СИНТЕЗ ШВИДКОДІЮЧИХ БАГАТОЗНАЧНИХ СТРУКТУР МОВНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

М.Ф Бондаренко, І.А Ревенчук, Г.Г Четвериков

ХТУРЕ
310166, Україна, м.Харків, пр.Леніна 14
тел. (380)-0572)-40-93-97
факс. (380)-0572)-40-91-13
E-mail: imd@kture.kharkov.ua
http://kture.cit-ua.net

Анотація.

Розвиток комп'ютеризації, проникнення обчислювальної техніки у всі сфери науки, промисловості, суспільного життя потребує від людини якісного описання та формалізації, проблеми, створення алгоритмів, розроблення програми, аналіз результатів, рекурсивно видозмінювати постановку проблеми та всі наступні компоненти. Однією з основних функцій інтелектуальної системи є функція підсистеми спілкування на природній мові. Традиційна схема аналізу та синтезу тексту на природній мові включає до свого складу наступні види оброблення: виділення слів та фраз у передредакторі, морфологічний аналіз, синтаксичний аналіз, семантичний аналіз, переведення у внутрішнє зображення, розуміння тексту. Процедури оброблення інформації природньої мови утворюють комплекс, що дозволяє як аналізувати, так і синтезувати текст й називають його лінгвістичним процесором.

ВСТУП.

Паралельно з аналізом розвивається й створення електронних моделей голосового тракту, здатного штучно генерувати голос людини, тобто створюються ситеми синтезу штучної мови. Між іншим, людська мова – дискретна та багатозначна і повинна описуватись засобами дискретної математики та утворюватись за допомогою логічних численнь висловлювань та численнь предикатів, які дають можливість описати мову за допомогою апарату рівнянь.

1. НЕОБХІДНІСТЬ АСП ДЛЯ ОПИСУ БАГАТОЗНАЧНОЇ ОБЧИСЛЮВАЛЬНОЇ СТРУКТУРИ.

Наявність алгебри скінченних предикатів (АСП) відкриває можливість переходу від алгоритмічного опису інформаційних процесів до опису їх у вигляді рівнянь. Усі змінні в рівнянні рівноправні, - будь-які з них можуть виступати як в ролі незалежних, так і в ролі залежних. Алгоритми та програми описують незворотній процес роботи систем від входів до виходів, а рівняння АСП - забезпечують у процесі роботи систем й зворотній процес: вхідні сигнали можна подати на будь-які полюси системи і, відповідно, зняти результуючий сигнал теж можна з будь-яких полюсів. При цьому рівняння дають ту перевагу перед алгоритмами, що можна розрахувати реакцію системи навіть при неповному визначенні вхідних сигналів, у той час як неповністю розроблений алгоритм є непрацездатним. По-друге, за умов зміни знань про об'єкт система рівнянь АСП, покладених на структуру системи, завжди готова до використання, а алгоритм часто вимагає докорінної зміни її структури.[3]

Єдиним недоліком опису системи за допомогою рівнянь АСП є те, що коли число змінних у рівняннях велике, а в рівняннях інтелектних процесів це число без сумніву буде великим, то й комбінаторно зростає число способів розподілу вхідних та вихідних сигналів між полюсами системи. За цих умов практично неможливо створити повний набір алгоритмів, кожний з яких обчислював би реакції системи при певному способі розподілі вхідних та вихідних сигналів між полюсами системи.[5]

Отже необхідно щоб програми та дані не протистояли одне одному, а утворювали сумісну нерозчленовану структуру опрацювання даних і якраз в інтегруванні, єдності програм та даних скритії потенційні можливості росту систем штучного інтелекту, як наслідок, продуктивності цифрових систем та мереж. При цьому програма як така не повинна уводитися в систему, а сама система служить породжуючою програму структурою. Просто в кожний вузол мережі вбудовані універсальні функціональні перетворювачі з k-значним кодуванням, що можуть гнучко переналагоджуватись на виконання будь-яких необхідних функціональних перетворень.

Кожний вузол семантичної мережі повинен здійснювати аналіз завдання (інтерпретувати його) та зв'язувати один універсальний перетворювач з іншим, якщо на вході дані, що потребують опрацювання - вузол ці дані опрацьовує, якщо нітранслює дальше мережею. Ці хвилі, породжені керуючими командами розповсюджуються активним середовищем з універсальних перетворювачів, взаємодіють між собою (інтерферують) і тим дають можливість створювати нові алгоритми. В мережі залишаються сліди цих інформаційних припливів й відпливів і наступні хвилі течуть своїм, несподіваним шляхом. Дані, що оточують універсальний перетворювач, надходять до нього, перетворюються у відповідності з закладеними перетвореннями й розповсюджуються дальше комутаційною мережею, змінюючи заодно й саму мережу і породжуючи децентралізоване керування в результаті колективної взаємодії елементів.

2. КОНЦЕПТУАЛЬНА СТРУКТУРНО – ФУНКЦІОНАЛЬНА МОДЕЛЬ БАГАТОЗНАЧНОЇ КОМІРКИ.

Теоретичні та експериментальні дослідження й виникаючі під час створення систем ШІ ускладнення сприяють висуненню концепції адекватності багатозначної логіки та структур завданню створення систем ШІ з очікуваними властивостями та можливостями щодо підвищеного захисту. В цьому контексті, для розкриття шляхів побудови і паралельно-об'ємних k-значних структур, розглянемо концептуальну структурно-функціональну модель багатозначної комірки.

Довільна система ШІ на системному рівні характеризується набором функцій, що реалізуються нею та функціональними вузлами, які реалізують ці функці, а також інформаційним обміном під час функціональних перетворень. У відповідності з задачами, що вирішуються структурно-функціональна комірка узагальненого виду на рівні системного підходу декомпозується на три ієрархічних рівні: функціональний (аналітико-синтетичний); тактичний (аналізаторно - координаційний) ;стратегічний (координаційний).

Відповідно на функціональному рівні до складу k-значної об`ємно - просторової комірки входять: n-вимірний комутатор сигналів; комплекс порогових пристроїв, дешифратори просторових проміжних ознак та формувачі k-значних функцій. Комутатор сигналів є керуючим пристроїм входу системи ШІ, що визначає з яким вхідним сигналом працює комірка: зовнішнім чи від стратегічного рівня.

Комплекс порогових пристроїв дозволяє здійснювати перетворення неперервних чи дискретних за часом та за рівнем k-значних сигналів (семантичне опрацювання вхідного сигналу системи), а також формування простору проміжних ознак (простору k-значних за суттю характеристичних функцій), як результату семантичного опрацювання.

Проміжні ознаки дешифруються, у подальшому, у керуючі сигнали вихідного комплексу формувача k-значних функцій, що здійснює аналітичні функціональні перетворення. Результат перетворень на функціональному рівні надходить на вихід комірки, а також надходить для оцінювання, з точки зору семантичного змісту, на стратегічний рівень.

ВИСНОВКИ.

Виходячи з того, що викладено вище стає очевидним.

1. Побудова багатозначної обчислювальної структури чи системи передбачає створення базового набору типових уніфікованих компонент просторового типу, які складають наступний конгломерат засобів пристрої зовнішнього обміну, що перетворюють двозначні коди в багатозначні; універсальні багатозначні функціональні перетворювачі, комутаційні елементи на декілька напрямків. В зв'язку з орієнтацією на мікроелектронне виконання компонентів мереж усі ці засоби можуть бути класифіковані згідно наступних ознак: вид сигналу, що несе інформацію (інформаційну ознаку), галузь застосування, вид схемотехніки та технологія виготовлення.

2. Актуальність задачі створення основ теорії синтезу надшвидкодіючих k-значних структур для мовних систем штучного інтелекту з використанням математичних моделей української мови, що базуються на методах теорії інтелекту і базовій основі - алгебрі скінченних предикатів.

ЛІТЕРАТУРА.

1. Шабанов-Кушнаренко Ю.П. *Теория интеллекта.- Ч.1: Математические средства* - Харьков: Вища школа, 1982. - 240 с.

2. Шабанов-Кушнаренко Ю.П. *О проблемах теории интеллекта//* Пробл. бионики, 1990.- Вып. 44. - С. 3 - 10.

3. Бондаренко М.Ф. *О решении уравнений алгебры конечных предикатов//* Локальные автоматизированные системы автоматики.- Киев: Наукова думка, 1983.- С. 138-144.

4. Коноплянко З.Д. *Принципы построения многозначных систем искусственного интеллекта//* Проблемы бионики. -1990.- Вып. 45.- С.27-35.

5. Коноплянко З.Д., Четвериков Г.Г. *Анализ и синтез k-значных структур.* - Рук. деп. в ДНТБ України 5.12.94 р., N 2294 - Ук94. - 258 с.

6. Четвериков Г.Г. *О математическом описании арифметических отношений десятичных кодов//* АСУ и приборы автоматики.-Харьков: Вища шк., 1980. - Вып. 58. - С.22-26.

БАГАТОЗНАЧНА СМИСЛОВА ІНТЕРПРЕТАЦІЯ УСНОМОВНОГО СИГНАЛУ

Тарас ВІНЦЮК

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН і Міносвіти
40, просп. Академіка Глушкова, Київ 252022
Тел.: +380 44 266-4356 Факс: +380 44 266-1570
Електронна пошта: vintsiuk@uasoiro.freenet.kiev.ua

Abstract

Significant continuous speech understanding algorithm is proposed. It consists in that that for current discrete time step it is being found not only the $N>>1$ best initial word subsequences, which are permissible in the speech dialogue language, but at the next discrete time steps it is being examined only such words that could continue these $N>>1$ already accumulated permissible in the speech dialogue language initial word subsequences. By this way it is being removed lacks and united advantages of two earlier mentioned speech understanding models [1—4].

1. Вступ

В [1] запропоновано дві технології (моделі) машини для усного перекладу та/або диктування (МУПД). Вони є ієрархічно організованими. Перша модель базується на так званій генеративній моделі розпізнавання, розуміння та синтезу усномовного сигналу. При цьому синтез мовлення використовується як зворотній зв'язок в процесі розпізнавання. Три основні блоки МУПД такі: 1) модель зовнішнього світу, яка описує всі можливі смисли, що передаються в процесі діалогу, 2) генератор текстів або речень природної мови і 3) генератор фонетично-акустичних прототипів (модельних сигналів) усної мови. Проблема МУПД сформульована як 1) проблема *знаходження* для сигналу, що розпізнається, *найбільш схожого на нього модельного сигналу злитої мови* з множини всіх можливих сигналів прототипів, породжених третім блоком, для всіх можливих текстів і речень природної мови, що генеруються другим блоком для всіх можливих смислів, що передаються першим блоком, а також як 2) проблема *аналізу (розбору) останнього* як послідовності слів та канонічної форми смислу, що передаються мовним сигналом. Отримані послідовності слів є граматично і семантично правильними, тому можуть бути надруковані диктувальною машиною або синтезовані каналом третього блоку для іншої мови в машині для усного перекладу.

Другий шлях створення МУПД — так звана багаторівнева модель з багатозначними рішеннями. Тут не використовується синтез мовлення в якості зворотнього зв'язку. Натомість вводяться багатозначні рішення за спрощених умов на всіх рівнях ієрархії оброблення мовного сигналу. Наприклад, на першому рівні вирішується узагальнена проблема розпізнавання злитого мовлення, яка полягає в тому, що виходячи з припущення про вільний порядок слів знаходяться $N1>>1$ найкращих послідовностей слів. Потім на другому рівні ці багатозначні послідовності аналізуються, допоки не знайдуться послідовності слів, що збігаються з породженими генератором текстів або речень природної мови. Таким чином, на другому, найвищому рівні отримуємо $N2>>1$ найкращих результатів розуміння, з них остаточно обирається один найкращий.

Далі пропонується алгоритм багатозначної смислової інтерпретації усної мови для МУПД. Він усуває недоліки та поєднує переваги двох вищезгаданих моделей.

Цей алгоритм полягає в тому, що, наприклад, у дворівневій багатозначній МУПД на першому рівні вирішується узагальнена проблема розпізнавання не за умов припущення про вільний порядок слів, але в умовах розгляду $N>>1$ найкращих послідовностей слів, або точніше $N>>1$ початкових підпослідовностей слів, що мають бути допустимими в мові усного діалогу, тобто відповідати синтаксису, семантиці та прагматиці усного діалогу. Тому для поточного відліку часу будемо знаходити лише $N>>1$ найкращих початкових послідовностей слів, котрі допустимі в мові усного діалогу, а на наступному часовому кроці розглядатимемо тільки ті слова, які можуть продовжувати вже відібрані підпослідовності слів, які є допустимими в мові усного діалогу. Таким чином, вводимо багатозначну смислову інтерпретацію усномовного сигналу.

В результаті досягається така ж точність розуміння, як і в другій моделі, але за значно менших обсягів обчислень та пам'яті.

Далі в деталях описується основний алгоритм багатозначної смислової інтерпретації.

2. Специфікації природної мови

Описуватимемо усну природну мову за допомогою семантичної мережі. Використаємо найпростіший метод. Всі можливі речення

розміститимо в предметних полях. В свою чергу всі речення кожного предметного поля (ПП) розділимо на категорії на основі смислів, що передаються. Кожному предметному полю відповідає порівняно невелике число категорій смислів (КС).

Наприклад, для довідкової служби аеропорту можна виділити такі категорії смислів: питання про приліт та відліт літаків; питання про наявність квитків; питання стосовно маршруту; питання про розташування служб аеропорту тощо.

Кожна категорія смислів (КС) складається з власної множини типів речень. Тип речення (ТР) — це конструкція, що економно задає множину речень, які отримуються з одного речення шляхом незалежних замін та переставлянь як окремих слів, так і словосполучень. Базовим елементом ТР є підсловник. Підсловники в ТР іменуються згідно семантики предметного поля.

Кожна КС має порівняно невелику кількість ТР. Вочевидь, КС можуть при потребі поповнюватися новими ТР.

Всі ТР легко задаються списочними мовами, наприклад *LISP*.

Наводимо приклад ТР для питань, що стосується різниці двох чисел для української мови:

((ЧОМУ / TO WHAT / СКІЛЬКИ / HOW MUCH) (СКАЖІТЬ / SAY / ДАЙТЕ ВІДПОВІДЬ / ANSWER / --) (БУДЬ ЛАСКА / PLEASE / БУДЬТЕ ЛАСКАВІ / PLEASE / БУДЬТЕ ДОБРІ / BE KIND TO / --))

(БУДЕ ДОРІВНЮВАТИ / WILL BE EQUAL / ДОРІВНЮВАТИМЕ / WILL BE EQUAL / ДОРІВНЮЄ / IS EQUAL) (РІЗНИЦЯ / THE DIFFERENCE / ЗАЛИШОК / THE RAMAINDER)

(ДВОХ ЧИСЕЛ / OF THE TWO NUMBERS / ЧИСЕЛ / OF THE NUMBERS / --) (ВІД *A* / FROM *A*) [ВІДНЯТИ / SUBTRACT / ЗАБРАТИ / TAKE AWAY] *B* / *B* / *A* МІНУС *B* / *A* MINUS *B*)

Круглі дужки ( ) містять підсловники, які можна переставляти, тоді як квадратні [ ] містять неінверсні підсловники. Підсловники можна переставляти лише в межах "старших" дужок ( ). Як правило, ТР параметризуються. В цьому прикладі параметрами є операнди *A* і *B*. Символ "--" означає порожнє слово.

Неважко переконатися, що навіть якщо не брати до розгляду розмаїтість операндів *A* і *B*, наведений ТР задає загалом $6!\cdot(2\cdot(2\cdot(3\cdot4))\cdot3\cdot2\cdot3\cdot3) = 1\,866\,240$ різних допустимих в усній українській мові речень стосовно питання про різницю двох чисел. Серед них, зокрема, такі речення:

ЧОМУ СКАЖІТЬ ДОРІВНЮЄ РІЗНИЦЯ ЧИСЕЛ *A* МІНУС *B*,

ЧОМУ ДОРІВНЮВАТИМЕ РІЗНИЦЯ *A* МІНУС *B* СКАЖІТЬ БУДЬТЕ ЛАСКАВІ,

СКІЛЬКИ ВІД *A* ВІДНЯТИ *B* БУДЕ ДОРІВНЮВАТИ ЗАЛИШОК ДВОХ ЧИСЕЛ,

ВІД *A* ВІДНЯТИ *B* ЧОМУ ДОРІВНЮЄ ЗАЛИШОК ДАЙТЕ ВІДПОВІДЬ.

Нижче наведено інший приклад ТР для КС стсовно різниці двох чисел:

(α [ЗМЕНШУВАНЕ *A* / MINUEND *A*] β [ВІД'ЄМНИК *B* / SUBTRAHEND *B*] δ (γ [ЧОМУ / TO WHAT / СКІЛЬКИ / HOW MUCH] [БУДЕ ДОРІВНЮВАТИ / WILL BE EQUAL / ДОРІВНЮВАТИМЕ / WILL BE EQUAL / ДОРІВНЮЄ / IS EQUAL] ξ (ε [РІЗНИЦЯ / THE DIFFERENCE / ЗАЛИШОК / THE RAMAINDER] [ДВОХ ЧИСЕЛ / OF THE TWO NUMBERS / ЧИСЕЛ / OF THE NUMBERS / --]))

Для цього ТР підсловники іменуються так: α — зменшуване, β — від'ємник, γ — питальне слово, δ позначає дію, ε — операція і ξ — об'єкт дії.

КС і ТР використовуватимуться в процесі багатозначної смислової інтерпретації злитого мовлення. Тут слід підкреслити, що структури ТР зручні для генерування слів, що продовжують допустимі початкові підпослідовності слів.

