад, надиктування окремих команд. Неперервна мова в свою чергу не має значної паузи між словами. Природний людський режим спілкування є неперервною мовою.

За ознакою множини дикторів системи розпізнавання мови поділяють на дикторозалежні, тобто такі, якість розпізнавання яких залежать від індивідуальних особливостей вимовляння диктора та дикторонезалежні. Дикторозалежні системи розпізнавання звичайно розпізнають мовленнєву інформацію сказану не тільки одним конкретним диктором, просто ймовірність вдалого розпізнавання слова, сказаного одним конкретним диктором, вища за середню ймовірність вдалого розпізнавання цього слова, сказаного іншими дикторами. У відповідності до цього положення дикторонезалежною називають таку систему, ймовірність вдалого розпізнавання слова якою однакова для усіх дикторів.

За об'ємом словника системи розпізнавання мови класифікуються на дві категорії: системи з малими та системи з великими словниками. Ці системи значно відрізняються одна від одної. Так, систему з малим словником можна навчити послідовно, вимовляючи кожне слово зі словнику.

Систему з великим словником необхідно навчати синтезованими акустичними ознаками слів (фонеми, трифони), адже в цьому випадку неможливо надиктувати системі весь словник.

Повнота словника полягає в тому, що кожне слово вхідного мовленнєвого сигналу має бути присутнім в словнику. Як правило повний словник мають лише системи розпізнавання мови з малим словником.

Первинна обробка мовних сигналів

Первинна обробка мовних сигналів полягає у видаленні шумів, виокремлення значимої для розпізнавання інформації, усунення варіантності диктора і навколишнього середовища, стиснення сигналу, сегментація на фонеми.

Методи аналізу мовних сигналів, тобто безпосереднього виділення ознак, поділяються на три групи [1]:

1. Фонетичні методи, які спираються на теорію мовотворення. Суть їх полягає у виділенні ознак, що характеризують спосіб артикуляції. Фонетичні методи аналізу мовних сигналів можуть розглядатися як перший рівень розпізнавання мови, тому що більшість з них засновано на деяких перетвореннях первинних ознак мовних сигналів.

2. Неакустичні методи, що по своїй меті примикають до фонетичних і полягають у виокремленні інформації про процеси, що супроводять артикуляцію.

3. Параметричні методи, що засновані, по-перше на представленні мовного сигналу як реалізації деякого процесу в часі і, по-друге, виокремлення деяких параметрів цього процесу, найчастіше пов'язаних з його спектральними характеристиками. Відомими параметричними методами аналізу мовленнєвої інформації є:

— спектрально-смугові, кореляційні і ортогональні методи;

— цифрові фільтри;

— методи обчислення спектра за допомогою швидкого перетворення Фур'є;

— методи, що застосовують вейвлет-перетворення для моделювання мовних сигналів

— методи лінійного передбачення мови (ЛПМ);

— методи, пов'язані з виділенням миттєвої частоти переходів через нуль;

— часові методи, засновані на аналізі розподілу тривалості інтервалів між переходами через нуль або екстремумами мовного сигналу;

— використання нелінійного перетворення і фазові співвідношення мовного сигналу;

Більшість з цих методів заснована на загладжувані сигналу та обчисленні спектру, чи кепстру.

Як правило, якісна первинна обробка мовленнєвої інформації включає в себе 5 етапів: видалення шуму, сегментація, виокремлення вокалізованих ділянок, вимір частоти основного тону та параметризація.

Існує багато методів, що реалізують ці етапи. Деякі методи можуть одночасно реалізовувати декілька етапів первинної обробки сигналів.

Для розпізнавання мови, як ізольованої, так і злитої, необхідно заздалегідь визначити її межі в контексті навколишніх даремних звукових сигналів. Складність визначення меж мови пов'язана з особливостями вимови конкретного диктора, наявністю в мовному сигналі різних видів сторонніх шумів, а також звукових артефактів процесу артикуляції (придих, чмокання і т.п.).

Методи, засновані на обчисленні короткочасної енергії сигналу, спектральної енергії та кількості нуль-перетинів нестійко працюють в умовах, коли з'являються шуми з динамічним спектром або відносно сильні стаціонарні шуми.

В будь-якій мові існує деякий набір звуків, який бере участь при формуванні звукового обрису слів. Як правило, звук поза мовою не має значення, він набуває його лише як складова частина слова, допомагаючи відрізнити одне слово від іншого. Елементи цього набору звуків називаються фонемами.

Звуки, що беруть участь у формуванні мови, мають дві основні класифікації: за ознаками артикуляцій і за акустичними ознаками.

Класифікація звуків за ознаками артикуляцій (екскурсія, витримка, рекурсія) є вкрай важливою при використанні методів розпізнавання мови за допомогою моделювання носоглотки, але для вирішення завдань ділення на фонеми цікавіший розгляд акустичних відмінностей звуків. За акустичними ознаками звуки підрозділяються на:

Тональні звуки — утворюються голосом при повній відсутності шумів, що забезпечує хорошу чутність звуку. Тональними звуками є усі голосні.

Сонорні (звучні) — чия якість визначається характером звучання голосу, який грає головну роль в їх формуванні, а шум бере участь в мінімальному ступені. Сонорними приголосними є: й, л, р.

Шумні — їх якість визначається характером шуму. Шумні приголосні поділяються на дзвінкі тривалі (в, з, ж), дзвінкі миттєві (вибухові) (б, д, г), глухі тривалі (ф, с, ш, х) та глухі миттєві (вибухові) (п, т, к).

Різниця між звуками різних видів є дуже великою і теоретично, це значно полегшило-б завдання поділу звуків, але в реальних умовах, коли ми маємо справу зі злитою мовою (мовним ланцюгом), етап екскурсії артикуляції наступного звуку накладається на рекурсію, чи навіть на етап витримки артикуляції попереднього звуку, що нівелює цю різницю.

Видалення шуму з мовних сигналів

В загальному випадку видалення шуму полягає в виокремленні значущого мовного сигналу від фонового шуму, та видаленні визначеного шуму з вхідного сигналу. Фоновий шум буває стаціонарний і нестаціонарний.

В [2] ділянки мовленнєвого сигналу, на які накладено *стаціонарний* шум *знаходять* за допомогою дискретного перетворення Фур'є та імовірнісних розподілів шуму та мови. Спочатку записується тільки шум і оцінюються параметри його рівномірного та Гауссівського розподілу. Потім, за допомогою порогової умови на щільність ймовірності нормального розподілу, із вхідного мовного сигналу виділяють ділянки з можливою сумішшю корисного сигналу і шуму. Для отриманих ділянок сигналу знову оцінюють параметри Гауссівського розподілу і, ґрунтуючись на теоремі Байеса, приймають рішення про наявність корисного сигналу в виокремленній ділянці.