3. Узагальнена проблема розпізнавання послідовності слів для злитого мовлення

Спершу розглянемо узагальнену проблему розпізнавання злитого мовлення, що складається зі слів вибраного словника [1, 2]. Потім цей результат узагальнимо для знаходження $N>>1$ смислів, що передаються злитим мовленням [3, 4].

3.1. Постановка задачі

Узагальнена проблема розпізнавання злитого мовлення полягає в тому, що виходячи з припущення про вільний порядок слів, знаходяться $N>>1$ найкращих різних послідовностей слів, впорядкованих за спаданням подібності до оброблюваного сигналу.

Нехай задані такі дані та знання:

A. Скінченна множина E елементарних прототипів усномовного сигналу (ЕПУС) $e(k^1) \in E$, де $k^1 \in K^1$ — ім'я ЕПУСу в алфавіті імен K^1. Наприклад, маємо $|K^1| = |E| = 2^{10}$ елементів у E і K^1. Отже, множина K^1 складає 1-й рівень модельних сигналів (мікрофонем), а пара (K^1, E) є кодовою книгою.

B. Скінченна множина K^2 — множина фонем-трифонів (ФТ) $k^2 \in K^2$ (ФТ формують 2-й рівень ієрархії модельних сигналів). ФТ є базовою фонемою, що розглядається в контексті впливу сусідніх фонем: першої, що передує, і другої, наступної. Для кожної природної мови фіксується біля 2000—3000 базових ФТ. Кожний ФТ k^2 з K^2 задається своєю транскрипцією в алфавіті K^1:

$$k^2 = \left(k_1^1, k_2^1, \ldots, k_s^1, \ldots, k_{q(k^2)}^1\right),$$

де s позначає порядковий номер в транскрипції та $q(k^2)$ — тривалість транскрипції для k^2.

C. Словник K^3 слів $k^3 \in K^3$, кожне з яких описане фонемною транскрипцією k^3 або фонемно-трифонною транскрипцією k^3, заданою в алфавіті K^2:

$$k^3 = \left(k_1^2, k_2^2, \ldots, k_s^2, \ldots, k_{q(k^3)}^2\right),$$

де $q(k^3)$ — довжина транскрипції слова.

D. Розподіли $P(x/k^1)$ спостережуваних елементів x для всіх $k^1 \in K^1$, зокрема для всіх $k^1 \in K^1$:

$$P(x/k^1) = P(x/e(k^1)).$$

Дані та знання, розглянуті в A, B, C, D, знаходяться в режимі навчання розпізнаванню [3]. При цьому формується так званий усномовний файл диктора.

Сигнал, який слід розпізнати, позначається послідовністю X_{0l} спостережуваних елементів x_i:

$$X_{0l} = (x_1, x_2, \ldots, x_i, \ldots, x_l),$$

де елементи x_i спостерігаються в рівновіддалених або майже рівновіддалених часових відліках i, а l — тривалість сигналу. Сегмент

$$X_{uv} = (x_{u+1}, x_{u+2}, \ldots, x_v), \quad 0 \le u < v \le l$$

розглядається як реалізація сигналу фонеми-трифона $k^2 = \left(k_1^1, k_2^1, \ldots, k_s^1, \ldots, k_{q(k^2)}^1\right)$ з імовірністю

$$P\left(X_{uv}/k^2\right) = \max_{\{t_s\}} \prod_{s=1}^{q(k^2)} \prod_{i=t_{s-1}+1}^{t_s} P\left(x_i/k_s^1\right),$$

де $t_0 = u$, $t_{s-1} < t_s$, $t_{q(k^2)} = v$. Отже, ця імовірність обчислюється згідно з фонемно-трифонною транскрипцією k^2 та як згортка за границями мікрофонем $\{t_s\}$.

Стохастична автоматна породжувальна граматика (граф) для порівняння спостережуваного сегмента X_{uv} з усіма прототипами, що генеруються для ФТ k^2, зображена на рис. 1a. Цей граф має $q(k^2)$ станів. Кожному станові s приписується мікрофонема k_s^1 або $k^1(s)$. Переходи здійснюються згідно стрілок за 0 або 1 часових кроків.

Зображуючи граф фонеми-трифона пунктирним прямокутником і об'єднуючи графи фонем-трифонів в лінійну послідовність згідно транскрипції слова k^3, отримуємо стохастичну автоматну породжувальну граматику для генерування прототипів (модельних сигналів) слів і їх порівняння з розпізнаваним сигналом. Граф слова зображено як "пелюстку" "квітки" на рис. 1b. Граф-квітка на рис. 1b представляє стохастичну автоматну породжувальну граматику для складання модельних сигналів злитого мовлення за умови вільного порядку слів. Кількість пелюсток збігається з кількістю слів у словнику.

Імовірність сегмента X_{uv} за умови слова k^3 обчислюється згідно з транскрипцією слова $k^3 = \left(k_1^2, k_2^2, \ldots, k_s^2, \ldots, k_{q(k^3)}^2\right)$ як згортка за границями фонем-трифонів $\{w_s\}$:

$$P\left(X_{uv}/k^3\right) = \max_{\{w_s\}} \prod_{s=1}^{q(k^3)} P\left(X_{w_{s-1}w_s}/k_s^2\right),$$

де $w_0 = u$, $w_{s-1} < w_s$, $w_{q(k^3)} = v$.

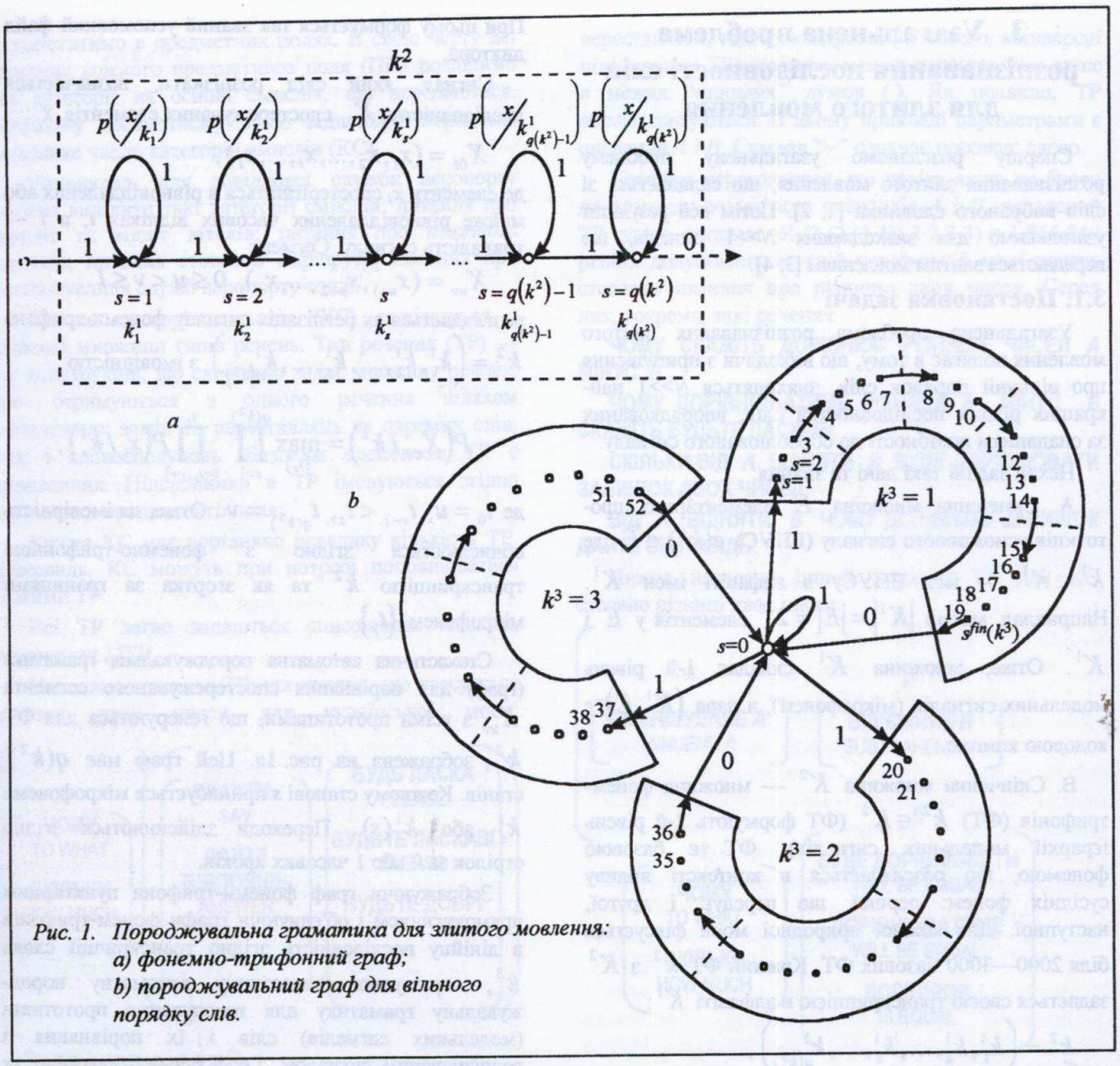


Рис. 1. Породжувальна граматика для злитого мовлення:
a) фонемно-трифонний граф;
b) породжувальний граф для вільного порядку слів.

Відповідно, найкраща відповідь розпізнавання у вигляді послідовності слів за умови їх вільного порядку і невідомої кількості слів Q у послідовності визначається максимізацією наступного виразу правдоподібності (1):

$$P\left(X_{0l}/k_1^3,\ldots,k_s^3,\ldots,k_Q^3\right)=\max_{\{m_s\}}\prod_{s=1}^{Q}P\left(X_{m_{s-1}m_s}/k_s^3\right)$$

де $m_0=0$, $m_{s-1}<m_s$, $m_{q(k^3)}=l$ — границі слів в усномовному сигналі X_{0l}.

3.2. Узагальнений алгоритм розпізнавання усної мови

Щоб сформулювати узагальнений алгоритм розпізнавання і таким чином знайти $N>>1$ найкращих послідовностей слів, введемо наскрізну нумерацію станів для графа злитого мовлення (див. рис. 1b). Розрізнятимемо головний стан s=0 для паузи, а також опустимо індекс 3 в записі k^3. Отже, лінійки станів s = 1:19, s = 20:36 і s = 37:52 представляють слова k = 1, k = 2 і k = 3 відповідно.

Окремо виділимо вхідні стани слів $s^{in}(k)$ і вихідні $s^{fin}(k)$. Це s = 1, 20, 37 and s = 19, 36, 52 відповідно.

Нехай $\Omega_i(s)$ позначає множину модельних сигналів злитого мовлення довжини i, що генеруються графом злитого мовлення в результаті переходів зі стану s=0 до стану s за i кроків часу. Позначимо через $\left(F_i^r(s),\mathcal{K}_i^r(s)\right)$, $r=1:N$ N-ку найкращих імовірностей $F_i^r(s)$, $r=1:N$ (1), що досягаються на множині $\Omega_i(s)$, але для початкового сегмента сигналу $X_{0i}=(x_1,x_2,\ldots,x_i)$, а через $\mathcal{K}_i^r(s)$, $r=1:N$ позначимо відповідну оптимальну

підпослідовність слів. В N-ці $\left(F_i^r(s), \mathcal{K}_i^r(s)\right)$, $r=1:N$ всі $\mathcal{K}_i^r(s)$ є різними. Нехай пари $\left(F_i^r(s), \mathcal{K}_i^r(s)\right)$ в N-ці $\left(F_i^r(s), \mathcal{K}_i^r(s)\right)$, $r=1:N$ розташовані в порядку спадання величини $F_i^r(s)$:

$$F_i^1(s) \geq F_i^2(s) \geq \ldots \geq F_i^r(s) \geq \ldots \geq F_i^N(s).$$

Нехай N-ка $\left(F_v^r(s), \mathcal{K}_v^r(s)\right), r=1:N$ вже обчислена для всіх станів s і для всіх відліків часу $v<i$, котрі передують i.

Тоді після появи наступного елемента x_i в момент часу i для всіх станів s обчислюється нова N-ка $\left(F_i^r(s), \mathcal{K}_i^r(s)\right), r=1:N$:

а) по-перше, для всіх внутрішніх станів слів s (окрім вхідних станів слів $s^{in}(k)$, це s=1, 20, 37 на рис. 1b, і головного стану s=0):

спочатку обчислюються всі $2N$ можливі добутки $F_{i-1}^w(s-1)P(x_i / k^1(s))$, $F_{i-1}^t(s)P(x_i / k^1(s))$, $w,t=1:N$ і виписуються відповідні $2N$ послідовності слів $\mathcal{K}_{i-1}^w(s-1)$, $\mathcal{K}_{i-1}^t(s), w,t=1:N$, і тоді формується нова N-ка $\left(F_i^r(s), \mathcal{K}_i^r(s)\right)$, $r=1:N$ шляхом вибору N найкращих добутків $F_i^r(s)$, що впорядковуються, з відповідними різними $\mathcal{K}_i^r(s)$ у цій новій N-ці;

b) по-друге, для всіх вхідних станів слів $s^{in}(k)$, $k \in K$, це s=1, 20, 37 на рис. 1b:

спочатку обчислюються всі $2N$ можливі добутки $F_{i-1}^w(0)P\left(x_i/k^1(s)\right)$, $F_{i-1}^t\left(s^{\text{in}}(k)\right)P\left(x_i/k^1(s)\right)$, $w,t=1:N$ і виписуються відповідні $2N$ послідовності слів $\mathcal{K}_{i-1}^w(0)$, $\mathcal{K}_{i-1}^t\left(s^{\text{in}}(k)\right)$, $w,t=1:N$, і тоді нова N-ка $\left(F_i^r(s), \mathcal{K}_i^r(s)\right)$, $r=1:N$ формується шляхом вибору N найкращих впорядкованих добутків $F_i^r(s)$ з відповідно різними $\mathcal{K}_i^r(s)$ в цій новій N-ці;

c) по-третє, для головного стану s=0:

спершу виписуються та обчислюються всі можливі $(|K|+1)N$ значення $F_i^w\left(s^{fin}(k)\right)$, $w=1:N$, $k=1:|K|$ та $F_{i-1}^u(0)P\left(x_i/k^1(0)\right)$, $u=1:N$ ($s^{fin}(k)$ є вихідними станами слів, це s=19, 36, 52 на рис. 1b) і формуються $(|K|+1)N$ нових послідовностей слів $\mathcal{K}_i^w\left(s^{fin}(k)\right)\oplus k$, $w=1:N$, $k=1:|K|$ і виписуються $\mathcal{K}_{i-1}^u(0)$, $u=1:N$, відповідно (тут позначка $\oplus$ означає додавання нового слова до послідовності слів), і тоді нова N-ка $\left(F_i^r(0), \mathcal{K}_i^r(0)\right)$, $r=1:N$ знаходиться шляхом вибору N найкращих і впорядкованих $F_i^r(0)$ з відповідно різними $\mathcal{K}_i^r(0)$ в новій N-ці.

Узагальнена відповідь розпізнавання визначається N-кою, обчисленою для головного стану s = 0 в момент часу i = l: $\mathcal{K}_i^r(0)$, $r=1:N$.

На початку алгоритму покладаємо: $F_0^1(0)=1$, $F_0^r(0)=0$ для $r=2:N$, $F_0^r(s)=0$ для всіх $r=1:N$ і всіх $s\neq 0$, $\mathcal{K}_0^r(s)=\varnothing$ для всіх $r=1:N$ і всіх s.

Неважко помітити, що поки розглядаються внутрішні стани слова, нові слова лише "розгортаються", а не додаються до підпослідовностей слів $\mathcal{K}_i^r(s)$, тоді як тільки при переході з вихідного стану слова $s^{fin}(k)$ до головного стану $s=0$ трапляється, що нові слова додаються до вже накопичених підпослідовностей слів $\mathcal{K}_i^r(s)$.

4. Алгоритм багатозначної смислової інтерпретації

Алгоритм багатозначного розпізнавання злитого мовлення подано в такій формі, що робить можливим поширення цього алгоритму на смислову інтерпретацію злитого мовлення.

Використовуватимемо поняття типів речень (ТР) і категорій смислів (КС), введених в розділі 2. Зокрема, структура ТР дає просте правило розпізнавання, чи може якесь слово продовжити вже накопичену послідовність слів, або чи є допустимою певна послідовність слів для даного ТР.

Нехай $\Lambda\left(\mathcal{K}_i^r(s)\right)$ — підсловник слів, що можуть продовжити початкову послідовність слів $\mathcal{K}_i^r(s)$. Коли $\mathcal{K}_i^r(s)$ заповнена і не може бути продовжена, тоді $\Lambda\left(\mathcal{K}_i^r(s)\right)$ дорівнює всьому словнику K, оскільки можливе започаткування нових послідовностей слів.

Аналізуючи підпослідовність слів $\mathcal{K}_i^r(s)$ за всіма структурами ТР і КС, неважко знайти послідовність смислів $\mathrm{M}\left(\mathcal{K}_i^r(s)\right)$, що передається сигналом злитого мовлення X_{0i}.

Тепер, посилаючись на розділ 3, стисло подамо алгоритм багатозначної смислової інтерпретації злитого мовлення.

Нехай $\left(F_i^r(s),\mathcal{K}_i^r(s),\mathrm{M}_i^r(s)\right), r=1:N$ є N-кою найкращих імовірностей $F_i^r(s), r=1:N$ (1), котрі досягаються на множині прототипів злитого мовлення $\Omega_i(s)$ для початкового сигналу $X_{0i}=(x_1,x_2,\ldots,x_i)$, і $\mathcal{K}_i^r(s), r=1:N$ — відповідні підпослідовності слів, що є різними, і $\mathrm{M}_i^r(s)=\mathrm{M}\left(\mathcal{K}_i^r(s)\right), r=1:N$ — відповідні оптимальні результати багатозначної смислової інтерпретації $\left(F_i^r(s),\mathcal{K}_i^r(s),\mathrm{M}_i^r(s)\right), r=1:N$ в N-ці, впорядкованій за спаданням $F_i^r(s)$.