Виокремлення ділянок з корисним сигналом, на який накладається *нестаціонарний* шум в [2] вирішують представленням шуму і мови в якості статистичних моделей прихованої Марківської мережі.

У загальному випадку *фільтрація* сигналу полягає в тому, щоб виокремити корисну складову сигналу і видалити сторонні шуми і спотворення. В залежності від характеру шуму виникає декілька завдань:

— Необхідно виокремити корисний сигнал з високочастотного або смугового шуму. В цьому випадку фільтрація полягає у виборі типу фільтру і розрахунку його параметрів [3];

— Потрібно виокремити мовний сигнал з мово-подібного шуму. Наприклад, два або більше дикторів можуть говорити одночасно, а потрібно отримати розбірливу мову тільки одного з них. Це одне з найскладніших завдань фільтрації, загальних методів рішення якої поки не існує;

— Потрібно відновити сигнал, що зазнав нелінійних спотворенн. Це завдання виникло з появою цифрових телефонних ліній, які ущільнюють і тим самим спотворюють початковий сигнал.

Видалення стаціонарного шуму в [2] реалізують фільтрацією компонент ДПФ вхідного сигналу синтезованими нерекурсивними фільтрами.

Задача *видалення* нестаціонарного шуму, в загальному випадку, поділяється на два великих класи задач, в залежності від того, чи відома нам заздалегідь фізична модель нестаціонарного шуму.

Перший напрям ґрунтується на тому, що інформації про фізичний процес генерації нестаціонарного шуму досить для побудови його моделі, що дозволяє за деякими параметрами відрізняти компоненти шуму від компонентів корисного сигналу.

Другий напрям використовують в випадках, коли фізичну модель шуму побудувати не можливо. Розвиток цього напряму ґрунтується на тому, що фільтр проходить стадію навчання, де за допомогою реалізацій незашумленої мови створюються її стани. Процес фільтрації схожий з процесом розпізнавання, де визначається стан мови, з максимальною ймовірністю схожий на ділянку зашумленої мови, і який згодом замінює собою цю ділянку. Оскільки варіативність мови від диктора до диктора висока, то така фільтрація припускає або шумоочистку сигналу наперед відомого диктора, яким і проводилося навчання фільтру, або створення достатньо великого банку голосів дикторів і відповідних ним станів в надії, що вдасться перебрати всі типи голосів.

В [2] можна знайти моделі для видалення деяких нестаціонарних шумів першого класу.

Сегментація мовних сигналів

Сегментація мовного сигналу полягає у виділенні ділянок сигналу, що відповідають окремим структурним одиницям мовного сигналу. Якщо в якості таких одиниць розглядати фонеми, то завдання сегментації зводиться до виявлення міжфонемних переходів. В рамках традиційних підходів розв'язання цієї задачі вельми проблематично.

В системах розпізнавання мови для визначення меж мови традиційно використовуються методи (наприклад, Voice Activity Detector), засновані на обчисленні короткочасної енергії сигналу або спектральної енергії. Крім того, додатково застосовуються методи, що використовують кількість нуль-перетинів сигналу і інформацію про тривалість мовних фрагментів. Недоліком цих алгоритмів є ненадійності в умовах нестаціонарного шуму, а також при виникненні різних звукових артефактів (придих, чмокання і таке інше). Також існують алгоритми, засновані на адаптивних порогових значеннях, але при виникненні звукових артефактів, а також відносно високому рівні шуму або незначному рівні корисного сигналу вони також стають нестійкими.

Надійний метод сегментації мовленнєвої інформації має задовольняти такі вимоги:

— забезпечення мінімальної вірогідності помилкового спрацювання при дії тільки шуму з високим рівнем;

— висока вірогідність правильного виділення мови навіть в умовах сильного шуму;

— висока швидкодія для виключення затримок включення і виключення алгоритму розпізнавання мови.

Для практичних цілей кожна фонема може бути представлена квазістатичним спектром, в якому передавальна функція, не змінюється в часі [4]. Явища, що обумовлені швидкими змінами функції джерела, можуть служити для розмежування окремих фонем в мовному потоці. Різкі зміни передівальної функції, що пов'язані з швидкою зміною положення артикулюючих органів, також вказують на межу (початок або кінець) фонеми. Звичайно, мінімальна швидкість зміни повинна бути визначена експериментально для кожного випадку. Додатковим засобом для визначення межі фонеми є швидкі флуктуації загальної інтенсивності звукової хвилі.

Розподіл сигналу на вокалізований та невокалізований. Вимір частоти основного тону

В задачах якісної обробки мовних сигналів важливу роль грає розділення ділянок вокалізованого і невокалізованого мовного сигналу. Фонеми, що мають вокалізовану ознаку — це фонеми з одним періодичним джерелом (голос), що не має різкого включення. Як правило, перші три форманти голосних для чоловічого голосу знаходяться в частотах нижче 3200 Гц. Форманти голосних характеризуються невеликим загасанням; на спектрограмі це виявляється в тому, що смуга частот кожною з формант відносно вузька. Внаслідок від'ємного схилу голосового спектру нижні форманти мають вищу інтенсивність. Проте, зважаючи на те що вухо більш чутливе до частот між 1000 і 2000 Гц, при сприйнятті це пониження інтенсивності спектру врівноважується [4].

Аналіз процесу мовотворення і реальних мовних сигналів показує, що вокалізовані ділянки мовного сигналу складаються з квазіперіодичних послідовностей асиметричних, швидко згасальних коливань [5] [6]. Тривалість квазіперіодів, визначувана параметрами голосового тракту, і в першу чергу, голосових зв'язок, є основним тоном (ОТ) мовного сигналу. Під виділенням частоти ОТ зазвичай розуміють визначення як миттєвої частоти коливань голосових складок диктора, так і форму цих коливань. Миттєва частота ОТ є значущим параметром практично у всіх завданнях класифікації мови. В задачах ідентифікації дикторів по значенню середньої частоти ОТ диктора можна створити, принаймні, два добре помітних класу дикторів (чоловіки і жінки). У завданнях розпізнавання мови за наявності ОТ, її можна класифікувати на вокалізовані (голосні і дзвінкі приголосні) і невокалізовані (шиплячі і глухі приголосні) звуки. У завданнях класифікації емоційних станів диктора використовується динаміка частоти ОТ і різних показників форми імпульсу ОТ (тривалість переднього фронту, показник асиметрії і т.п.). Таким чином, основний тон також є важливим для прогнозу тих варіацій в положенні формант, котрі зумовлені відмінностями в довжині голосового тракту у різних індивідів. Крім того, основний тон служить як точка відліку для розрізнювальної ознаки напруженості.