Нехай N-ки $\left(F_v^r(s),\mathcal{K}_v^r(s),\mathrm{M}_v^r(s)\right), r=1:N$ вже обчислені для всіх станів s і для всіх відліків часу $v<i$, що передують i.

Тоді після появи наступного спостережуваного елементу x_i в момент часу i для всіх станів s обчислюється нова N-ка $\left(F_v^r(s),\mathcal{K}_v^r(s),\mathrm{M}_v^r(s)\right)$, $r=1:N$:

a) по-перше, для внутрішніх станів слів s спочатку обчислюються всі можливі $2N$ добутки $F_{i-1}^w(s-1)P(x_i/k^1(s))$, $F_{i-1}^t(s)P(x_i/k^1(s))$, $w,t=1:N$ і виписуються відповідні $2N$ послідовності $\left(\mathcal{K}_{i-1}^w(s-1),\mathrm{M}_{i-1}^w(s-1)\right)$, $\left(\mathcal{K}_{i-1}^t(s),\mathrm{M}_{i-1}^t(s)\right)$, $w,t=1:N$ слів і смислів, і тоді нова N-ка $\left(F_i^r(s),\mathcal{K}_i^r(s),\mathrm{M}_i^r(s)\right), r=1:N$ формується шляхом вибору N найкращих добутків $F_i^r(s)$, що впорядковуються, з відповідно різними $\mathcal{K}_i^r(s)$ в цій новій N-ці;

b) по-друге, для всіх вхідних станів слів $s^{in}(k)$, $k\in K$ спершу обчислюються всі можливі $2N$ добутки $F_{i-1}^w(0)P\left(x_i/k^1(s)\right)$, $F_{i-1}^t\left(s^{\mathrm{in}}(k)\right)P\left(x_i/k^1(s)\right)$, $w,t=1:N$, і виписуються відповідні $2N$ послідовності слів і смислів $\left(\mathcal{K}_{i-1}^w(0),\mathrm{M}_{i-1}^w(0)\right)$, $\left(\mathcal{K}_{i-1}^t\left(s^{\mathrm{in}}(k)\right),\mathrm{M}_{i-1}^t\left(s^{\mathrm{in}}(k)\right)\right)$, $w,t=1:N$, і тоді нова N-ка $\left(F_i^r(s),\mathcal{K}_i^r(s),\mathrm{M}_i^r(s)\right)$, $r=1:N$ формується шляхом вибору N найкращих і впорядкованих добутків $F_i^r(s)$ з відповідними різними $\mathcal{K}_i^r(s)$ в цій новій N-ці;

c) по-третє, для головного стану s=0 спершу виписуються та обчислюються всі можливі $(|K|+1)N$ значень $F_i^w\left(s^{fin}(k)\right)$, $w=1:N$, $k=1:|K|$ і $F_{i-1}^u(0)P\left(x_i/k^1(0)\right)$, $u=1:N$, і формуються та беруться до уваги всі можливі нові підпослідовності слів і результатів смислової інтерпретації $\left(\mathcal{K}_i^w\left(s^{fin}(k)\right)\oplus k,\mathrm{M}\left(\mathcal{K}_i^w\left(s^{fin}(k)\right)\oplus k\right)\right), w=1:N$, $k\in\Lambda\left(\mathcal{K}_i^w\left(s^{fin}(k)\right)\right)$ і $\left(\mathcal{K}_{i-1}^u(0),\mathrm{M}_{i-1}^u(0)\right)$, $u=1:N$ відповідно, і тоді знаходиться нова N-ка $\left(F_i^r(0),\mathcal{K}_i^r(0),\mathrm{M}_i^r(0)\right), r=1:N$ шляхом вибору N найкращих і впорядкованих $F_i^r(0)$ з відповідними різними $\mathcal{K}_i^r(0)$ в цій новій N-ці.

Відповідь багатозначної смислової інтерпретації визначається N-кою, обчисленою для головного стану $s=0$ в момент часу $i=l$: $\left(\mathcal{K}_l^r(0),\mathrm{M}_l^r(0)\right), r=1:N$.

На початку алгоритму багатозначної смислової інтерпретації покладаємо: $F_0^1(0)=1$, $F_0^r(0)=0$ for $r=2:N$, $F_0^r(s)=0$ для всіх $r=1:N$ і всіх $s\neq 0$, $\mathcal{K}_0^r(s)=\varnothing$ та $\mathrm{M}_0^r(s)=\varnothing$ для всіх $r=1:N$ і всіх s.

5. Прикінцеві положення

Алгоритм багатозначної смислової інтерпретації злитого мовлення оперує лише з допустимими в діалозі початковими підпослідовностями слів, і навіть більше: щоб уникнути факту локального рішення, багатозначний розв'язок $N>>1$ вводиться в кожний момент часу.

Очевидно, що більш раціонально змінювати кількість багатозначих рішень N, починаючи з великих чисел, зменшуючи їх в процесі накопичення інформації при обробленні усномовного сигналу.

Алгоритм багатозначної смислової інтерпретації злитого мовлення не гарантує глобального розв'язку проблеми розуміння, але за певного вибору кількості смислів N в реальному комп'ютерному середовищі знаходиться прийнятний прагматичний результат.

ЛІТЕРАТУРА

1. T.K. Vintsiuk. *Speech Recognition and Understanding.* — Kibernetika, 1982, No. 5, pp 101-111.
2. Т.К. Винцюк. *Обобщенная задача распознавания слитной речи.* — Труды VIII Всесоюзного семинара "Автоматическое распознавание слуховых образов 1982", Киев, 1982, с. 345-348.
3. Т.К. Винцюк. *Анализ, распознавание и смысловая интерпретация речевых сигналов.* — Киев: Наукова думка, 1987, 264 с.
4. Taras K.Vintsiuk. *Two Approaches to Create a Dictation/Translation Machine.* Proceedings of the Second International Workshop "Speech and Computer", Cluj-Napoca, 1997, pp 1-6.

Робастні алгоритми верифікації особи за голосом, що призначені для роботи в умовах сильних завад та спотворень мовних повідомлень

І. І.Горбань, А. В. Клименко

Інститут проблем математичних машин і систем Національної академії наук України
252187, Україна, Київ, Пр. Академіка Глушкова, 42
Тел. (044) 2666174, Факс (044) 4468129, E-mail: gorban@immsp.kiev.ua, klimenko@immsp.kiev.ua

У цій доповіді представлені нові робастні алгоритми обробки мовних сигналів, спрямовані на підвищення якості роботи алгоритмів автоматичної верифікації особи за голосом в умовах сильних частотних спотворень сигналу та потужних завад. Доповідь складається з чотирьох розділів. У вступній частині наведено обгрунтування необхідності застосування робастних алгоритмів обробки сигналу для здійснення ефективної верифікації в реальних умовах експлуотації систем автоматичної верифікації особи за голосом. Описання робастних алгоритмів верифікації особи за голосом наведено в другому розділі. Третій розділ присвячено результатам тестування цих алгоритмів. Краткі висновки наведено у четвертому розділі.

1. Вступ

Умови експлуатації систем верифікації особи за голосом дуже часто далекі від ідеальних. Мовні повідомлення, за якими здійснюється верифікація, як правило, сильно спотворені та зашумлені. Більшість відомих алгоритмів верифікації здатні працювати лише в умовах незначних завад та частотних спотворень [1-4].

В процесі розробки нової криміналістичної автоматичної системи верифікації та ідентифікації особи за голосом (CASVI) [5-8] авторами цієї доповіді були запропоновані і досліджені нові робастні алгоритми верифікації, що забезпечують стійку роботу в умовах сильних частотних спотворень сигналу та адитивних завад.

Мета доповіді – представити ці алгоритми і описати результати їх тестування.

2. Опис алгоритмів

2.1. Принципова схема роботи системи

При верифікації основними етапами обробки сигналів є: попередня обробка порівнюваних повідомлень (ПОП), розрахунок інформаційних ознак (РО), порівняння цих ознак (ПО) і прийняття рішення (ПР). На етапі порівняння інформаційних ознак та приняття рішення використовуються дані з бази даних (БД) (Рис. 1)

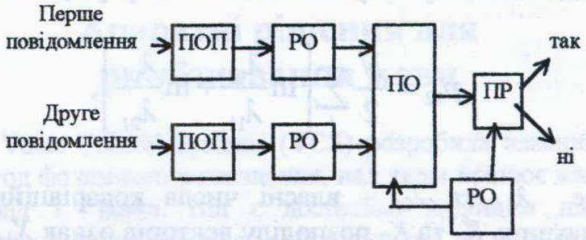


Рис. 1. Принципова схема роботи підсистеми верифікації системи CASVI

2.2. Попередня обробка повідомлень

На етапі попередньої обробки проводиться розбиття кожного повідомлення на фрейми тривалістю 20 мс. Ці фрейми розділяються системою на сигнальні, що підлягають подальшій обробці, і шумові, що виключаються. Ця процедура виконується на основі розрахунку оцінок верхньої та нижньої меж спектру сигнала і робастної обробки спектрів. Завдяки цим алгоритмам забезпечується відбір ідентичних фреймів як при відсутності так і при наявності спотворень та завад.

2.3. Розрахунок інформаційних ознак

Розрахунок інформаційних ознак передбачає для кожного фрейма обчислення 16-ти кепстральних компонент, а також двох перших кепстральних похідних. Для виключення залежності кепстрів від частотних спотворень сигналу введено їх центрування на вектор, що являє собою кепстр оцінки верхньої межі спектру сигналу. Для виключення розбіжностей в ознаках, що зумовлені різними шумовими умовами, введено нормування кепстрів. Суть нормування кепстрів полягає в штучному підвищенні рівня завад в кожному фреймі до

відношення сигнал-завада 12 дБ на кожній частоті з подальшим розрахунком кепстральних компонент.

2.4. Порівняння ознак та прийняття рішення

Порівняння ознак полягає в розрахунку відстані між порівнюваними повідомленнями. Ця відстань визначається шляхом об'єднання відстаней між окремими інформаційними ознаками.

Базою для вимірювання відстаней між ознаками обрано відношення правдоподібності. При цьому припускається, що ознаки мають гаусовий розподіл. При такому припущенні відстані можуть бути обчислені за формулою

$$p_{12} = \frac{N}{2}\sum_{l}\left[\ln\frac{\lambda_{0l}}{\lambda_{1l}} + \ln\frac{\lambda_{0l}}{\lambda_{2l}}\right],$$

де λ_{1l} та λ_{2l} – власні числа коваріаційних матриць K_1 та K_2 розподілу векторів ознак X_{1n} та X_{2n} (n = 1,...,N) відповідно для 1-го та 2-го повідомленнь. λ_{0l} – власні числа коваріаційної матриці K_0 сукупного розподілу векторів X_{1n} та X_{2n} (n = 1,...,N), l – порядковий номер кепстральної компоненти, N – кількість сигнальних фреймів, що обробляються.

Об'єднання одержаних відстаней здійснюється за допомогою бази даних.

Рішення приймається при порівнянні сукупної відстані між повідомленнями і порогом, що розрахований за допомогою бази даних.

3. Результати тестування алгоритмів

Для проведення комплексного тестування алгоритмыв верифікації використовувались 4 набора записів голосів 8 дикторів (усі диктори чоловічої статі) тривалістю від 1 до 2,5 хв. Використовувались записи, зроблені в разний час з перервами між записами від 2 тижднів до 3 місяців.

Базова тривалість фрагментів, по яких проводилась верифікація – 11,6 с чистого сигналу.

Перед початком проведення тестування робочі записи були спотворенні до 35 дБ, а також зашумлені адитивною завадою до відношення сигнал-завада 12дБ.

Всього було проведено 972 тести, з них 849 для різних дикторів і 123 для однакових. При цьому ймовірність правильного приняття рішення про співпадіння порівнюваних голосів виявилася рівною 93%, при цьому помилка ложної тревоги зафіксована на рівні 10%.

4. Висновки

Проведені дослідження дозволили сформувати алгоритми верифікації для системи CASVI. Завдяки використанню нових процедур була забезпечена ефективна робота системи в складних умовах: при наявності великих спотворень й сильних завад.

Література

1. B. S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proc. IEEE*, 64.– April.–1976.–P. 460-475.
2. S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. ASSP*, 29(2).–April.–1981.–P. 254-272.
3. H. Gish, M. Shmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, Oct.–1994.–P. 18-31.
4. I. I. Gorban, "Crime Automatic Speaker Verification and Identification (CASVI) System," *134th Meeting of ASA*, 102(5).–Pt.2.– Nov.–1997.–P. 3165.
5. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition," *IEEE Signal Processing Magazine*, Sept.–1996.–P. 58-71.
6. И. И. Горбань, Н. И. Горбань, А. В. Клименко, М. С. Хазанович, "Подсистема верификации новой криминалистической автоматической системы верификации и идентификации личности по голосу (CASVI)," *Математические машины и системы*, 2.–1997.–С. 61-64.
7. I. I. Gorban, "Text-Independent Speaker Verification Algorithms for Corrupted Signals," *Proc. Of 4th International Conference on Application of Computer Systems (ACS'97)*.–1997.–Oct.–P8.
8. I. I. Gorban, "Crime-Detection Speaker Verification and Identification System," *Proc. Of 9th annual International Conference on Signal Processing Application and Technology (ICSPAT'98)*.–1998.–Sept.–P5.

Комп'ютерні засоби розпізнавання усної мови

Вадим Данник
Міжнародний науково-навчальний центр інформаційних технологій та систем
40, просп. Акад. Глушкова, Київ 252022, Україна
Електронна пошта: vadym@technopark.kiev.ua

Абстракт

Описуються впроваджені програмні та апаратні системи для розпізнавання усної мови, коротко описано методи, на яких вони базуються та сфери застосування. Наведено порівняльні характеристики, вказано основні тенденції розвитку.

Вступ

Використання технології розпізнавання мови почалось з невеликих програм, що працювали з десятком слів та могли керуватись голосовими командами. Розвиток цієї технології сьогодні дозволяє без довгої процедури настроювання на голос диктора, без використовування коштовних мікрофонів чи спеціалізованого обладнання усно керувати побутовими пристроями, вдосконалювати системи обмеження доступу, розширяти можливості людей з фізичними вадами тощо. Системи зі спеціалізованими словниками "розв'язують руки" медикам, журналістам, юристам, секретарям та рядовим користувачам сучасних персональних комп'ютерів. Активно розробляються та набувають поширення автоматичні довідникові системи з доступом з телефоного апарату.

Так, наприклад, фірма Sharper Image виробляє голосові номеронабирачі для телефонів, AutoLink OnBoard випускає автомобільні прилади з голосовим керуванням. Наявність голосового інтерфейсу надає "інтелектуальності" автомобільним бортовим системам, побутовій електроніці - диктофонам, телефонам, відеосистемам, дитячим іграшкам і навіть вимикачам.

Програми розпізнавання усної мови для ПК

Найпопулярнішим на сьогоднішній день програмним продуктом для персональних комп'ютерів залишається DragonDictate від Dragon Systems. Невибагливий до апаратних ресурсів, цей високоінтегрований програмний продукт з нескладним інтерфейсом підтримує 6 мов, включно з російською. Як і для інших подібних програм, необхідно застосовувати низькошумові гостроспрямовані мікрофони та стежити за чіткою вимовою слів. Оскільки для здійснення неперервного надиктовування необхідна не тільки значна адаптація до вимови та інтонації диктора, але й урахування контексту, додатково випускаються спеціалізовані модулі словників для різних застосувань - для офісу та банків, радіологічної та екстремальної медицина, телефонії. Для розробників напрацьовано багатий програмний та апаратний інструментарій для створення мовного інтерфейсу для прикладних систем (навігація по комп'ютерному інтерфейсу, ввід данних, телефонні транзакції тощо).

Апаратні рішення для розпізнавання мови

Voice Control Systems (VCS) розробила власний метод фонемного розпізнання, над яким працює вже понад 17 років. Він є достатньо надійним для керування автомобільним обладнанням чи роботи з телефоном "без рук". Базові DSP – TI-C3X та C4X.

Для порівняння, остання версія методу неперервного розпізнавання усної мови працює на 1/2 33МГц TMS320C31. Реалізовано метод для сотового та безпровідного зв'язку, включено розпізнавання чисел на 15 мовах. Метод базується на модифікованому алгоритмі Маркова, за базову мовну одиницю прийнято фонему. Програма реалізації методу коштує від $500 в роздріб до $5 оптом.

Розроблено також фонетичний словниковий розпізнавач, де за базову мовну одиницю прийнято звукові прототипи фонем кожної мови. Наприклад, для американського діалекту англійської мови це 43 одиниці. Словник для розпізнавання складається з моделей слів, побудованих з цих одиниць, тому він може бути легко розширений. Набір фонем може бути також використаний для визначення меж слів. Програмне забезпечення працює на DSP TMS320C30, на частоті 55МГц може одночасно працювати два розпізнавачі.

Для прикладних задач з високими вимогами надійності (автомобільне обладнання, телефонний зв'язок) розроблено метод розпізнавання окремо вимовлюваних слів у дикторонезалежному, дикторозалежному та адаптивному варіантах. Є готові словники для більш як 45 мов, можливе створення розробниками власних словників. Ця технологія застосовується в розширювачах клавіатур та інтерактивній мультимедії, в телефонії - для автоматизації операторних функцій та голосового набору, в комп'ютерній телефонії - для голосового термінального доступу до персональних комп'ютерів, в побутовій електроніці - голосове

Таб. 1. Порівняльні характеристики апаратних та програмних систем розпізнавання голосових команд та мови.