В більшості випадків для розрізнення голосних звуків досить двох перших формант, при цьому F1 (діапазон 150…850 Гц) співвідноситься з ознакою артикуляції підйому (розчину), тобто з розрізненням голосних верхнього і нижнього підйому (вузьких ― широких; закритих — відкритих). Як правило для вузьких голосних значення F1 нижче. Форманта F2 (діапазон 500…2500 Гц) співвід-

носиться з ознакою ряду. Для «передніх» голосних значення F2 вище, для «задніх» — нижче. Сумарне значення частот F1+F2 співвідноситься з ознакою лабіалізації під час артикуляції фонем, адже лабіалізація звуку викликає пониження частот, відповідних F1 і F2. [5]

Параметризація мовних сигналів

Головна роль процедури параметризації полягає у виокремленні найінформативніших параметрів мови задля вирішення конкретної задачі. Залежно від типу завдання класифікації може бути застосований той чи інший вид параметризації. Наприклад, основна вимога до виду параметризації при дикторонезалежному розпізнаванні мови в тому, щоб вона якомога сильніше «згладжувала» індивідуальні особливості голосів дикторів, і зворотне завдання повинне вирішуватися при параметризації в системі текстовонезалежній ідентифікації дикторів.

Головним при параметризації є припущення про стаціонарність мовного сигналу на проміжках часу близько декількох мілісекунд [2]. Таким чином, в ході аналізу мова розбивається на блоки даних, які зазвичай називають вікнами. На основі даних вікна обчислюється вектор ознак, який є основою для вирішення будь-якого завдання обробки мови.

Важливою також є оцінка значущості спектральних компонент мовленнєвого сигналу. Ця оцінка тісно пов'язана з суб'єктивним сприйняттям мовленнєвого сигналу людиною, що, в свою чергу, зумовлено властивостями людського слуху. Характер суб'єктивного сприйняття частоти тональних сигналів прийнято описувати масштабною «мел» шкалою [2], що описує емпірично-отриману залежність відчуття частоти тонального сигналу.

Власне методів параметризації існує багато:

1. Обчислення логарифмічно масштабованих кепстральних коефіцієнтів. Метод ґрунтується на обчисленні дискретного косинус-перетворення згладжених компонентів спектру Фур'є. Вихідний вектор ознак має невелику розмірність, отже метод власне стискає спектральні дані.

2. Обчислення масштабованих коефіцієнтів лінійного передбачення (мел-коефіцієнтів). Відповідно до [2] метод полягає у: обчисленні компонентів спектра Фур'є вхідного сигналу; згладжуванні спектру операцією згортки в трикутних вікнах; помноження отриманих коефіцієнтів на емпірично-отриману функцію рівної гучності; розрахунок оберненого перетворення Фур'є; розрахунок коефіцієнтів лінійного передбачення отриманого згладженого сигналу; розрахунок кепстральних коефіцієнтів, на основі коефіцієнтів лінійного передбачення.

3. Вибір значущих компонент спектральної оцінки. Цей метод дозволяє отримати уяву про те, які з компонент спектру є значимими і які функціональні залежності існують між компонентами спектра. Виявлення функціональних залежностей між спектрами сигналу, зокрема дозволяє добре відокремити один стан сигналу від іншого. Іншими словами, це дозволяє, наприклад, зробити вибір між парою фонем чи парою дикторів.

4. У [7] розглядається модель мовотворення, яка описує мовленнєвий сигнал положенням частотних моментів енергії сигналу в широких формантних діапазонах, що дає змогу знизити варіацію ознак за рахунок спектральних варіацій. Двійкове кодування положення цих моментов на основі відношення енергій в частотних піддіапазонах допомагає уникнути впливу амплітудних варіацій. Метод описує мовленнєвий сигнал двійковими восьмирозрядними векторами частотно-детектувальної і частотно-сегментувальної функцій, одержаними кодуванням положення частотних моментів у визначених діапазонах частот. Це дає змогу підвищити інваріантність мовних образів до диктора і голосності вимовляння, дещо знизити (порівняно з відомими методами) надлишковість відображення мовної інформації, здійснити процеси сегментації і маркування сигналу на звукотипи паралельно в часі і тим самим дещо збільшити швидкість розпізнавання.

Сегментація сигналу на основі спектральної ентропії

Метод спектральної ентропії, це метод визначення меж мови, заснований на обчисленні ентропії (як міри невизначеності або безладу в деякому розподілі) спектра сигналу. Для стійкого виділення мови використовується властивість відмінності значень ентропії для мовних сегментів і для фонових шумів. Відмінна риса даного підходу полягає в тому, що цей показник є малочутливим до змін амплітуди сигналу і, отже, дозволяє більш стійко і точно визначати межі мови.

Експериментальні результати по застосуванню методу спектральної ентропії показали, що мовні фрагменти успішно виділяються із звукових сигналів, що містять різні види сильних шумів (білий, рожевий, вузькосмуговий шум і т. д.) і звукових артефактів. Крім того, метод спектральної ентропії має прийнятну обчислювальну складність, що дозволяє його ефективно використовувати в мобільних системах розпізнавання мови реального часу.

Вперше ентропію спектру було запропоновано використовувати для даного завдання в 1998 році [8, 9]. Алгоритм спектральної ентропії здійснюється таким чином. Сигнал, що надходить з мікрофону, конвертується з аналогового в цифрове представлення і ділиться на короткі сегменти. При цьому перекриття сусідніх сегментів складає трохи більше 25 %. Далі, використовуючи алгоритм швидкого перетворення Фур'є (ШПФ), обчислюється короткочасний спектр сегмента сигналу. Потім проводиться нормалізація обчисленого спектра по всіх частотних компонентах. Кількість використовуваних спектральних компонент може обиратися від декількох десятків до декількох сотень. На даному етапі важливо знайти компроміс між необхідною чутливістю і обчислювальним навантаженням.

Для того, щоб вже на цьому етапі аналізу відкинути деякі види шумів, вводяться деякі обмеження, а саме: смуга частот обмежується значеннями 200 … 8000 Гц. Ця смуга частот охоплює практично всі частотні компоненти, присутні в мовному сигналі людини, отже таке обмеження дозволяє виключити дію як дуже низькочастотних, так і високочастотних шумів (наприклад, внутрішніх шумів звукової карти або мікрофона).

Можливі значення щільності вірогідності звичайно теж обмежуються як зверху так і знизу, що дозволяє виключити шуми, зосереджені у вузькій окремій області, а також шуми, що мають приблизний однаковий розподіл частотних компонент по всьому спектру (наприклад, білий шум). Додатково можуть використовуватися різні методи очищення сигналу від шуму (наприклад, адаптивний фільтр Калмана або методи спектрального віднімання) [10]

На наступному етапі проводиться обчислення спектральної ентропії, отриманого нормованого спектра. Далі застосовується медіанне згладжування послідовності набутих значень спектральної ентропії. На відміну від багатьох інших методів згладжування (наприклад, методу ковзного середнього), даний метод є значно стійкішим до окремих викидів і випадкових спотворень даних. В основу методу покладено обчислення ковзної медіани. Для того, щоб знайти значення ковзної медіани в точці t, обчислюється медіана значень ряду в часовому інтервалі $[t-q, t+q]$, яка визначається як центральний член послідовності значень ряду, що входять в цей часовий інтервал, впорядкованою за збільшенням.

Для задач, в яких вид шуму і його спектр мало змінюються з часом, є дуже ефективним додаткове обчислення ентропії спектра для короткої ділянки звукового сигналу, що містить тільки акустичний фоновий шум без включення мовних або інших звукових фрагментів, і віднімання ентропії для шуму з набутих значень ентропії для аналізованого сигналу.

На останньому етапі застосовується логіко-часова обробка, що враховує допустиму на практиці тривалість мовних і немовних фрагментів. Спочатку обчислюється адаптивний поріг значення ентропії, який служить для виділення крайових точок (початок і кінець) гіпотези фрагмента мови. На основі цього порогу обираються акустичні сегменти аналізованого сигналу, які належать до мови людини. Після цього проводиться логіко-часова обробка виділених ділянок сигналу. Ця обробка необхідна, оскільки у багатьох випадках через виникнення різних звукових артефактів немовні ділянки сигналу помилково приймаються за мову, і навпаки, деякі ділянки, що містять мову, відкидаються через специфічні акустичні характеристики. При логіко-часовій обробці застосовуються два основні показники: мінімальна тривалість виділених фрагментів, що містять мову; та максимальна тривалість безмовної ділянки.

Враховуючи, що людина не може мати дуже короткі мовні фрагменти, а також те, що в мові завжди присутні певні паузи (наприклад, змички перед вибуховими приголосними), експериментальним шляхом визначається порогове значення для мінімальної тривалості мовної ділянки і максимальної тривалості безмовної ділянки.

Кореляційний аналіз

Цей метод розпізнавання ґрунтується на порівнянні деяких характеристик мови (енергетичних, спектральних тощо). В більшості випадків за еталони приймаються цілі слова. Цей метод зручний для використання в системах розпізнавання дискретної мови, тобто в системах з обмеженим слов-

ником (командні системи). Словник в таких системах може бути дуже великий, але платою за це є дикторозалежність системи, тобто система має бути попередньо натренована на кожного користувача.

На рис. 1 зображена структурна схема обробки мовного сигналу в системі розпізнавання мови на основі кореляційного аналізу [11], що була вперше запропонована Фумітадом Ітакурою в 1975 році [12]. Для збільшення ефективності і зменшення об'єму обчислень Ітакура, розуміючи що полоса частот вхідного сигналу складає 3000 Гц, використовував частоту дискретизації 6670 Гц. Згідно з запропонованим методом, після визначення моментів початку і закінчення слів на основі методів обробки в часовій області оцінюються перші вісім коефіцієнтів кореляції зі швидкістю 67 оцінок/с. Для компенсації спотворень спектра, що з'являлися внаслідок недосконалостей приладів передачі мови, Ітакура обчислював усереднений спектр на великому інтервалі часу, за допомогою усереднення коефіцієнтів кореляції по всій фазі і припасуванням отриманого спектра до усередненого по фазі спектра двополюсної моделі. Параметри двополюсної моделі використовувалися для побудови зворотного фільтра. Середній по фазі спектр потім нормувався по входу шляхом згортки початкових коефіцієнтів і коефіцієнтів кореляції імпульсної характеристики зворотного фільтра. Перші шість нормованих автокореляційних коефіцієнтів потім використовувалися для створення і розпізнавання еталонних зразків.

Рис. 1. Структурна схема СРМ на основі кореляційного аналізу

Після нормалізації спектра починається процедура розпізнавання. Невідома фраза порівнюється з кожним наявним у файлі еталоном. Порівняння відбувається на основі міри розрізнення в просторі параметрів лінійного передбачення мови. Ця міра використовується і для динамічного узгодження часового масштабу вхідної фрази при мінімізації відстані з кожної з еталонних фраз.

На основі обчислення відстаней до кожного слова зі словника вибирається те слово, для якого отримана мінімальна відстань. Якщо абсолютне значення відстані перевищує деякий поріг, то рішення не приймається. У цьому випадку вибирається інше слово з мінімальною відстанню, воно приймається як рішення і надходить на вихід системи розпізнавання.

Ця система досліджувалася [12] з використанням двох різних словників. Перший, обсягом приблизно 120 слів (назви міст Японії) давав 97,3 % правильного розпізнавання слів, а другий словник, що складався із 26 букв і десяти цифр давав точність 88,6 %. Загалом система не відрізнялася високою швидкодією і вимагала тривалого і ретельного навчання.

Фонемно-орієнтований метод

Цей метод ґрунтується на виокремленні акустичних ознак слів (фонеми, трифони) з вхідного мовленнєвого сигналу. Текст, як відомо, складається з букв, слів, речень, — тобто він дискретний. Мова ж в нормальних умовах звучить неперервно. Людська мова, на відміну від тексту, зовсім не складається з букв. Наприклад відомо, що слово «папа» складається з чотирьох букв, однак на фонограмі (рис. 2) чітко видно, що насправді, воно складається не з чотирьох, а тільки з двох звуків (фонем): «па» і «па».

Рис. 2. Фонограма слова «папа»

Добре відомим є той факт, що елементарні звуки, з яких складається мова, не еквівалентні буквам. Тому ввели поняття фонеми для позначення елементарних звуків мови. Дотепер фахівці ніяк не можуть вирішити — скільки ж усього різних фонем існує. У лінгвістичній науці є цілий розділ — фонетика. Для кожної мови існує свій кінцевий набір фонем. В українській мові за одними даними їх близько 40, за іншими — більше сотні.