Розмір словника	Підтримка злитного мовлення	Багато-дикторність	Точність розпіз-навання	Аналіз грама-тики	Працездатність в шумному середовищі	Виробник, ресурси, вартість
Інтегровані схеми та портативні пристрої						
40	Ні	Так			Так	Summa Group, HM2007, $16
25	Ні		97%		Так	OKI VRP6679, $20
50	Так	Так			Так	VCS, Голосові номеронабирачі 2060 на TMS320C5X
14-25	Так	Так	96%-99%		Так	Sensory Inc., RCS-164, до $10
300-10000	Так	Ні			Ні	Verbex Voice Systems, Speech Commander - портативний пристрій для ПК
Програмні продукти						
62000 … 230000	Так, 160 слів/хв	Адаптивне	95%-98%	Так	Ні	Dragon Systems, Dragon NaturallySpeaking, $60
20-100000	Ні	Так	95%-99%	Ні	Ні	IBM VoiceType, $100
20000-60000	Так, 140 слів/хв	Так	95%	Так	Ні	L&H VoiceXpress, $200
500-20000	Так	Так		Так	Ні	Cognitive Technologies

керування для іграшок, відеомагнітофонів, телевізорів. Цей метод реалізовано для широкого спектру мікропроцесорів та мікроконтролерів – Intel-X86, TI-C5X, C3X, C4X та C2X, OKI 6679, NEC-V20 та V30. Для порівняння, на 486DX-33 може одночасно працювати 8 розпізнавачів. Вартість - від $500 у роздріб до $1 великим оптом.

Голосовий номеронабирач 2060 від VCS - автономне обладнання для дикторонезалежного пофонемного розпізнавання окремих слів та компресії сигналів по CELP на базі DSP TMS320C5X. Модель 2060 розпізнає 50 імен та набирає відповідний телефонний номер. Інтерфейс розпізнає команди "call", "program", "list" та має голосовий зворотній зв'язок.

Фірма Sensory Inc пропонує рішення на базі RISC-мікроконтролерів без використання DSP. Застосування апарату нейроних мереж, на відміну від статистичних методів, дозволяє ефективніше вирішити проблему багатодикторності, діалектів мови, акценту чи навіть стану диктора - додатковим стендовим тренуванням системи, а не збільшенням обчислювальних ресурсів, додатковою адаптацією до голосу диктора чи модифікацією алгоритмів.

Використання узагальнених нейроних мереж для розпізнавання мови дає надто малу точність - до 80%, але застосування мережи спеціальної архітектури дозволяє підняти точність до 95% і навіть до 99% для спеціальних застосувань.

Головна перевага методу нейроних мереж - мінімальні вимоги до обчислювальних ресурсів та низька вартість апаратних реалізацій.

Інструментарій для розробників

Користувачам систем розпізнавання мови вже не треба "винаходити велосипед", в Інтернеті доступні вільно розповсюджувані програмні засоби для проведення експериментів та дослідження різноманітних алгоритмів. Доступні базові алгоритми розпізнавання за моделлю Маркова, рекурентними та нерекурентними нейронними мережами, програми для дослідження спектрограм слів та побудови словників фонем тощо.

Для ефективної інтеграції мовних технологій в різні операційні системи розроблені та активно поширюються серед розробників відповідні інтерфейси:

- ASAPI: Advanced Speech API (AT&T),
- SAPI: Microsoft Windows Speech API,
- SRAPI: Speech Recognition API,
- TAPI: Microsoft Windows Telephony API.

Підсумок

Загалом, розпізнавання мови досягло значних успіхів у певних застосуваннях, проте більшість важливих застосувань не може бути реалізовано на базі сьогоднішніх технологій. Розробляються напрямки нотування діалогів, синхронного усного перекладу, формування вимови для тих, хто вивчає іноземні мови. На жаль, переважна більшість розробок стосується у першу чергу англійської мови та вимагає значної адаптації або фундаментальної переробки для інших мов.

Важливим фактором для успішної реалізації мовних проектів залишаються проведення фундаментальних досліджень мовних сигналів та створення великого структурованого та фонетично розміченого тестового та навчального мовного матеріалу. Існуючі рішення для слов'янських мов ще не забезпечують точність більше ніж 95%, щоб отримати широке розповсюдження. Відсутність суттєвих досягнень за останні роки вказує на необхідність додаткової підтримки з боку фундаментальних наук, що досліджують людську мову.

МОДЕЛЮВАННЯ ПЕРЕКЛАДУ УСНОМОВНИХ СЛІВ

Пилипенко Валерій

Міжнародний науково-навчальний центр інформаційних технологій та систем

40, просп. Академіка Глушкова, Київ 252022

Електронна пошта: pilipenk@uasoiro.freenet.kiev.ua

Абстракт

Описується програмний комплекс, в який входять програми розпізнавання усної мови, перекладу її на іншу та озвучування відповіді. Комплекс дозволяє користувачу вимовляти в мікрофон слово рідною мовою, побачити переклад слова в текстовому редакторі і почути його. Також є сервісні функції, що дозволяють ввести нову мову для перекладу, словник і настроїтися на голос диктора. Окрім цього, є можливість контролювати процес введення звукового сигналу.

Реалізація програми перекладача

Програма, що моделює переклад усномовних слів, реалізована у вигляді текстового редактора, який дозволяє користувачеві створювати текст не лише шляхом введення символів за допомогою клавіатури, але і промовляючи слова в мікрофон. Використовується стандартний текстовий редактор, вбудований в операційну систему Windows 95.

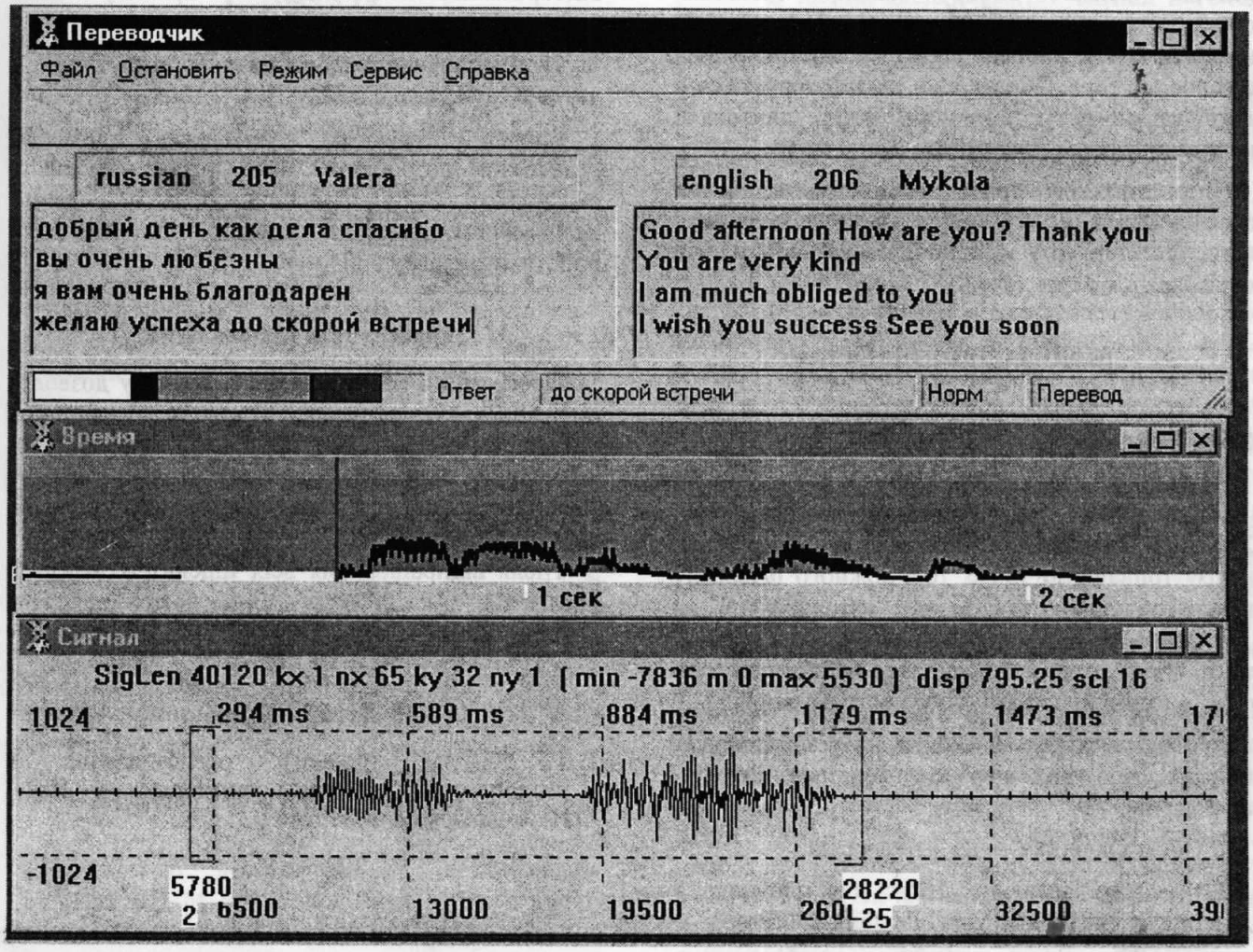


Рис. 1. Загальний вигляд програми перекладача усномовних слів. Панель текстового редактора поділена на дві частини – вхідний текст і переклад. Для контролю гучності вимовляння виведений індикатор гучності. При необхідності можна контролювати рівень гучності (середнє вікно). Також можна розглянути введений сигнал за допомогою нижнього вікна.

Вікно редактора поділене на дві частини — в лівій відображається текст фрази тією мовою, якою його вимовив користувач, в правій частині дається переклад на потрібну мову (рис. 1). В результаті, завжди можна водночас бачити текст на вхідній мові та перекладений текст.

Результат перекладу озвучується за допомогою заздалегідь записаних слів або синтезатора усної мови.

Настроювання на мову, словник і голос користувача

Для введення в систему нової мови і нового словника достатньо набрати в довільному текстовому редакторі словник потрібною мовою і зберегти його в текстовому режимі під іменем, яке буде відображати мову і словник згідно погоджень, прийнятих у системі.

Програма настроюється на мови розпізнавання і перекладу, на робочий словник і на голос диктора. Після вибору мови, словника, користувача і кількості слів пропонується вимовити кожне вибране слово для настроювання на голос диктора. Образи слів запам'ятовуються в індивідуальному усномовному файлі диктора і використовуються при розпізнаванні мови.

Після того, як проведене настроювання на голос диктора, є можливість використати базу даних його голосу для відповідей в процесі розпізнавання-перекладу.

Модуль розпізнавання

Програма реалізована мовою C/C++ в операційній системі Windows 95.

Вхідний сигнал, введений за допомогою Sound Blaster, перетворюється в цифровий вигляд (16 біт, 22 кГц) і рівномірно розбивається на інтервали аналізу тривалістю 15 мс. Для кожного інтервалу аналізу обчислюються вектор автокореляції та міри схожості з еталонними елементами з кодової книги. Після цього обчислюється найкраще слово згідно алгоритму динамічного програмування [1].

Проводилися експерименти на різноманітних вибірках, для яких програма показала наступні параметри:

1. Обсяг словника — до 4 тисяч слів.

2. Час розпізнавання 0.25 с для словника в одну тисячу слів для IBM PC Pentium-200 Pro.

3. Середня надійність розпізнавання не менше 90% для словника в одну тисячу слів.

Архітектура комплексу

Моделювання процесу розпізнавання, перекладу з однієї мови на іншу та видачі відповіді проводиться за допомогою комплексу, що складається з таких програмних блоків:

1. Введення-виведення мовного сигналу, визначення моментів початку та кінця мовного сигналу, забезпечення оброблення мовного сигналу в реальному часі на фоні введення-виведення мовного сигналу.

2. Формування індивідуального усномовного файлу диктора за навчальними вибірками.

3. Формування бази даних для відповідей.

4. Автоматичне розпізнавання усномовного сигналу.

5. Відображення відповіді розпізнавання в лівому вікні текстового редактора та часового сигналу в окремих вікнах.

6. Переклад з вхідної мови на потрібну і відображення перекладу в правому вікні текстового редактора.

7. Відтворення або синтез перекладу у вигляді звукового сигналу.

Блочна структура комплексу дозволяє незалежно розробляти його різні частини і об'єднувати їх у відповідності з фіксованим інтерфейсом взаємодії модулів для перевірки роботи комплексу в цілому.

Прикінцеві положення

Програмний комплекс для перекладу дозволяє відпрацьовувати різноманітні модулі у взаємодії з метою з'ясування вузьких місць і відпрацьовування нових алгоритмів розпізнавання мови, перекладу з однієї мови на іншу, а також синтезу усної мови.

Після відпрацювання всіх підсистем можлива реалізація комплексу для промислового використання.

Література

1. Т.К. Винцюк. Анализ, распознавание и интерпретация речевых сигналов. — Киев: Наукова думка, 1987, 264 с.

ЧАСОВА ТРАНСФОРМАЦІЯ МОВНИХ СИГНАЛІВ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Юрій Рашкевич, Роман Ткаченко, Зореслава Шпак

Державний університет "Львівська політехніка"
290646, м. Львів-13, вул. Ст. Бандери, 12, тел.398-793
електронна пошта: rashkev@polynet.lviv.ua

Наведено результати використання штучних нейронних мереж для задач перетворення часового масштабу мовних сигналів. Описана структура мережі та представлені результати експериментів прогнозування зміни тривалості мовних одиниць при сповільненні темпу відтворення мовної інформації.

1. ВСТУП.

Починаючи із кінця 80-х років, штучні нейронні мережі (ШНМ) знаходять широке застосування для розв'язування багатьох типів задач оброблення мовних сигналів, включаючи задачі розпізнавання слів та фраз, верифікації дикторів, виділення ключових слів тощо. Особливо перспективним є застосування ШНМ для оброблення сигналів із змінними в часі параметрами, оскільки завдяки особливій прогностичній здатності ШНМ часто з дивовижною точністю передбачають значення сигналу, чи його параметрів.

Дуже важливою для задач регулювання темпу мови, є властивість ШНМ акумулювати та використовувати в процесі роботи інформацію про кореляційні залежності між сусідніми сегментами сигналу, тобто відслідковувати взаємовплив тривалостей сусідніх сегментів, що необхідно для збереження збалансованої темпоральної структури слова в цілому не тільки на вході моделі (що є властивим для регресійних моделей), але й на виході.

2. МОДЕЛЬ ЕКСПЕРИМЕНТУ.

Метою дослідження є встановлення здатності ШНМ відслідковувати закономірності у зміні тривалостей звуків різних класів при сповільненні темпу мови з урахуванням кореляційного впливу як тривалості попереднього звуку, так і звуку наступного. Така постановка задачі у випадку прискорення розглянута в [1], де наведені результати прогнозування тривалостей стаціонарних ділянок мовного сигналу 5-шаровою ШНМ, в якій функції перетворення входів у виходи відрізняються від сигмоїдних і зображуються сплайном Ерміта. Мовний сигнал подавався у вигляді тривалостей біжучої і двох сусідніх (попередньої і наступної) стаціонарних ділянок. Незважаючи на те, що середня похибка дещо перевищувала 20 %, отримані результати підтвердили можливість і перспективність використання ШНМ в такого типу задачах.

В наших експериментах сигнал на вході ШНМ подавався у вигляді послідовності векторів:

$$(l_{i-1}, k_{i-1}, l_i, k_i, l_{i+1}, k_{i+1}),$$

де символом l позначені тривалості біжучого, попереднього та наступного звуків, а символом k - відповідні ознаки класифікації. На основі цих даних ШНМ прогнозувала тривалість біжучого звуку в сповільненому темпі.

Мовний сигнал сегментувався на окремі класи звуків згідно із запропонованим в [2] алго-ритмом сегментації та маркірування. Виділялися 5 класів звуків - наголошені голосні, ненаголо-шені голосні, вокалізовані приголосні, невокалі-зовані приголосні, вибухові звуки, а також міжслівні паузи.

В експериментах використана гетерогенна ШНМ з проективно-латеральними синаптичними зв'язками, яка відноситься до класу мереж прямого поширення Feed Forward. Мережа забезпечує відтворення складних поверхонь на навчальній множині даних як завгодно точно, однак оптимальні прогностичні властивості моделі встановлюються відповідним вибором параметрів ШНМ на основі зовнішнього критерію якості. Загалом процес вибору по суті відповідає концепції методу групового врахування аргументів О.Г.Івахненка та здійснюється в автоматичному режимі. Процес навчання такої ШНМ є неітераційним, здійснюється за час, що не перевищує декількох секунд, а принцип балансу точності відтворення забезпечує адекватне відображення закономірностей з одночасним вилученням чисто випадкових факторів.

3.РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТІВ.

Випробовування проводилися на мовних текстах, висловлених одним диктором у різних темпах: швидкому розмовному та сповільненому виразному. Загальний коефіцієнт зміни темпу мовлення складав 2,1. Обидва записи було просегментовано на ділянки, що відповідали звукам мови та паузам. Для кожної ділянки встановлювався її клас у відповідності із типом звуку.

Проведено дві серії експериментів: прогнозування зміни тривалостей звуків без вказування їх типів та передбачення з урахуванням класів звуків.