Розглянемо модель побудови системи розпізнавання мови заснованої на фонемно-орієнтованому підході (рис. 3).

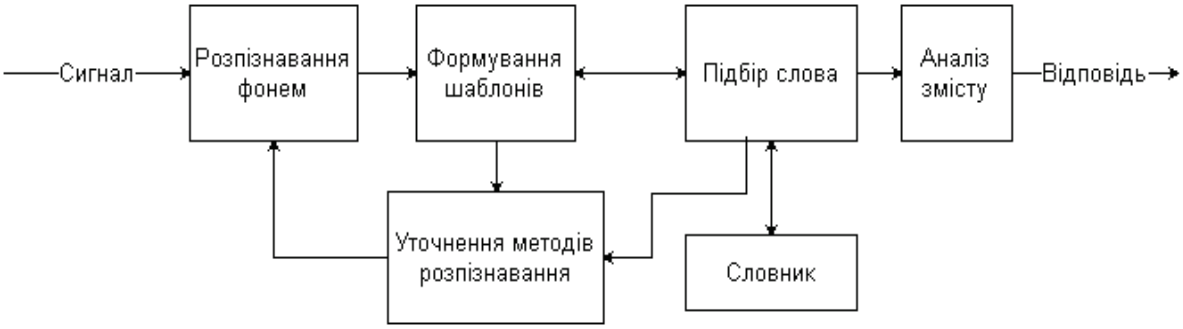

Рис. 3. Модель фонемно-орієнтованої СРМ

Зі списку розпізнаних з визначеною точністю фонем, складається шаблон, що передається на наступний рівень, де по ньому відбувається підбір найбільш вірогідного слова і передача інформації про вибір на більш високий рівень для подальшого аналізу і на нижній, за рахунок зворотного зв’язку, для налаштування системи на конкретного користувача. Перевага такої моделі побудови полягає в високій адаптивності, що дає можливість динамічного самоналаштовування системи на диктора, і багаторівнева система перевірок, що підвищує точність роботи.

Експертні системи

Експертні системи, це системи з різними засобами формування і обробки бази знань. Прикладом побудови таких систем може слугувати система ROBOTRON (рис. 4).

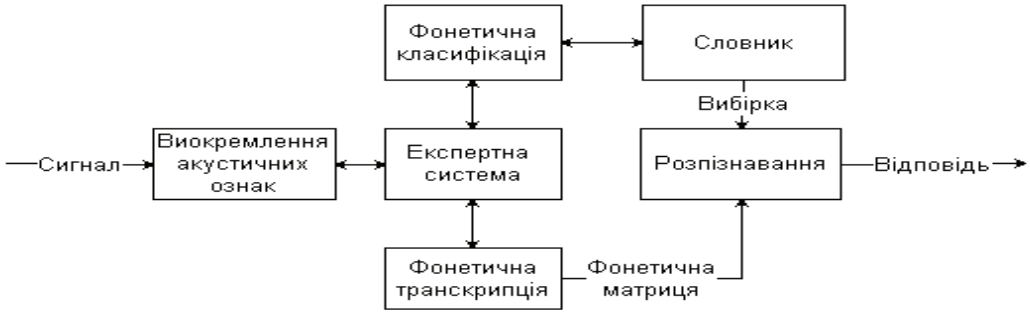

Рис. 4. Схема блоку розпізнавання системи ROBOTRON

Система розроблена в інституті інформатики Дрездена [13] та призначена для розпізнавання неперервної мови. Вона дозволяє обробляти мовну послідовність слів з необмеженою тривалістю без додаткової інформації про межі слів. Розроблена для дослідження та оптимізації процесів розпізнавання мови і навчання.

Система складається з аналізатора, математичного забезпечення, блоку стиснення інформації, блоку розпізнавання та алгоритму навчання. Побудова і обробка сугубо баз знань не є максимально ефективними щодо вирішення проблем розпізнавання, тому доводиться знаходити інші математичні підходи.

Приховані Марківські моделі

Приховані Марківські моделі в даний час є одним з найефективніших підходів до побудови систем автоматичного розпізнавання мови. Той факт, що мова призначена для передачі, а отже, і для захисту інформації, дозволяє розглядати її як деякий код, а мовний потік — як послідовність деяких кодових пакетів. Неважливо, що є елементом цього коду — фонеми, склади або цілі слова, значення має лише те, що імовірність появи будь-якого елемента коду залежить від деякого числа попередніх елементів. Таким чином, мова породжується Марківським джерелом, а мовний код є випадковим. Для реалізації цього підходу застосовується так звана «прихована Марківська модель» (Hidden Markov Model), в якій дозволяються переходи тільки у наступний чи поточний стан.

Прихованою Марківською моделлю (ПММ) зветься Марківський процес, що не спостерігається безпосередньо. Результати дії цього процесу спотворюються випадковим процесом і лише після цього стають доступними для спостережень. Параметрами прихованої Марківської моделі є:

— можливі стани процесу;

— ймовірність переходу з одного стану в інший;

— ймовірність спотворення стану, що спостерігається.

Загальна теорія прихованих Марківських моделей може бути легко адаптована для розв'язання окремих задач розпізнавання мови. Конкретні вимоги і умови кожної задачі знімають або, навпаки, додатково накладають деякі обмеження на ПММ, що дозволяє у визначених межах варіювати обчислювальну складність системи і вибирати ефективніший алгоритм навчання моделі. Адекватність ПММ мовному сигналові досягається за рахунок вдалого вибору стаціонарних ділянок сигналу і визначення відповідних розподілів імовірності векторів ознак [14].

Розглянемо систему розпізнавання мови, побудовану на застосуванні ПММ із використанням кодової книги (рис. 5).

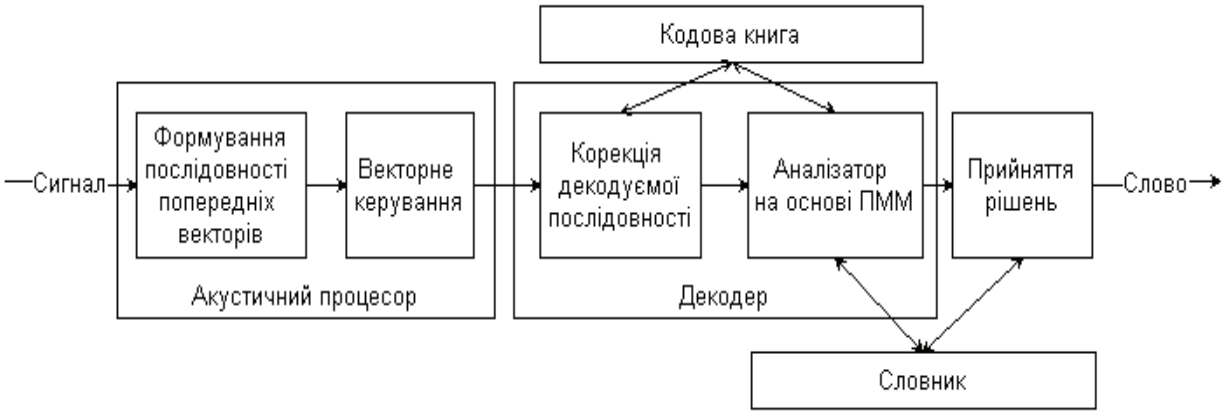

Рис. 5. Система розпізнавання мови на основі ПММ

Акустичний процесор виконує задачі дискретизації, попередньої обробки й виділення характерних ознак, тобто перетворює акустичний мовний сигнал в набір характери стичних векторів, що в

наступному використовуються для побудови кодової книги, навчання ПММ та безпосередньо розпізнавання.

Декодер здійснює розрахунок найбільш імовірних слів, що відповідають послідовності кластерів, які отримані на виході акустичного процесора. При цьому кожному слову відповідає власне Марківський ланцюг з N станів і вимог переходу між ними. Робота декодера полягає в безпосередньому звернені до кодової книги (набору обмеженої кількості еталонних ознак, що є словами кодової книги), згідно з алгоритмом розпізнавання. В якості алгоритму розпізнавання обирається алгоритм послідовного декодування зі зворотним зв'язком, що реалізує відношення максимальної правдоподібності. Ймовірності символів, що спостерігаються, визначаються функцією розподілу, в той час як імовірності переходу із одного стану в інший задаються дискретними значеннями з матриці розподілу. Процес розпізнавання полягає на застосуванні алгоритму Вітербі (цей алгоритм є варіантом методу динамічного програмування для ланцюгів Маркова, він складається з прямого та зворотного проходів).

Перевага алгоритму полягає в тому, що він дозволяє досягти компроміс між об'ємом пам'яті, що вимагається, та складністю обчислень, за рахунок визначення на кожному кроці функції правдоподібності та порівняння її з граничним значенням. Тестування таких систем із словником з 50 слів, що промовляються різними дикторами, які не приймали участі в процесі навчання системи, доводить вірогідність розпізнання не нижче 78 % [14].

Існує багато інших прикладів адаптування ПММ до розпізнавання мовленнєвої інформації. Так, в [2] наведено приклад розпізнавання мовленнєвої інформації за допомогою ПММ, заснованої на трифонах та алофонах. Фон, або акустичне представлення, що супроводжує фонему не є однорідним у часі, а має свою внутрішню структуру і для описання цієї неоднорідності фонему моделюють як послідовність станів. Кожна фонема розглядається, як ПММ, що визначена на своїй множині станів, зв'язаних переходами. Для розв'язання задачі розпізнавання проводиться зв'язування схожих станів трифонів, за допомогою фонетичного дерева розв'язків. Зв'язування дає набір кінцевих кластерів, в яких містяться достатня кількість даних для вдалої оцінки параметрів вихідної функції розподілу.

В [14] розглядається поєднання ПММ з нейронними мережами. Нейронна мережа використовується для акустичного моделювання вхідного сигналу, що дозволяє позбавити ПММ деяких вхідних обмежень. Наприклад, стає відсутнім припущення про евклідовість метрики простору ознак-векторів, що притаманно класичному ПММ-підходу з дискретними векторами ознак із векторної кодової книги. Нейромережева акустична модель компактніша і потребує меншої кількості параметрів для отримання порівняного за якістю моделювання.

Змішані Гаусівські моделі та відношення правдоподібності

На початку 2000-х років, змішані Гаусівські моделі (ЗГМ), стали домінувальним підходом для моделювання систем текстово-незалежного розпізнавання диктора [15]. В задачах розпізнавання диктора ЗГМ використовується, в якості багатовимірної загально-імовірнісної моделі щільності розподілу, здатної представляти довільні щільності, що робить її придатною до застосування в текстово-незалежних системах розпізнавання. Використання ЗГМ у задачі розпізнавання диктора було вперше описано в [16]. Модернізація ЗГМ-систем для верифікації диктора було описано і оцінено за кількома загальнодоступними наборами дикторських голосів в [17].

Модель застосовується у задачі розпізнавання диктора таким чином. *Нехай надано сегмент вхідного сигналу Y, і є гіпотетичний диктор S; завдання розпізнавання (верифікації) диктора, полягає у визначені, чи був сигнал Y, сказаний диктором S.* Таку задачу часто називають задачею виявлення єдиного диктора, бо в випадках, якщо не сказано інше, використовується неявне припущення, що *Y* містить сигнал тільки від одного диктора. Звичайно, модель може застосовуватися і для розв'язання задач виявлення багатьох дикторів.

Задача виявлення єдиного диктора може бути записана, в сенсі гіпотез таким чином:

H_0: *Y* був вимовлений гіпотетичним диктором *S*, і

H_1: *Y* не був вимовлений гіпотетичним диктором *S*

Гіпотеза приймається, якщо її відносна ймовірність більше за наданий поріг, тобто:

$$\frac{p(Y/H_0)}{p(Y/H_1)}\begin{cases}\geq\theta & \text{приймаємо } H_0;\\ <\theta & \text{відкидаємо } H_0,\end{cases}$$

де $p(Y/H_i)$, i = 0,1 — функція щільності ймовірності для гіпотези H_i, оціненої для сегмента спостережуваного сигналу *Y*, яку також називають ймовірністю гіпотези H_i для даного сегменту. θ — це поріг для прийняття або відкидання H_0. Основна мета системи виявлення диктора полягає у визначенні методики підрахунку значень двох ймовірностей — $p(Y/H_0)$ і $p(Y/H_1)$.