У першій серії при навчанні мережі на вхід ШНМ подавались трійки значень, які задавали тривалості трьох послідовних звуків у швидкому темпі, а на вихід - відповідні їм тривалості звуків при повільній вимові. Навчальна множина складалась із 100 наборів. Встановлено оптимальні для задачі такого класу параметри ШНМ: число асоціативних нейронів - 5, коефіцієнт нелінійності гіперповерхні процесу - 0,38. Результати передбачення для 25 наборів значень тривалостей звуків наступного за навчальним фрагменту мовного тексту наведені на рис. 1. Темні стовбчики відображають еталонні значення, я ясні - значення прогнозу. Середнє значення відхилення при передбаченні тривалостей звуків у випадку неврахування їх типів складало 31 мс, що відповідає 19,1% від усередненої тривалості звуків у повільному темпі. Проте, оскільки для різних звуків прогнозовані значення відрізнялись від еталонних як у сторону збільшення, так і в сторону зменшення, то загальне відхилення у передбаченні цілого фрагменту склало 7,2%.

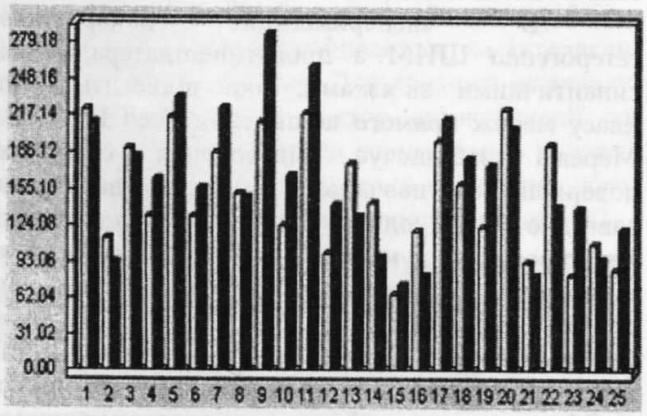


Рис.1. Прогонозування тривалостей звукових елементів без урахування ознак класифікації

На рис. 2 зображено гістограму результатів передбачення для тих самих навчальних та експериментальних наборів, але з вказанням класів звуків (на вході задавалися тривалість і клас кожного із трійки звуків). Аналіз результатів підтвердив, що введення класів звуків забезпечило вищу точність у передбаченні тривалостей як окремих звуків (середнє відхилення складало 21 мс або 12,9% від середньої тривалості звуку в еталонному тексті), так і для цілого фрагменту, використаного в експерименті (сумарна прогнозована тривалість відрізнялася від еталонної тільки на 2,4%). Тобто, з високою точністю був витриманий загальний коефіцієнт сповільнення темпу.

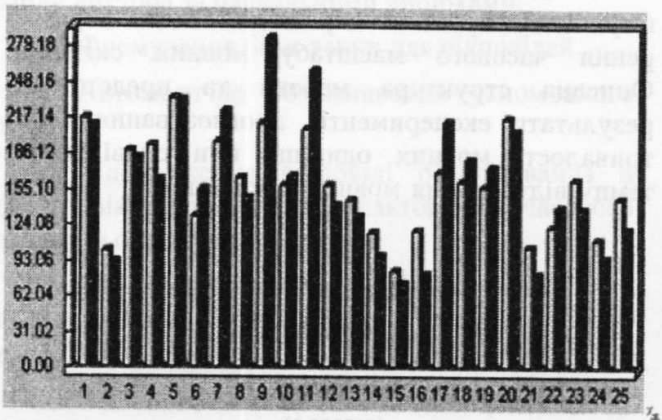


Рис.2. Прогнозування тривалостей звукових елементів з урахуванням класів звуків

При зменшенні навчальної вибірки до 50 наборів результати прогнозування різко погіршились - середнє відхилення у тривалостях звуків складало 23%, а для цілого есперимнтального фрагменту - 11,4%.

4. ВИСНОВКИ.

Отримані результати свідчать, що введення додаткової ознаки - класу звуку дозволяє суттєво підвищити прогностичні властивості нейронної мережі. Подальше покращення результатів може бути досягнуте шляхом введення в структуру ШНМ нелінійних синаптичних зв'язків. Передбачається також використання можливостей ШНМ для проведення класифікації звуків.

ЛІТЕРАТУРА.

1. Rashkevych Yu. Non-linear time-scale modification of speech by Neural Networks//Proc. of the Summer School on Neural Network Application to Signal Processing. - Czestochowa (Poland). - 1997. - P. 393-395.
2. Рашкевич Ю.М. Перетворення часового масштабу мовних сигналів. - Львів: Академічний експрес, 1997. - 140 с.

КОМП'ЮТЕРНІ ЗАСОБИ ДЛЯ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ МОВНОГО СИГНАЛУ

Микола Сажок

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 252022
Електронна пошта: mykola@uasoiro.freenet.kiev.ua

Program complex, presented in this paper, is a set of tools required for speech recognition, understanding, synthesis methods modelling and system architecture tests. All cycle of applied tasks is supported. Among them are speech signal recording, collecting, segmentation and natural languages knowledge accumulation. Implemented features allow setting experiments for speech signal transformations, pre-processing, recognition, understanding, and synthesis. The ways to make developed software open for custom research features are discussed as well as its compatibility with popular tools for scientific research.

1 Вступ

Комп'ютерне моделювання — необхідний етап у розробленні методів та тестуванні архітектур систем розпізнавання, розуміння та синтезу усної мови. Висновки щодо придатності, а також шляхи вдосконалення того чи іншого методу або архітектури системи випливають з результатів їх натурного моделювання [1].

Тому розроблення засобів експериментальних досліджень є життєво необхідним для розвитку усномовних технологій. На щастя, сучасні комп'ютерні технології якраз і дозволяють створювати складні, інтелектуалізовані дослідні станції. Отже, з'являється можливість зробити ці станції достатньо універсальними, придатними не лише для задач тої чи тої дослідницької групи, але й для будь-якого дослідника в ділянці усномовних технологій.

Програмний комплекс складається з двох основних підкомплексів: студія дослідника усномовного сигналу (ДУМС) та студія дослідника природніх мов (ДПМ).

Студія ДУМС — інтегроване робоче місце для опрацювання мовного сигналу, лінгвістично-фонетичних знань та індивідуальну інформацію щодо диктора. З іншого боку, за допомогою цього модуля проводяться експерименти з попереднього оброблення, розпізнавання та синтезу мовного сигналу.

Студія ДПМ забезпечує введення, накопичення та оброблення знань про природні мови, а також проведення експериментів з автоматичного розуміння мови та перекладу з однієї мови на іншу. Одна з компонент цієї студії відповідає за формування бази знань про конкретну природну мову, інша компонента задає певну предметну область, що є частиною моделі зовнішнього світу, яка вводиться для розуміння природної мови [2].

2 Студія дослідника усномовного сигналу

Студія ДУМС задумана як універсальне робоче місце для науковця-дослідника усномовниго сигналу. Студія дозволяє оперувати як самим акустичним сигналом, так і його лінгвістично-фонетичними атрибутами, враховувати індивідуальні особливості диктора.

Запис мовного сигналу проводиться в діалоговому режимі, частинами. Якщо якість певного блоку сигналу не задовольняє експерта, або ж про це сигналізують закладені в студію алгоритми, такий блок позначається як "поганий" з тим, щоби його перезаписати.

Наступним важливим кроком є автоматизована сегментація сигналу на синтагми, ритмогрупи, слова, склади, фонеми, мікрофонеми тощо. Експерт приймає, підправляє або відміняє результати автоматичного сегментування. Для підвищення ефективності роботи експерта забезпечується спрощений доступ до сегментів, напрклад, фонем (див. рис. 1). Отже, експерт може легко виділяти, проглядати та прослуховувати кожну ділянку сигналу, що відповідає певній фонемі.

Результати сегментації, коротка інформація про диктора зберігаються в стандартному звуковому (wave) файлі як окремий блок файлу (chunk), не пошкоджуючи, таким чином сам файл, залишаючи його доступним для інших програм. Більш докладна індивідуальна інформація може зберігатися в окремому файлі, причому посилання на нього зберігається в звуковому файлі.

Щойно усномовний сигнал відсегментовано, експерт накопичує сегменти фонем і оперує з ними як з окремими лінгвістино-акустичними одиницями, що економно задаються модельними елементами, якими є, наприклад, одно-квазіперіодні сегменти [3]. В бібліотеці сегментів фонем містяться посилання на акустичні дані, що зберігаються в звуковому файлі. Також передбачена можливість запису звукових даних безпосередньо в бібліотеку сегментів фонем.

Обслуговування всієї необхідної лінгвістично-фонетичної інформації відбувається наступним чином: кожному сегменту фонеми приписується ім'я фонеми, її фонемне оточення, що автоматично обчислюються за результатом сегментації. Крім того, запам'ятовуються фраза, позиція фонеми у фразі, найближчі границі синтагм, ідентифікатор диктора тощо.

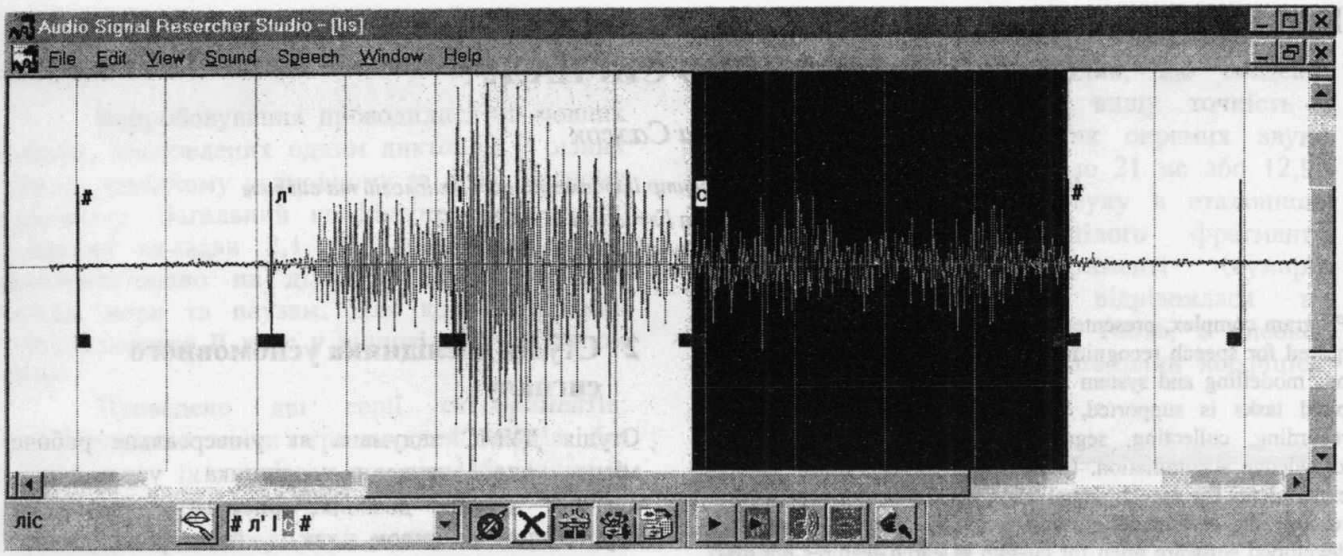


Рис. 1. Результат сегментації українського слова *ліс*. Кожний сегмент фонеми-трифона (ФТ) може бути виділений як за допомогою подвійного клацання мишкою на відповідній ділянці сигналу, так і шляхом вибору фонеми у вікні редагування (в даному випадку *с*). Потім експерт прослуховує виділене, коригує границі фонем, поповнює бібліотеку сегментів ФТ.

Експерт може вільно переміщати звукові файли до будь-якого іншого директорію, оскільки студія ДУМС за допомогою автоматизованої процедури поновлює зв'язки зі звуковими файлами.

Звукові сигнали можна редагувати, а саме: переставляти, вилучати, розтягувати в часі тощо.

Для обчислення границь одно-квазіперіодних сегментів (мікрофонем) використовується стійкий алгоритм, запропонований в [3]. Сегментація ж на ділянки, що відвідають окремим фонемам проводиться за аналізом одно-квазіперіодних сегментів, при цьому враховується інформацію про періодичність, тривалість, спосіб і місце творення фонеми.

Роль фонеми-трифона може виконувати не лише одна фонема в контексті з оточенням, а і послідовності фонем: склади, префікси, суфікси тощо.

Введена можливість одночасного візуального спостереження двох пар різних ділянок звукового сигналу, комбінуючи при цьому його різні представлення. Додатково забезпечується фільтрування сигналу, порівняння ділянок сигналу за різними мірами схожості, формантний аналіз, "зашумлення" сигналу тощо.

3 Студія дослідника природних мов

Студія ДПМ забезпечує введення, накопичення та оброблення знань про природні мови. Студія дозволяє також ставити експерименти в ділянці автоматичного розуміння мови та перекладу з однієї мови на іншу. Одна з компонент цієї студії відповідає за формування бази знань про конкретну природну мову, що фактично є словником з описом всіх словоформ, вживання слова в контесті з іншими тощо.

Кожній підтримуваній студією ДПМ мові відповідає динамічна бібліотека, що забезпечує функційні інтерфейси граматики та словника, роботу з текстами – всі операції, безпосередньо залежні від конкретної мови, а також взаємозв'язок з моделлю зовнішнього світу (МЗС). На даний момент лише для української мови існує версія такої динамічної бібліотеки.

За допомогою студії ДПМ експерт формує словник природної мови, зіставляючи слова зі складовими МЗС. Розроблено автоматизовану процедуру для знаходження в заданому тексті невідомих слів та додавання їх до словника в діалоговому режимі з користувачем.

МЗС породжує всі можливі канонічні форми заданої предметної області, за якими генеруються тексти вже для конкретної природної мови [1, 2].

4 Підсумок

Архітектура створеного програмного забезпечення дозволяє його широке використання в дослідженнях, тестах. Розробляється можливість додавати користувацькі функції, що дозволить будь-якій науково-дослідницькій групі проводити дослідження та тестувати свої алгоритми, методи.

Література

1. Т.К. Винцюк. *Анализ, распознавание и смысловая интерпретация речевых сигналов.* — Киев: Наукова думка, 1987, 264 с.
2. T.K. Vintsiuk. *Two Approaches to Create a Dictation/Translation Machine.* - Pros. of the 2nd Intern. Workshop "Speech and Computer", Cluj-Napoca, 1997, pp 1-6.
3. Taras K. Vintsiuk. *Optimal Joint Procedure for Current Pitch Period Discrimination and Speech Signal Partition into Qausi-Periodic and Non-Periodic Segments.* — Proc. Of the First Workshop on Text, Speech, Dialogue (TSD'98), Brno, 1998, pp. 135—140

Ідентифікація диктофонів за параметрами спрацьовування системи активації

Шевченко А. І., Старушко Д. Г.
Донецький державний інститут штучного інтелекту
Україна, 340048, м. Донецьк, вул. Артема, 118-б
Телефон, факс (0622) 926-082
E-mail: kis@iai.donetsk.ua

Article is denoted development and realization of strategy to identifications the dictaphones for the sound expert operations. For identifications is offered to use temporary parameters of operating a system to activation's by the voice. Principles of feature extraction are described In the article for the recognition, for what is offered single-purpose device on the base developed by authors numerical frequency meter. Article is kept test and experimental material.

ВСТУП

Під час проведення фоноскопічних експертиз перед експертами досить часто постає завдання ідентифікації технічних засобів, завдяки яким була виготовлена подана для дослідження фонограма. При цьому найбільш складним аспектом проблеми є вибір ідентифікуючих ознак, тобто компонентів фонограм і матеріалів, що досліджуються, які несуть інформацію про конструктивні та інші особливості технічних засобів запису. Досить часто в ролі подібних ознак постають характеристики носія запису, обумовлені геометричними параметрами записуючого пристрою, наприклад, формою й розташуванням записуючої магнітної голівки, відстанню між стираючою та записуючою голівками і таке інше. Наступним, не менш важливим джерелом інформації про засіб запису, є сам сигнал. Швидкість руху стрічки, коефіцієнт детонації, власні шуми механічних та електронних підсистем, характерні викривлення сигналу та інші параметри засобів запису знаходять своє відображення в структурі записанного сигналу і можуть бути використані при вирішенні завдань ідентифікації. Треба відзначити, що існуюча технологічна база дозволяє проводити вимірювання та аналіз ознак другої групи значно детальніше, ніж першої. Саме тому зусилля авторів були сконцентровані, в основному, навколо проблем, пов'язаних з виділенням ідентифікуючих ознак безпосередньо з сигналу, що досліджується. Подана робота розглядає проблему ідентифікації звукозаписуючого пристрою, застосовуючи її до фонограм, виготовлених за допомогою малогабаритних диктофонів.

ІДЕНТИФІКАЦІЯ ДИКТОФОНІВ

У фоноскопічній практиці досить часто в ролі об'єкта досліджень фігурують фонограми, виконані на мікрокасетах з використанням диктофонів. При цьому необхідно підтвердити або заперечити припущення про те, що фонограма була виготовлена за допомогою поданого експертам диктофону. Більшість сучасних диктофонів з метою економії магнітної стрічки обладнані системою активації запису за вхідним сигналом. Принцип роботи цієї системи простий: якщо під час запису протягом певного часу рівень вхідного сигналу не перевищує певного порогового значення, двигун диктофону вимикається аж до появи вхідного сигналу достатнього рівня, після чого двигун знову вмикається, і запис фонограми поновлюється. Проведені дослідження показали, що часові параметри розгону стрічки при спрацьовуванні системи мають досить стабільний характер, практично повністю визначаються характеристиками самої системи і досить слабко залежать від носія запису, що використовується. Таким чином, існують передумови щодо використання інформації про часові характеристики спрацьовування системи активації в ролі ідентифікуючої ознаки.