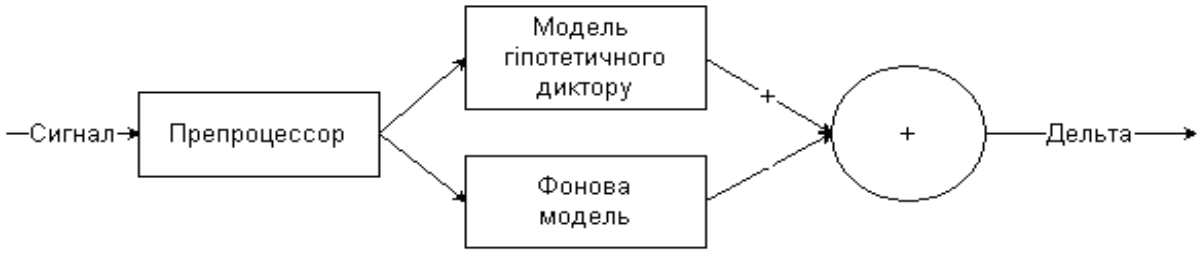

Рис. 6. Загальна схема алгоритму розпізнавання, побудованого на відношенні правдоподібності

На рис. 6 показано основні компоненти, що мають місце в системах розпізнавання диктора, побудовані на відношеннях правдоподібності. Завданням препроцесора є виділення особливостей мовленнєвого сигналу, що виявляють дикторозалежну інформацію. Крім того, в препроцесорі можуть бути реалізовані методи, що мінімізують ефекти змішування в виділених сигналах, як, наприклад, лінійне фільтрування або шум. Результатом роботи препроцесора, зазвичай є послідовність векторів особливостей, що представляють випробовуваний сегмент в дискретному часі. Алгоритм не накладає ніяких обмежень на послідовність видобутих ознак. Нема обмежень на те, чи в послідовних проміжках часу були видобуті ознаки, чи ні. Наприклад, рівень всього мовленнєвого сигналу може бути використаний, як ознака. Видобуті вектори особливостей використовуються, для розрахунку ймовірностей H_0 та H_1.

Тоді як модель для H_0 добре визначена і може бути оцінена, використовуючи навчальний мовленнєвий сигнал *S*, модель для H_1 визначена менше, тому що потенційно вона має представити повний простір можливих альтернатив для гіпотетичного диктора. Для моделювання альтернативних гіпотез застосовувалися два головні підходи.

Перший підхід полягає в використовуванні набору моделей інших дикторів для покриття простору альтернативних гіпотез. В різних контекстах, цей набір інших дикторів називався наборами відношень правдоподібності, компаньйонами, і другорядними дикторами.

Нехай надано N другорядних дикторів $\{\lambda_1,\ldots,\lambda_N\}$, моделлю альтернативної гіпотези називають

$$p(X/H_1) = F\left(p(X/\lambda_1),\ldots,p(X/\lambda_N)\right),$$

де $F()$ є деякою функцією, наприклад, середнє або максимальне значення з набору ймовірностей другорядних дикторів. Вибір, розмір, і комбінації другорядних дикторів були темою багатьох досліджень [17]. Взагалі, було виявлено, щоб отримати найкращий результат з цим підходом необхідно для кожного окремого диктора використовувати різний (індивідуальний) набір другорядних дикторів. Це може бути недоліком в програмах, які використовують великий набір гіпотетичних дикторів, тому що кожному потрібно підібрати його власний набір другорядних дикторів.

Другий підхід до моделювання альтернативних гіпотез полягає у об'єднанні деяких дикторів в одну модель і навчанні цієї єдиної моделі. Цю єдину модель називають по-різному: загальна модель, світова модель, і універсальна другорядна модель (УДМ). Нехай надано сукупність зразків мовленнєвих сигналів великої кількості дикторів, що входять до набору можливих дикторів, що

будуть розпізнані. Єдина модель λ_{bkg} — була навчена на представлення альтернативних гіпотез. Дослідження цього підходу зосередилися на виборі і компонуванні дикторів і мовленнєвих сигналів, що використовуються для навчання цієї єдиної моделі [18], [19]. Головною перевагою цього підходу є те, що єдина дикторонезалежна модель для конкретної задачі може бути навчена лише один раз, а потім бути використана для всіх гіпотетичних дикторів в рамках цієї задачі. Також можливо використовувати більше однієї другорядної моделі, спеціально пристосованих до конкретних наборів дикторів [19].

Важливим кроком у створенні згаданого вище детектора відношення правдоподібності є вибір фактичної функції правдоподібності $p(X/\lambda)$. Вибір цієї функції в значній мірі залежить від вибраних ознак і специфіки проблеми. Для текстово-незалежного розпізнавання диктора, де немає апріорно заданого тексту, що має сказати диктор, найкращими функціями правдоподібності виявилися змішані Гаусівські моделі. В текстово-залежних застосуваннях, коли текст, що має бути сказаний дикторами наперед задано, ЗГМ може надати додаткову інформацію для побудови функції ймовірності. Можна використовувати більше ускладнені функції правдоподібності, як наприклад функції, що засновані на прихованих Марківських моделях, але вони не надають ніякої переваги над ЗГМ для задач текстово-незалежного розпізнавання дикторів.

Переваги використання ЗГМ як функції правдоподібності полягають у тому, що вони мають невелику обчислювальну складність, побудовані на добре зрозумілій статистичній моделі, і, для текстово-незалежних задач, вони є нечутливими до тимчасових мовних аспектів, моделюючи тільки основний розподіл акустичних спостережень від диктора. Останнє одночасно є і недоліком, оскільки тут не використовуються вищі інформаційні рівні мовленнєвого сигналу диктора, що з'являються в тимчасовому мовленнєвому сигналі. До цього часу, проте, ці підходи (наприклад, великий словник або розпізнавання фонем) в основному використовувалися тільки для розрахунку значень правдоподібності, без явного використання будь-якого вищого рівня інформації, як, наприклад, дикторозалежне використання слів або розмовний стиль.

Переваги системи розпізнавання, що ґрунтується на змішаній Гаусівській моделі з універсальною другорядною моделлю (ЗГМ-УДМ) полягають в виборі оптимального відношення правдоподібності для розпізнавання мовленнєвого сигналу, використанні простих але ефективних ЗГМ для функції правдоподібності, використанні УДМ для представлення конкуруючих альтернативних дикторів, і Байесову адаптацію для отримання моделі гіпотетичного диктора.

Недоліки системи полягають у поганій обробці умов невідповідності, адже системи ЗГМ-УДМ залежать від низькорівневої акустичної інформації. Нажаль, інформація про диктора і канали сигналу, що зв'язані між собою в один потік невідомим чином впливають на мовленнєві спектральні ознаки і якість розпізнавання цього сигналу наявними системами значно погіршується при зміні мікрофону або навколишнього акустичного середовища. Були зроблені кроки в напрямку зменшення цих недоліків. Були намагання використовувати адресацію шумів лінійного каналу з відніманням кепстрального середнього значення і застосуванням фільтру RASTA, адресацію нелінійних ефектів з нормалізацією оцінки log-правдоподібності (HNORM) і за рахунок хвильової компенсації [20], але недоліки все ще присутні і мають бути вирішені для вдалого використання ЗГМ, як систем розпізнавання мовленнєвої інформації.