Опишемо детальніше, яким чином спрацьовування системи позначається на структурі сигналу, який записується. Розглянемо ідеалізовану модель стрічкопротягуючого механізму, яка складається з нерухомої голівки читання/запису H і носія запису - стрічки T, яка переміщується відносно голівки. Будемо й далі вважати процес запису/відтворювання лінійним, тобто стверджувати, що сигнал, який прочитано з частини носія запису, є пропорційним записанному на цій частині сигналу, причому коефіцієнт пропорційності не залежить від характеру переміщення стрічки відносно голівки. Це досить не точне припущення, виправдане, відверто кажучи, тільки для систем з магніточутливими голівками. Однак, на данному етапі воно може бути застосовано. Отже, в початковий момент часу стрічка нерухома, голівка розташована над її початком, і система активується гармонійним сигналом $S(t)$ з постійною частотою ω_0, тобто

$$S(t) = ACos(\omega_0 t + \varphi_0).$$

Після спрацьовування системи активації стрічка починає рухатися, причому відстань між поточною позицією записуючої голівки та її початковою позицією на стрічці змінюється за законом $l(t)$. Конкретний вигляд функції $l(t)$ визначається параметрами пристрою, апріорі про неї можна сказати наступне:

1. $l(0) = 0$.
2. $l(t)$ монотонно зростає на всій області визначення.

3. $l(t)$ асимптотично наближається до лінійної функції $y(t)=y_0+\tilde{\vartheta}t$ зі збільшенням t, де $\tilde{\vartheta}$ – установлена швидкість руху стрічки під час запису, y_0 – деяка константа.

4. Існує зворотна до $l(t)$ функція, позначим її $g(x)=l^{-1}(x)$.

Нехтуючи коефіцієнтом детонації, будемо вважати, що в процесі відтворення голівка, що зчитує, рухається з постійною швидкістю ϑ, при цьому відстань $l_2(\tau)$ між позицією голівки і початком стрічки змінюється за законом $l_2(\tau)=\vartheta\tau$, де τ – час, який минув від початку відтворення запису. Беручи до уваги припущення стосовно лінійності запису/відтворення, можна записати, що відтворений сигнал дорівнює

$$\tilde{S}(\tau)=kACos(\omega_0 g(\vartheta\tau)+\varphi_0).$$

Повна фаза такого сигналу дорівнює $\varphi(\tau)=\omega_0 g(\vartheta\tau)+\varphi_0$, а його миттєва частота

$$\omega(\tau)=\frac{d\varphi(\tau)}{d\tau}=\omega_0 g'(\vartheta\tau)\vartheta.$$

Застосовуючи теорему про похідну зворотної функції, маємо

$$\omega(\tau)=\frac{\omega_0\vartheta}{l'(g(\vartheta\tau))}.$$

Таким чином, похідний сигнал, записаний під час встановлення швидкості руху стрічки при спрацюванні системи активації, підлягає частотній модуляції згідно з наведеним виразом. Знаменник наведеного виразу повністю визначається динамічними характеристиками стрічки, що розгоняється, які залежать від конструкції системи активації, тому, при відомій швидкості зчитування сигналу, в ролі динамічної характеристики системи активації може поставати наступна функція

$$\Omega(\tau)=\frac{\omega(\tau)}{\omega_0}=\frac{\vartheta}{l'(g(\vartheta\tau))}.$$

Методика її експериментального визначення, згідно з вищенаведеним, постає у наступному:

1. Диктофон, увімкнутий в режимі запису і зупинений системою активації при відсутності вхідного сигналу, активується звуковим сигналом повної частоти (наприклад, 1 КГц).
2. Записується частина сигналу певної тривалості (2 – 3 сек.).
3. Стрічка перемотується назад, і отримана фонограма відтворюється з постійною швидкістю. Вихідний сигнал диктофону оцифровується для подальшого дослідження.
4. Визначається миттєва частота введеного сигналу, і формується динамічна характеристика системи $\Omega(\tau)$.
5. Якщо необхідно, експеримент можна повторити.

Найбільш складним етапом експерименту є етап визначення миттєвої частоти сигналу. Тут треба відзначити, що в реальних системах сигнал підлягає різним нелінійним викривленням, як на етапі запису, так і на етапі відтворення, що призводить до появи у спектрі сигналу вищих гармонік записаної частоти. Типовий приклад спектрограми подібного сигналу наведено на рис.1. На ньому виразно видно зону спрацьовування системи активації та зону усталеного руху стрічки. У тих випадках, коли рівень вищих гармонік досить невеликий, для визначення динамічної характеристики системи з успіхом може бути використаний розроблений авторами цифровий частотний детектор, детально описаний в [1].

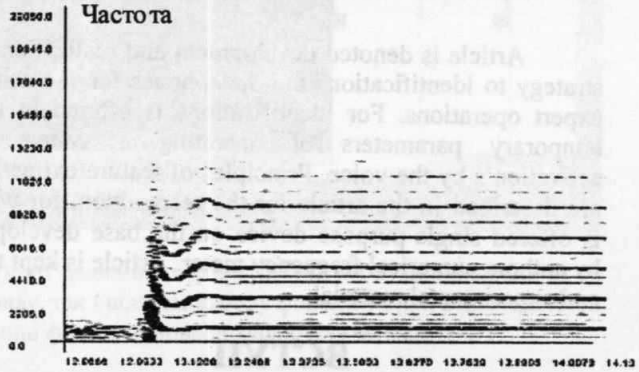


Рис.1 Спектрограма сигналу, записаного під час спрацьовування системи активації.

У тих випадках, коли рівень вищих гармонік та/або шумів занадто високий, найбільш доцільно застосовувати методику визначення динамічної характеристики системи, яка базується на обчисленні миттєвих спектрів сигналу. Нижче наведено спрощений опис методики:

1. У зоні усталеного руху стрічки береться відрізок тривалістю біля сотих часток секунди.
2. На обраному відрізці обчислюється миттєвий спектр сигналу. Для зменшення небажаних ефектів, пов'язаних з малою тривалістю відрізку, при обчислюванні спектру доцільно використовувати зважування відрізку підходящим цифровим вікном, наприклад, вікном фон Ханна або Блекмена.
3. Визначається місцезнаходження на осі частот локальних максимумів спектру потужності, відповідних до гармонік базової частоти.
4. Далі обробка відбувається у циклі, на кожній ітерації якого робочий відрізок зсовується з певним кроком (біля кількох мілісекунд) до початку сигналу. Цикл завершується, коли середній рівень сигналу в робочому відрізку стає нижче за порогове значення, тобто робочий відрізок виходить за межі зони спрацьовування системи активації.
5. Інакше знову обчислюється миттєвий спектр потужності робочого відрізку сигналу і визначається місцезнаходження локальних максимумів. Відбувається побудова траєкторій гармонік базової частоти, при якій продовженням траєкторії для кожної гармоніки вважається позиція на осі частот локального максимуму, найближчого до вже побудованого на попередній ітерації циклу.
6. Цикл повторюється.
7. Після закінчення циклу для кожної траєкторії обчислюється співвідношення у вигляді

$\Omega_i(n)=\frac{\omega_i(n)}{\omega_{i0}}$, де n – індекс ітерації, $\omega_i(n)$ – траєкторія i-ї гармоніки, отримана на n-й ітерації, ω_{i0} – значення траєкторії, обчислене на кроці 3.

8. На підставі отриманих даних обчислюється динамічна характеристика системи активації за формулою $\Omega(n)=\frac{1}{N}\sum_{i=1}^{N}\Omega_i(n)$, де N – кількість отриманих траєкторій.

Нижче, на рис. 2, подано динамічну характеристку системи активації диктофону фірми SONY, отриману під час обробки визначеним способом сигналу, спектрограма якого була наведена вище.

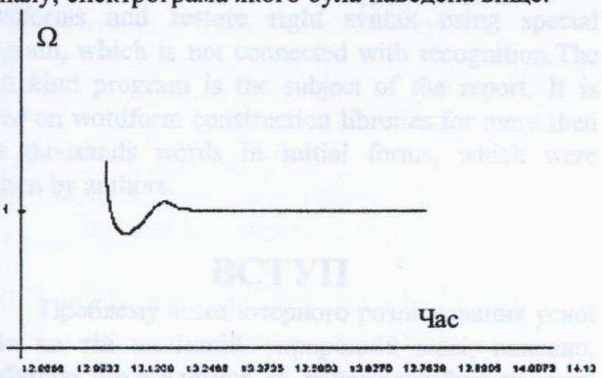


Рис. 2. Динамічна характеристика системи активації диктофону.

Таким чином, може бути сформована множинність, що навчає систему розпізнавати образи, які можна використати під час ідентифікації диктофонів. Наступним завданням є визначення динамічної характеристики системи активації диктофону, за допомогою якого була виготовлена фонограма, що досліджується. Для цього можна скористатися відомим з теорії мовоутворення [4] фактом, згідно якому спектри голосних звуків мають лінійчату структуру. Приклад спектрограми голосного звуку "О" наведено на рис. 3.

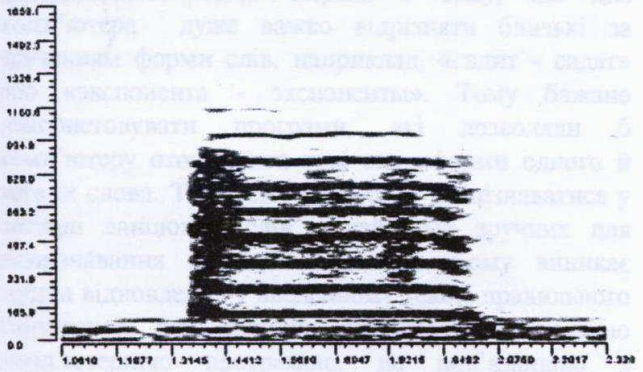


Рис.3. Спектрограма звуку «О»

Очевидно, що при активації диктофону голосним звуком, описаному вище викривленню підлягають усі його спектральні компоненти; окрім того, час спрацьовування системи активації замалий порівняно з часом інтонаційних змін основного тону мовлення. Тому для визначення динамічної характеристики невідомого диктофона достатньо застосувати описану вище методику з тією лишь різницею, що дослідженню підлягає відрізок фонограми, виконаний при активації диктофона голосним звуком або коротким приголосним з наступним голосним. Як приклад, на рис. 4 подано спектограму сигналу, отриманого при активації диктофону, який фігурував раніше, словом "ДА", а на рис. 5 - динамічна характеристика системи активації, отримана на основі цього сигналу.

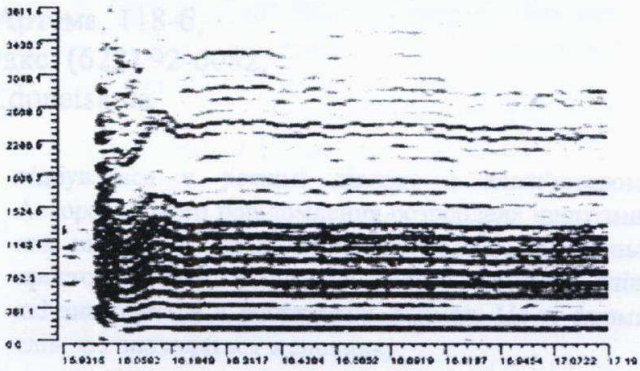


Рис. 4. Спектрограма сигналу активації

З цього видно, що отримані різними способами характеристики практично співпадають.

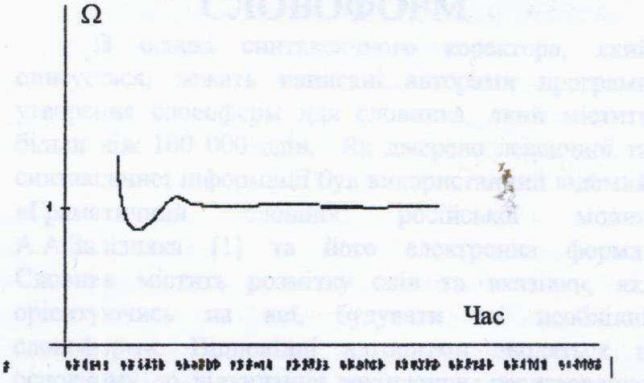


Рис. 5. Динамічна характеристика диктофону.

Отримані таким чином характеристики з успіхом можуть бути використані під час розв'язання завдання ідентифікації засобів запису. Вибір класифікатора, що використовується, визначається постановкою завдання ідентифікації у кожному конкретному випадку.

ВИСНОВКИ

У роботі запропоновано підхід до розв'язання проблеми ідентифікації диктофонів; обгрунтовано і розглянуто метод ідентифікації за динамічною характеристикою системи активації запису; стисло описано використані при цьому методики та алгоритми. Розглянуто приклади.

ЛІТЕРАТУРА

1. Старушко Д.Г., Шевченко А.И. Цифровой частотный детектор. Донецк, ДонГИИИ, *«Искусственный Интеллект»* №1, 1998.

2. Старушко Д.Г., *«Быстрый алгоритм дискретного преобразования Хартли»*. Ялта, Материалы конференции KDS – 97, 1997.

3. Н. Ахмед, К. Рао, *«Ортогональные преобразования при обработке цифровых сигналов»*. М.: Связь, 1980.

4. Фант Г. *«Акустическая теория речеобразования»*. М.: Наука, 1964.

ЩОДО ПРОГРАМИ СИНТАКСИЧНОГО КОРЕКТОРА

В.Ю. Шелепов, А.Н. Лимарь, А.В. Шевченко, Г.В. Саввіна,
Э.Н. Селищев, В.А. Грабовая
Інститут проблем штучного інтелекту
340048, Донецьк, Артема, 118-б,
Тел.: (0622) 92-6082, Факс: (622) 92-6082,
E-Mail:shel@iai.donetsk.ua

Computer russian speech recognition is complicated by great number of wordforms. It is appropriate the computer should identify close wordforms and restore right syntax using special program, which is not connected with recognition.The such kind program is the subject of the report. It is based on wordform construction libreries for more then 100 thousands words in initial forms, which were written by authors.

ВСТУП

Проблему комп'ютерного розпізнавання усної мови на тій чи іншій природній мові, напевно, необхідно роздивлятися як найважливійшу частину більш загальної задачі навчання комп'ютера відповідній мові. При цьому ситуація до деякої міри нагадує навчання людини іноземній мові. Але для викладача очевидно, що знання мови - це не тільки знання або розуміння слів, але й володіння мовними структурами. Серед останніх одне з найважливіших місць займають синтаксичні структури.

Російська мова, якою ми займаємося, відноситься до флективних мов й відрізняється великою кількістю форм слів. Тому для неї проблеми синтаксису пов'язані з задачею розпізнавання мови. Справа в тому, що для комп'ютера дуже важко відрізняти близькі за звучанням форми слів, наприклад, «садит - садят» або «экспонента - экспоненты». Тому бажано використовувати програми, які дозволяли б комп'ютеру ототожнити такі словоформи одного й того ж слова. Тоді усна мова буде розпізнаватися у вигляді ланцюжка слів у найбільш зручних для розпізнавання формах. Але при цьому виникає задача відновлення у введенному тексті правильного синтаксису, яка повинна вирішуватися окремою комп'ютерною програмою, не пов'язанною з розпізнаванням - синтаксичним коректором. Такий коректор буде принципово видрізнятися від програм, які використовуються в автоматичних перекладачах, тому що тут відсутній первісний синтаксично правильний текст, й відновлення синтаксису необхідно здійснювати практично з нуля. Очевидно, що рішення цієї задачі у повній мірі неможливо без враховування семантики і, більш того, без участі людини. Ця участь в ряді моментів повинна відбуватися у режимі діалогу з комп'ютером. Авторами цього повідомлення розроблена програма, яка відновлює правильний синтаксис достатньо простого речення російської мови в ланцюжці слів, які знаходяться в початкових формах. Це ті форми слів, які знаходяться в словнику.

1. ПРО УТВОРЕННЯ СЛОВОФОРМ

В основі синтаксичного коректора, який описується, лежать написані авторами програми утворення словоформ для словника, який містить більш ніж 100 000 слів. Як джерело лексичної та синтаксичної інформації був використанний відомий «Граматичний словник російської мови» А.А.Зализняка [1] та його електронна форма. Словник містить розмітку слів та вказівки, як, орієнтуючись на неї, будувати всі необхідні словоформи. Відповідні алгоритми зводяться в основному до відкидання закінчення, перетворення основи при наявності біглої голосної й додавання нового закінчення. Іноді основа повинна повністю замінюватися іншою (наприклад, «идти» - «шел»).

Трудність полягає у великій кількості різновидів конкретних реалізацій цих алгоритмів. Так, для відмінювання іменників використовуються близько двохсот різновидів алгоритму утворення п'яти непрямих відмінків однини й шести відмінків множини. Розмітка значної частини словника була нами змінена з причин, про які сказано нижче.

Метод, який застосовується при утворюванні словоформ, розглянемо на прикладі іменників. Іменники російської мови, що містяться в граматичному словникові А.А.Зализняка, були розподілені на наступні групи:

- Група А. Іменники, при утворенні форм яких їх основа не змінюється, а змінюється лише закінчення в кінці слова;
- Група В. Іменники, при утворенні форм яких основа змінюється;
- Група С. Іменники цієї групи позначають або людей за національною, географічною та соціальною належністю (наприклад, «южанин») або малят (наприклад, «бельчонок»)

• Група D. Іменники цієї групи мають нестандартне закінчення «-а» в називному відмінку множини (наприклад, «рукава»), або нестандартне «нульове» закінчення в родовому відмінку множини (наприклад, «грузин»).

Для кожної з вищеозначених груп була складена структура, яка містить всі можливі закінчення, які виникають при утворенні відмінкових форм в однині та множині і написана функція підстановки цих закінчень.

Всі іменники зі складу словника А.А. Зализняка були перерозмічені, тому що розмітка автора словника не підходила для роботи з ряду причин.

Приклад 1. (приклад нової розмітки)

абажур mAa0_0

З цієї розмітки бачимо, що:

• слово «Абажур» має чоловічий рід. На це вказує літера 'm';

• слово «Абажур» - неістота. Це видно з того, що літера 'm' маленька. Якби слово вказувало на істоту, розмітка починалася б з великої літери;

• Закінчення додається в кінець слова без знищення останньої літери слова. На це вказує наявність в розмітці літери 'a'. В противному випадку в розмітці замість литери 'a' стояла б літера 'b'. (Наприклад, слово «портфель mAb2_1»);

• При підстановці закінчень однини використовується структура за номером '0', при підстановці закінчень множини - також структура за номером '0';

Як можна бачити з описання нової розмітки, на неї покладене практично все інформаційне навантаження про утворення форм слів, які відрізняються від початкових, і, отже, нова розмітка більш громіздка порівняно з розміткою А.А.Зализняка. І хоча, казалося б, словник, який містить слова з новою розміткою, повинен займати більше місця на диску, ніж вихідний, його розмір приблизно дорівнює розміру вихідного словника, тому що з останнього було викинуто безліч службової інформації, на яку в багатьох випадках спирається використання словника А.А.Зализняка.

Крім того, робота з новою розміткою дозволяє отримати значний виграш у часі на стадії утворення словоформ. Це можна продемонструвати на наступному прикладі.

Приклад 2. Розглянемо наступні два слова: слово «Площадь» и слово «Мышь». В словнику А.А.Зализняка вони мають однакову розмітку - «8е». Але як наслідок того, що основа другого слова закінчується на шиплячий, а основа першого - ні, в утворенні форм «площадям - мышам», «площадями - мышами» є різниця. Таким чином, утворюючи словоформи за розміткою А.А.Зализняка й зустрівши одне з вищезгаданих (або аналогічних) слів з розміткою «8е», ми змушені робити перевірку на наявність шиплячої в основі слова й в залежності від результату перевірки підставляти необхідне закінчення. Ці дії неможливо реалізувати одним або декількома операторами алгоритмічного язика. Крім того, оператори перевірки й порівняння, особливо строкових констант, займають одне з перших місць по захвату процесорного часу.

Ми розглянули випадок, коли тільки дві дещо відмінні групи слів віднесені до одного класу. В вихідному словнику зустрічаються випадки, коли до одного класу віднесені 4-5 груп іменників, які різняться між собою (див. [1], с. 48). В зв'язку з цим кількість перевірок на стадії утворення словоформ збільшується. Нова розмітка, як нам здається, вільна від подібних дефектів.

Утворення словоформ для інших самостійних частин мови здійснюється за аналогічними алгоритмами. Програми утворення словоформ оформлені у вигляді динамічних бібліотек (DLL), які можна підключати в готовому вигляді до проектуємих граматичних програм і які представляють, таким чином, самостійний програмний продукт.

2. ПРО ПРОГРАМУ КОРЕКТОРА

У нас вони використовуються в програмі «Синтаксичний коректор», робота якої базується на використанні апарату дерев синтаксичного підпорядкування (див. [2]). Після того, як ланцюжок початкових форм введено у вікно редактора, комп'ютер через діалогове вікно питає про спосіб і час, в які необхідно поставити дієслово, яке виражає присудок. Те ж саме він робить відносно числа, в яке необхідно поставити підмет (відповіді - натискування відповідних кнопок за допомогою миші). Після цього комп'ютер автоматично узгоджує підмет і присудок. У випадку наявності додатку він ставить іменник у відмінку, який визначається прийменником. Нарешті, він узгоджує наявні прикметники й числівники з тими іменниками, до яких вони відносяться, й ставить у потрібній формі займенники.

ВИСНОВКИ

В результаті ми отримуємо можливість обробити ланцюжок слів типу: "Усталый конь медленно пробирается сквозь высокий трава", перетворивши її в речення: «Усталые кони медленно пробирались сквозь высокую траву».

ЛІТЕРАТУРА

1. Зализняк А.А. Грамматический словарь русского языка. Москва: "Русский язык", 1977, 880с.

2. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. Москва: Наука, 1985, 144с.

Speech Recognition Algorithm Implemented with a DSP

Authors: ***Constantin FILOTE***, e-mail: filote@eed.usv.ro
Adrian GRAUR, e-mail: adriang@eed.usv.ro
Gabriel ANTONESEI
Ovidiu OBADĂ
tel:(40) 30.216297, fax: (40) 30.520277,
Faculty of Electrical Engineering, *"Ştefan cel Mare"* University of Suceava,
1 University Street, 5800, Suceava, România.

Abstract: The aim of this paper is to describe a speech recognition algorithm. We describe the structure of a speaker-dependent system for sounds recognition. The implementation of this algorithm with a DSP allow to design and construct a vehicle that can be fully controlled from human generated sounds. More exactly we can command to our car to go forward, to left or right and to stop.

Keywords: Speech recognition , DFT algorithm, DSP.

1. INTRODUCTION

Most application of today involve the use of discrete time technological of processing continuous-time signals. One important class of signal processing problems is signal interpretation. The objective of the processing is not to obtain an output signals but to obtain a characterization of the input signal.

In a speech recognition system, the objective is interpret the input signal or extract information from it. Typically, such a system will apply preprocessing (filtering, parameter to estimation, Fourier transform etc.) followed by a pattern recognition system.

Speech recognition research and development has several goals. Simplifying the interface between users and machine is one major goal. Just as many users consider the mouse an improvement to the user interface on a personal computer, machine speech recognition and understanding has the potential to gre atly simplify the way people works with machines. Examples of this emerging technology include dialing telephones and controlling consumer electronic through voice-activation. As voice input and output become further integrated into the everyday machines, many advances will be possible.

Speech recognition systems fall into two categories:

- *Speaker dependent systems* that are used (and often trained) by one person (our car);
- *Speaker independent systems* that can be used by anyone.

2. VOICE PRODUCTION & MODELING

You can separate human speech production into two distinct sections: sound production and sound shaping. Sound production is caused by air passing across the vocal chords (as in "a", "e " or "o") or from a constriction in the vocal tract (as in "sss", "p" or "sh"). Sound production using the vocal chords is called voiced speech; unvoiced speech is produced by the tongue, lips, teeth, and mouth. In signal processing terminology, sound production is called excitation.

Sound shaping is a combination of the vocal tract, the placement of the tongue, lips, teeth, and the nasal passage. For each fundamental sound, or phoneme, the shape of the vocal tract is somewhat different, leading to a different sound. In signal processing terminology, sound shaping is called filtering.

3. ALGORITHM DESCRIPTION

The theory behind speech recognition is relatively simple. First, the DSP acquires an input sound (word) and compares it to a library of stored sounds. Then the DSP selects the library sound that most closely matches the unknown input sound. The selected sound is the recognition result. Systems that follow this model have two distinct phases: training phase and recognition phase.

3.1 Training phase

When you train a system to recognize sounds, you first create a library of stored sounds. Each sound to be recognized is stored in a library. Once the library is built, the system training is complete, and the task of recognition can begin.

3.2 Recognition phase

Figure 1 show a block diagram of the recognition algorithm:

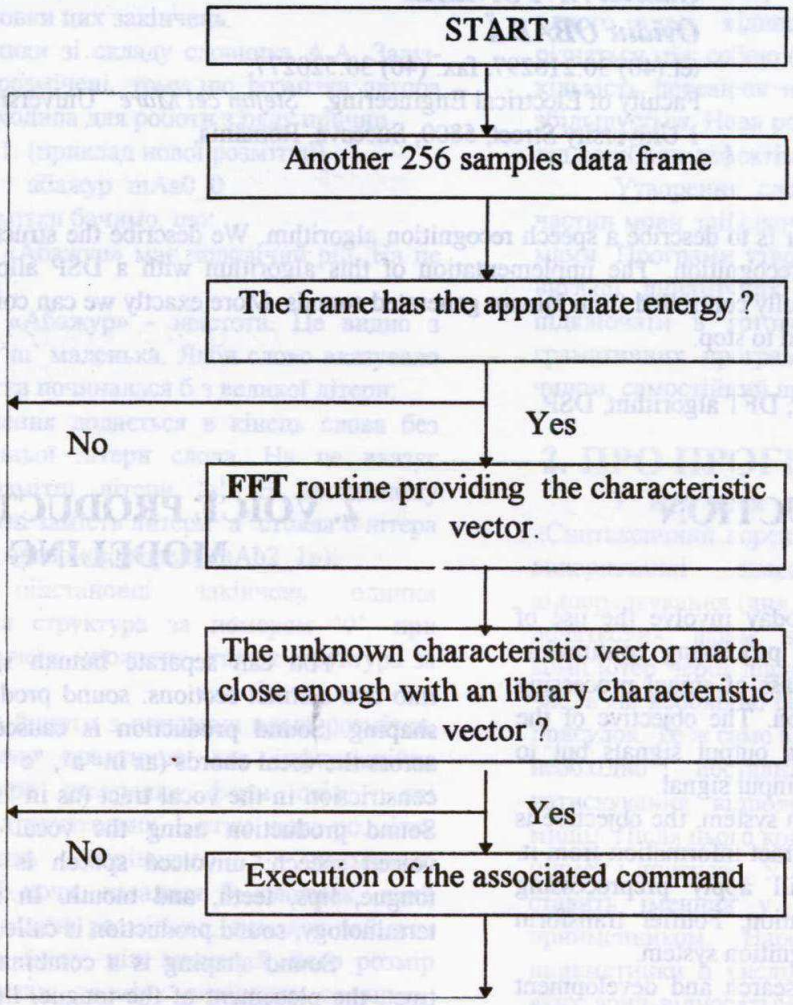


Figure 1 Speech Recognition Algorithm Diagram.

For the instant we resume to recognize fundamentals sounds such are the voices. We must mention that the same voice, but pronounced in two different words can be in her tour different. So, for the beginning we can recognize words by founding different voices.

The analogue input signal provided by a microphone is converted into a digital one as a result of hold and sampling process at 8 kHz frequency.

The processor receive this signal, split him in frames of 256 data (32 ms) and process this frames.

Because the microphone provide signals continuously and this signals are not all the time human sounds (like the noise, the smashes, etc.), it will be inefficient to process each frame. That's the reason why we make a test of this frames, more exactly we check if the energy of frame is high enough. If she is we go further, if not we take another frame.

Almost all speech recognition process use a parametric representation of speech rather than the waveform itself as the basis for pattern recognition. These parameters usually carry the information about the short time spectrum of the signal.

Among the most popular representation, produced by various forms of signal analysis, are spectral coefficients (**DFTC**), cepstral coefficients (**CEPC**), and linear predictive coding coefficients (**LPC**):

- Fourier analysis (**DFT**) yields discrete frequencies over time:

$$S(f) = \sum_{n=0}^{Ns-1} s(n) e^{-j2\pi n \frac{f}{f_s}} \quad (1)$$

where f_s is the sampling frequency and N_s is the length of the analysis sequence;

- linear predictive coding (**LPC**) yields coefficients of a linear equation that approximate the recent history of the raw speech values;

- cepstral analysis [1] calculates the inverse Fourier transform of the logarithm of the power spectrum of the signal:

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left|\hat{X}(e^{j\omega})\right| e^{j\omega n} d\omega \quad (2)$$

$$c(n) = \frac{1}{N_S} \sum_{k=0}^{Ns-1} \log|S(k)| e^{\frac{2\pi}{Ns} kn} \quad (3)$$

A signal observed for a finite interval of time (window) may have distorted spectral information in the Fourier transform. To avoid or minimize distortion, a signal is multiplied by a window-weighting function before the DFT is performed. Window choice is crucial for separation of spectral components. Our application use a 32 ms Hamming window to weight samples toward the center of the window.

The coefficients for the Hamming window [3] are obtained from the formula:

$$w(n) = \alpha + (1-\alpha)\cos\left(2\pi \frac{n}{N}\right) \quad (5)$$

commonly, $\alpha = 0.54$.

N = number of coefficients

Range: $$n = -\frac{N}{2} \text{ to } n = \frac{N}{2} - 1 \quad (6)$$

The DFT of Hamming Window Function is:

$$W(\theta) = \alpha D(\theta) + \frac{1}{2}(1-\alpha)\left[D(\theta - \frac{2\pi}{N}) + D(\theta + \frac{2\pi}{N})\right] \quad (7)$$

where,

$$\theta = 2\pi \frac{k}{N} \quad (8)$$

and

$$D(\theta) = exp\left(\frac{j\theta}{2}\right) \frac{sin\left(N\frac{\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right)} \quad (9)$$

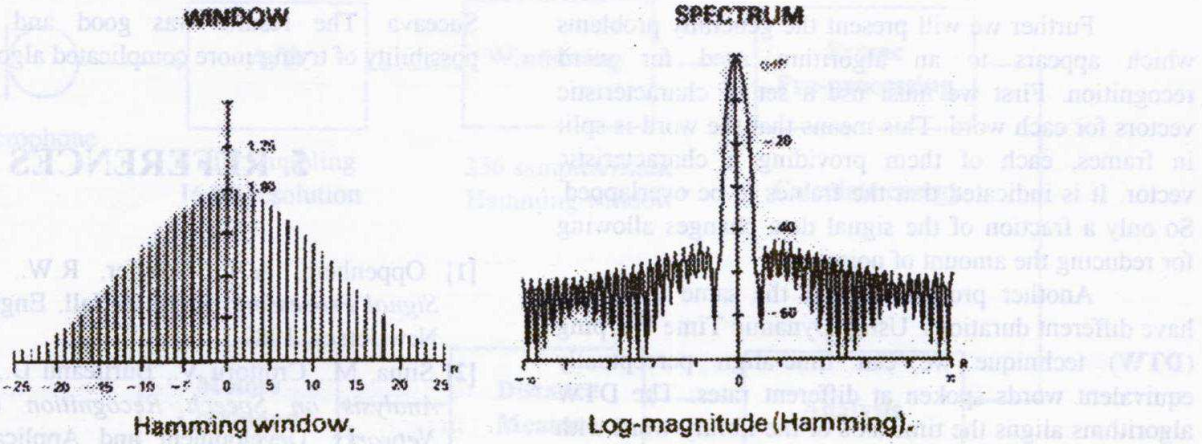


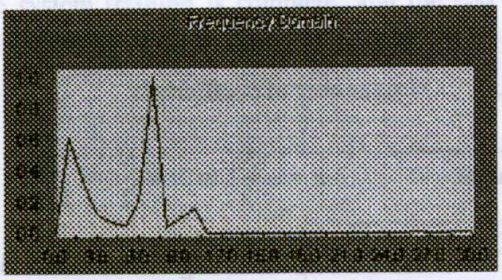
Figure 2 Characteristic of the Hamming Window.

Then we perform a 256 points FFT routine (Fast Fourier Transform). This routine provide a 128 points data vector containing information about the frequencies of the frame (in fact it contains the coefficients of the FFT). We do some operations with this vector in order to remove unnecessary information. We will name it the characteristic vector.

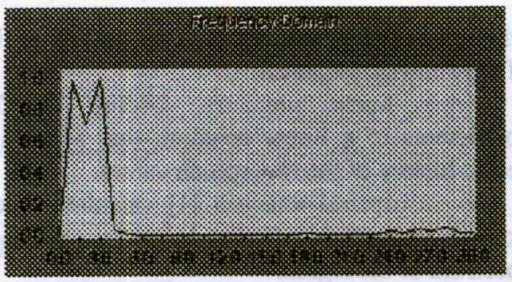
We compare this vector with the library created in the training phase. The distance between vectors can be determinate in many ways: the absolute distance, the euclidian distance, etc. I used the absolute distance.

Then the DSP selects the library vector that most closely matches the unknown vector and perform the associated command.

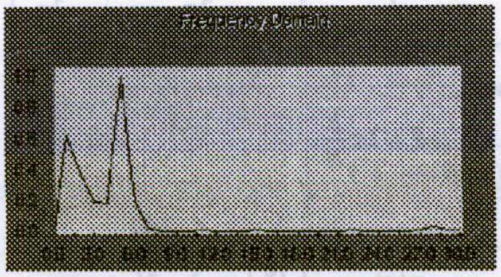
In figure 3 you can see the characteristic vectors of the voices that we use in our application.



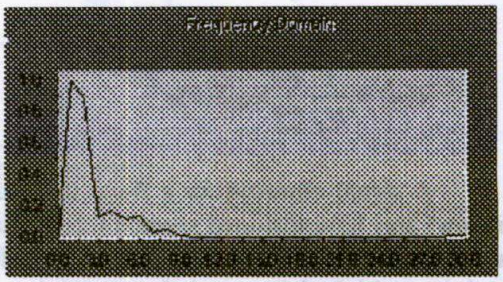
Characteristic vector of sound "a" from word UP



Characteristic vector of sound "i" form word PICK



Characteristic vector of sound "e" from word LEFT



Characteristic vector of sound "o" from word STOP

Figure 3 DFT caracteristic of some sounds.

Further we will present the generally problems which appears to an algorithm used for word recognition. First we must use a set of characteristic vectors for each word. This means that the word is split in frames, each of them providing a characteristic vector. It is indicated that the frames to be overlapped. So only a fraction of the signal data changes allowing for reducing the amount of noise.

Another problem is that the same word can have different durations. Using Dynamic Time Warping (**DTW**) technique we can time-align perceptually equivalent words spoken at different rates. The DTW algorithms aligns the time axis of the library word with the time axis of the unknown word.

4. CONCLUSIONS

The presented speech recognition algorithm a was implemented on an fully controlled by human sounds vehicle realized in the Digital Signal Processing Laboratory of Electrical Engineering Department from Suceava. The results was good and give us the possibility of trying more complicated algorithms.

5. REFERENCES

[1] Oppenhein, A.V., Schafer, R.W., *Discrete-time Signal Processing,* Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[2] Sima, M., Croitoru V., Burileanu D., *Performance Analysis on Speech Recognition using Neural Networks,* Development and Application Systems (D&AS '98), Suceava, may 21-23, 1998, pp. 259-266.