Висновки

Сьогодні найскладнішими елементами побудови систем розпізнавання мови є алгоритми відтворення послідовності вимовлених слів, визначення ефективної інваріантної акустичної моделі й формування мовних моделей для тих, чи інших мов, що потребує багаторічної праці спеціалістів різних галузей науки: інженерів-акустиків, дослідників мовних технологій, нейрофізіологів, лінгвістів.

На акустичному рівні дуже важливим є якісне й водночас досить компактне представлення звукового сигналу в багатомірному просторі ознак, що містять значиму для розпізнавання інформацію. Для побудови векторів ознак використовуються методи спектрального аналізу (лінійне передбачення мови, гомоморфний аналіз), однак вони мають ряд недоліків. В частині II буде описа-

но метод створення акустичної моделі сигналу за допомогою вейвлетного базису і перетворення Фур'є.

Ефективність використання статистичного підходу — прихованих Марківських моделей для встановлення послідовностей слів безперечна, однак існує низка обмежень цього підходу, які можна частково компенсувати за рахунок використання нейромереж чи штучних імунних систем, що будуть описані у наступній частині.

Таким чином, ефективна система розпізнавання мови має містити в собі такі етапи обробки вхідного сигналу, як видалення шуму, сегментація, виділення вокалізованих ділянок, параметризація, розпізнавання, коригування за словником з оберненим зв'язком. Зрозуміло, що не один метод не може покрити усі етапи. Ефективна система має поєднувати в собі найкращі методи виконання кожного етапу, використовуючи їх переваги.

Також треба зазначити особливість сприйняття мови людиною — вона не все те чує, що сприймає, більшість інформації вона домислює. Отже, створюючи системи розпізнавання мови, треба виходити за рамки самої мови. В майбутньому можна очікувати появу не тільки систем мовного діалогу, а й створення інтерпретаторів, спроможних правильно передати зміст й семантичне наповнення спотвореного мовного повідомлення чи інформації, отриманої в умовах наявності стаціонарних і нестаціонарних перешкод.

СПИСОК ЛІТЕРАТУРИ

1. Плотников В. Н. Речевой диалог в системах управления / В. Н. Плотников, В. А. Суханов, Ю. Н. Жигулевцев. — М. : Машиностроение, 1988. — 223 с. — ISBN 5-217-00148-8.

2. Аграновский А. В. Теоретические аспекты алгоритмов обработки и классификации сигналов / А. В. Аграновский, Д. А. Леднов. — М. : Радио и связь, 2004. — 164 с.

3. Большаков И. А. Статистические проблемы выделения потока сигналов из шума / И. А. Большаков. — М. : Советское радио, 1969. — 464 с.

4. Jakobson R. Preliminaries to speech analysis. The distinctive features and their correlates / R. Jakobson, C. Gunnar, M. Fant, M. Halle. — Cambridge : Massachusetts Institute of Technology, 1961. — ISBN 978-0-262-60001-9.

5. Фант Г. Акустическая теория речеобразования / Гуннар Фант. — М. : Наука, 1964. — 284 с.

6. Рабинер Л. Теория и применение цифровой обработки сигналов / Л. Рабинер, Б. Гоулд. — М. : Мир, 1978. — 848 с.

7. Биков М. Дикторонезалежне описання образів в системах розпізнавання сигналів мови / Микола Биков, Абдурахман Раїмі, Максим Биков // Вимірювальна техніка та метрологія. Збірник наукових праць. — 2006. — № 66. — С. 13—17.

8. Waheed K. A robust algorithm for detecting speech segments using an entropy contrast : праці міжн. конф. 45th IEEE International Midwest Symposium on Circuits and Systems MWSCAS'2002, 4-7 серп. 2002, Oklahoma (USA). — С. 328—331, III.

9. Shen J.-L., Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments : праці міжн. конф., 30 лист. — 4 груд. 1998, 5th International Conference on Spoken Language Processing, Sydney (Australia).

10. M. Fujimoto. Evaluation of noisy speech recognition based on noise reduction and acoustic model adaptation on the AURORA2 tasks : праці міжн.конф., вер. 2002, Spoken Lang. Processing' ICSLP'2002, Denver (USA), 2000 — С. 465—468, I.

11. Рабинер Р. Л. Цифровая обработка речевых сигналов / Р. Л. Рабинер. — М. : Радио и связь, 1981. — 495 с.

12. Itakura F. Minimum Prediction Residual Principle Applied to Speech Recognition : праці наук. конф., Лютий 1975, IEEE Trans. Acoustics, Speech, and Signal Proc, 1975. — Т. 23, № 1. — С. 67—72.

13. Потапова Р. К. Речевое управление роботом / Р. К. Потапова. — М : Радио и связь, 1989. — 328 с.

14. Бовбель Е. И. Статистические методы распознавания речи: скрытые Марковские модели / Е. И. Бовбель, И. Э. Хейдеров // Зарубежная радиоэлектроника. Успехи современной радиоэлектроники. — 1998. — № 3. — С. 45—65.

15. Reynolds D. Speaker verification using adapted gaussian mixture models / Douglas A. Reynolds, Thomas F. Quatieri, Robert B. Dunn // Digital Signal Processing. — 2000. — № 10. — С. 19—41.

16. Rose R. Text-independent speaker identification using automatic acoustic segmentation : праці міжн. конф., 3 — 6 квіт. 1990, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1990. — С. 293—296, I.

17. Reynolds D. Speaker identification and verification using Gaussian mixture speaker models / D. A. Reynolds // Speech Communication. — 1995. — № 17. — С. 91—108.

18. Matsui T. Likelihood normalization for speaker verification using a phoneme and speaker-independent model / T. Matsui, S.Furui // Speech Commun. — 1995. — № 17. — С. 109—116.

19. Rosenberg A. E. Speaker background models for connected digit password speaker verification : праці міжн. конф. International Conference on Acoustics, Speech, and Signal Processing, 1996. С. 81—84.

20. Quatieri T. Magnitude-only estimation of handset nonlinearity with application to speaker recognition : праці міжн. конф. International Conference on Acoustics, Speech, and Signal Processing, May 1998.

Рекомендована кафедрою інтелектуальних систем

Надійшла до редакції 15.06.09
Рекомендована до друку 25.06.09

Добрушкін Григорій Олександрович — аспірант, ***Данилов Валерій Якович*** — професор.

Кафедра математичних методів системного аналізу Національного університету України «Київський політехнічний інститут»