[3] Higgins, R. J, *Digital Signal Processing in VLSI,* Prentice-Hall, Enlewood Cliffs, New Jersey, 1990.

[4] *Digital Signal Processing Applications using the ADSP-2100 Family*, Prentice-Hall, Inc. A division of Simon & Scuster Englewood Cliffs, New-Jersey 07632, 1992.

Voice Controlled Vehicle

Authors: ***Adrian GRAUR***, e-mail: adriang@eed.usv.ro
Constantin FILOTE, e-mail: filote@eed.usv.ro
Ovidiu OBADĂ
Gabriel ANTONESEI
tel:(40) 30.216297, fax: (40) 30.520277,
Faculty of Electrical Engineering, *"Ştefan cel Mare"* University of Suceava,
1 University Street, RO-5800, Suceava, România.

Abstract: The aim of this paper is to describe a speech recognition hardware. The implementation of this algorithm with an DSP allow to design and construct a vehicle that can be fully controlled from human generated sounds.

Keywords: Speech recognition, DSP, voice controlled vehicle.

1. INTRODUCTION

The results obtain in the speech recognition had leading to application in drive system. Because the speech recognition domain is very interesting we have decided to make a project with this theme. Our application consist in fully controlled by human sounds.

The most suitable for this application is DSP. The vehicle can receive commands to go forward, to left or right and to stop.

Limited by the memory size and the power of computing we decided to use the simple sound speech recognition algorithm (see block diagram in figure 1).

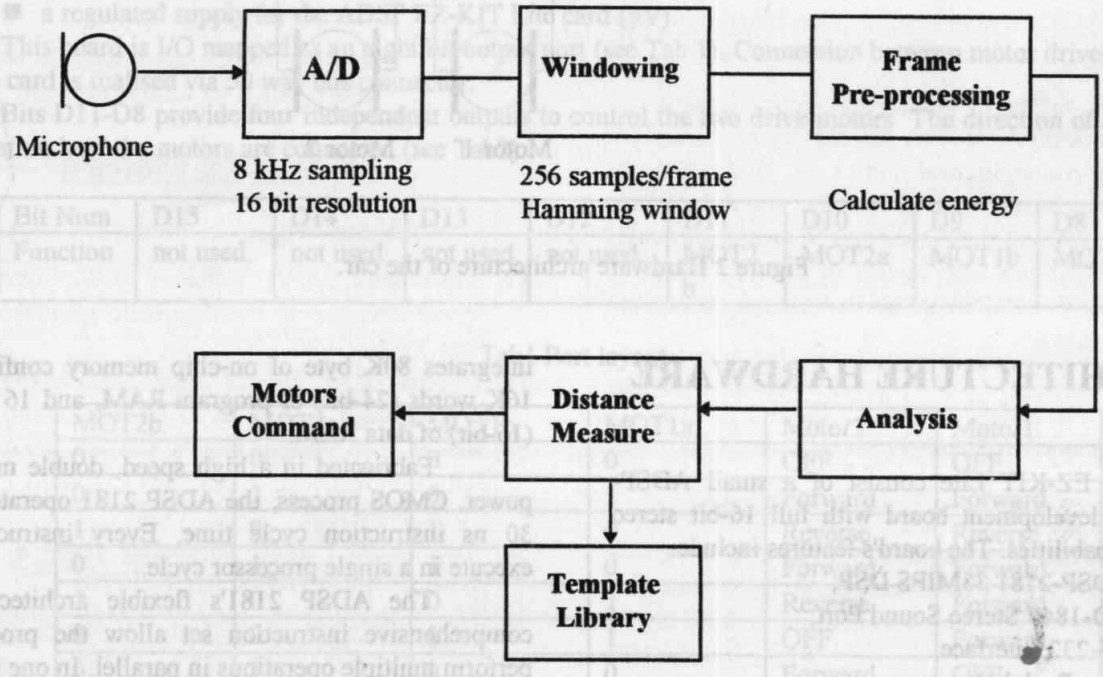


Figure 1 Block diagram of speech recognition algorithm.

2. VEHICLE HARDWARE

Overview: The vehicle will comprise of a pair geared dc motors to provide propulsion and steering, a microphone to monitor any sound, a EZ-KIT Lite board to process and analyse these sounds and a motor card to provide adequate power for the motors. Six 1.5 volt AA batteries will provide the power source. The hardware architecture is show in figure 3.

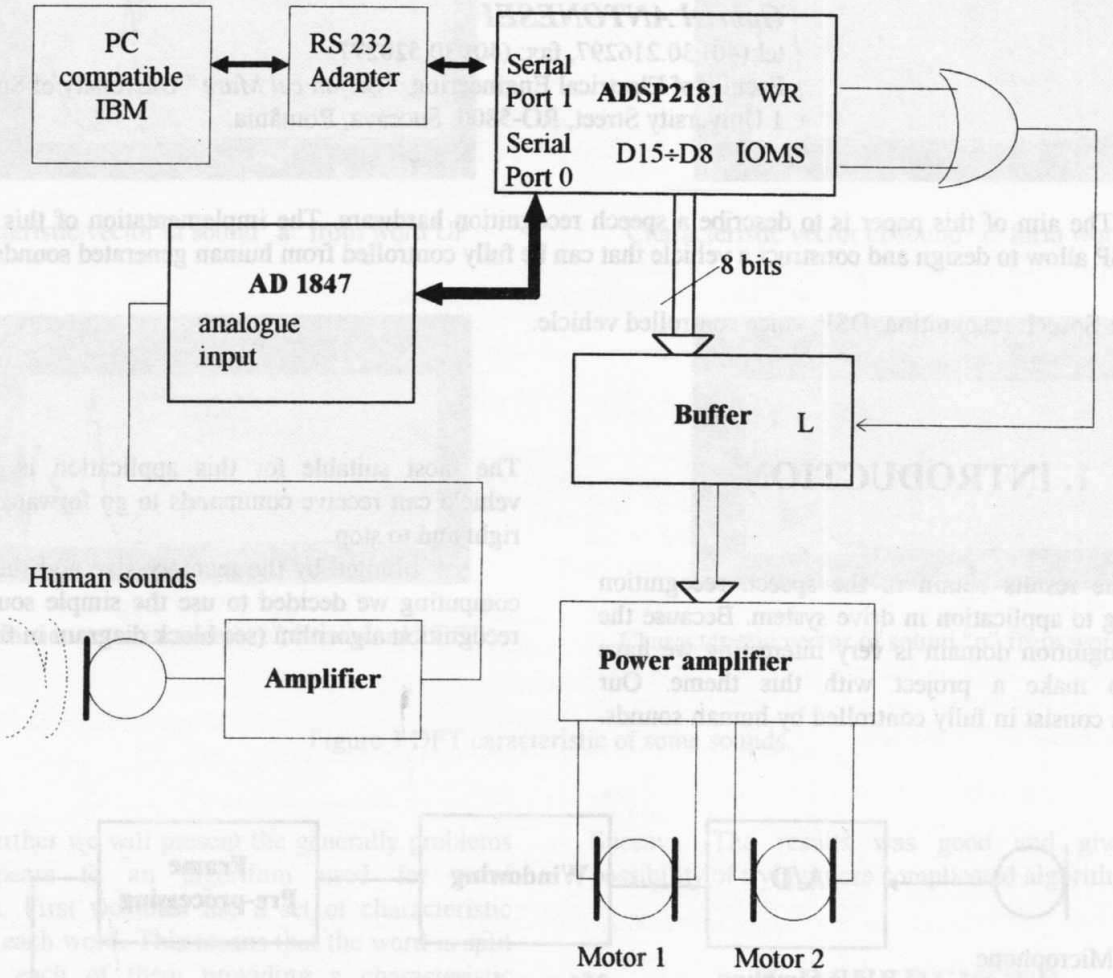


Figure 2 Hardware architecture of the car.

3. ARCHITECTURE HARDWARE

The EZ-KIT Lite consist of a small ADSP-2181 based development board with full 16-bit stereo audio I/O capabilities. The board's features include:

- ADSP-2181 33MIPS DSP;
- AD-1847 Stereo Sound Port;
- RS-232 Interface;
- User Push buttons;
- Expansion Connector.

The ADSP-2181 is a single-chip microcomputer optimised for digital signal processing and other high speed numeric processing applications. It integrates 80K byte of on-chip memory configured as 16K words (24-bit) of program RAM, and 16 K words (16-bit) of data RAM.

Fabricated in a high speed, double metal, low power, CMOS process, the ADSP 2181 operates with a 30 ns instruction cycle time. Every instruction can execute in a single processor cycle.

The ADSP 2181's flexible architecture and comprehensive instruction set allow the processor to perform multiple operations in parallel. In one processor cycle the ADSP 2181 can:

- generate the next program address;
- fetch the next instruction;
- perform one or two data moves;

- update one or two data address pointers;
- perform a computational operation.

The AD1847 integrates key audio data conversion and control function into a single integrated circuit. Dynamic range exceeds 70 dB over the 20 kHz audio band. Sample rates from 5.5 to 48 kHz are supported from external crystal. The sample rates in our program is 8 kHz. We need to build two others board: microphone adapter board and motor drive board. This two boards are used to interface the microphone and the motors with EZ-KIT Lite board.

Microphone board its a simply amplifies the input from the microphone to a suitable level for the codec input of the EZ-KIT Lite board. The schematic of this board is show in figure 4.

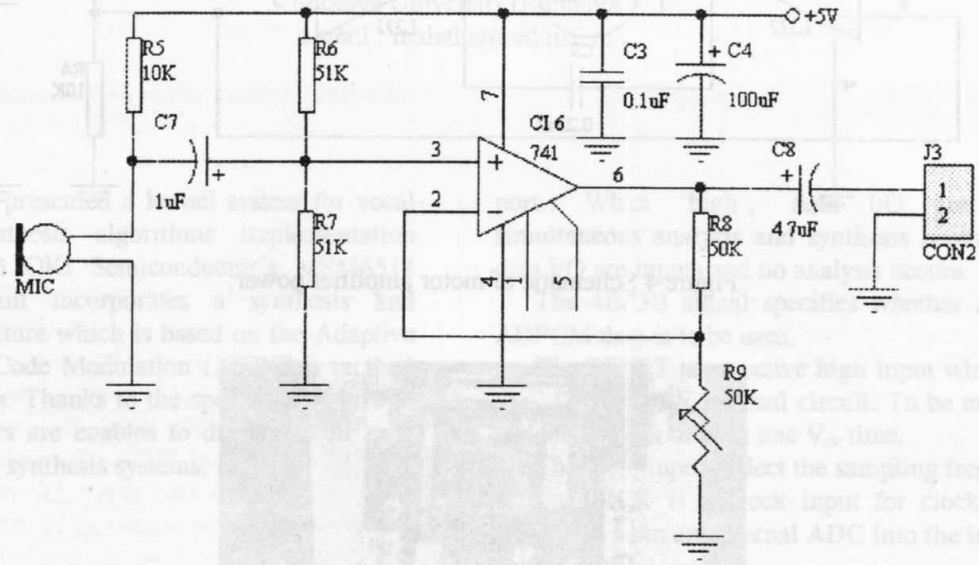


Figure 3 Schematic of microphone ampifier.

The motor drive card is a custom made board for the vehicle project and provides the following facilities:

- four high current outputs (6V @ 500mA) to control the two drive motors;
- a regulated supply for the microphone board (5V);
- a regulated supply for the ADSP EZ-KIT Lite card (9V).

This board is I/O mapped as an eight bit output port (see Tab 1). Connection between motor drive card and EZ-KIT Lite card is realised via 50 way bus connector.

Bits D11-D8 provide four independent outputs to control the two drive motors. The direction of each motor is dependent on how the motors are connected (see Tab2).

Bit Num	D15	D14	D13	D12	D11	D10	D9	D8
Function	not used	not used	not used	not used	MOT2b	MOT2a	MOT1b	MOT1a

Tab1 Port layout.

MOT2b	MOT2a	MOT1b	MOT1a	Motor2	Motor1
0	0	0	0	OFF	OFF
0	1	0	1	Forward	Forward
1	0	1	0	Reverse	Reverse
0	1	1	0	Forward	Forward
1	0	0	1	Reverse	Forward
0	0	0	1	OFF	Forward
0	1	0	0	Forward	OFF
1	1	1	1	Brake	Brake

Tab.2 Pins used for direction of motors.

The command of the motors was implemented with operational amplifier which able to support 1A on its output (L272, SGS Thomson), like in figure 5.

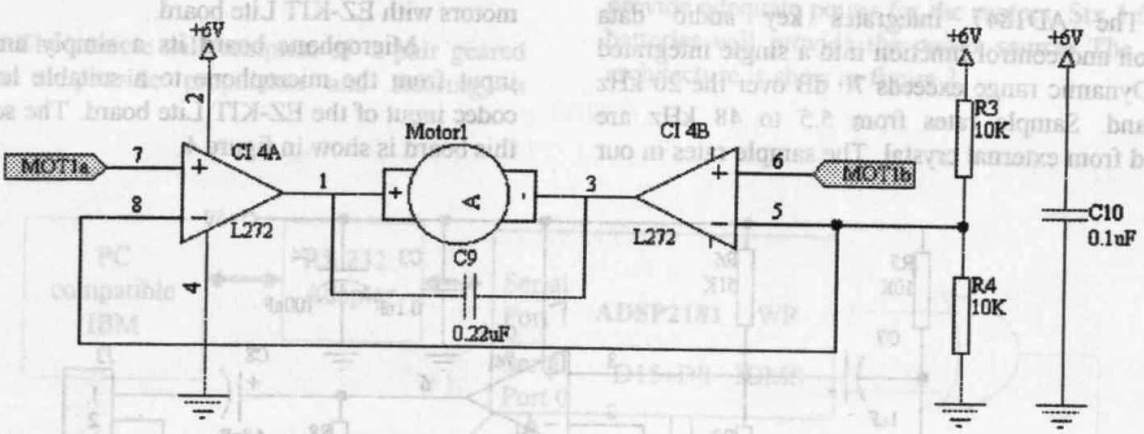


Figure 4 Schematic of motor amplifier power.

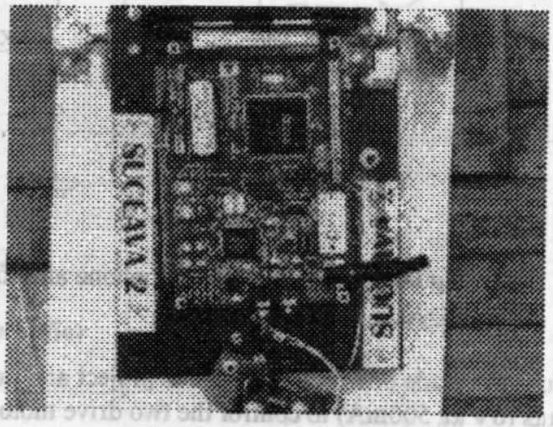


Figure 5 Photography of the car.

The software application was written in assembler language. A very useful feature was the possibility to communicated with a PC through serial port. Also we take advance of the simulator tools provided the EZ-KIT Lite. The object code program is realised with PC and then loaded in the DSP program memory through serial port. The photography of the car is show in figure 6.

4. CONCLUSIONS

The presented speech recognition algorithm and hardware a was implemented on an fully controlled by human sounds vehicle realised in the Digital Signal Processing Laboratory of Electrical Engineering Department from Suceava. The results was good and give us the possibility of trying more complicated algorithms.

5. REFERENCES

[1] Oppenhein, A.V., Schafer, R.W., *Discrete-time Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[2] Higgins, R. J, *Digital Signal Processing in VLSI*, Prentice-Hall, Enlewood Cliffs, New Jersey, 1990.

[3] *ADSP-2100 Family User's Manual*, PTR Prentice-Hall, Inc. A Simon & Schuster Company Englewood Cliffs, New Jersey 07632, 1994.

[4] *ADSP-2100 Family, Assembler Tools & Simulator Manual*, ANALOG DEVICE, 1994.

[5] *ADSP-2100 Family, EZ-KIT Lite Reference Manual*, ANALOG DEVICE, 1995.

[6] SGS-THOMSON Microelectronics.

The Kernel Board for Vocal Signal Analysis/Synthesis Algorithms Implementation

George MAHALU
Suceava University, Romania
e-mail : mahalu@eed.usv.ro

Abstract

In this work is presented a kernel system for vocal signal analysis/synthesis algorithms implementation around of the LSI OKI Semiconductor's MSM6518 circuit. This circuit incorporates a synthesis and analyses stage structure which is based on the Adaptive Differential Pulse Code Modulation (ADPCM) method of data compression. Thanks to the specific structure of the circuit the users are enables to develop their own speech analysis and synthesis systems.

1. INTRODUCTION

In the pattern recognition and/or synthesis speech signals domain these are numerous hardware and software techniques to resolving specific questions. In a lot of cases are necessary very complex electronic equipments and/or sophisticated programme environments. This work become to suggest a more simple method of testing and to put in point for varied algorithms of implementation with a low price.

2. SCHEME DESCRIPTION

In figure 1 is presented the principle scheme of the development kernel system for some specific applications. The structure contain an ADPCM speech analysis/synthesis IC, an AD converter structure with serial data output and a logic control structure to ensure the communication protocol between the MSM5218 IC and other devices. To understand the manner working of the system is necessary to describe the exchange signals on the board.

The V_{ck} is refereed like a signal whose frequency is equal to the sampling frequency selected by S_1 and S_2 inputs.

The D_0 up to the D_3 are four data bits through is exchanged the ADPCM cod between a local processing subsystem and the memory system. When data is rather ADPCM coded through three bits the D_0 is not used.

The ANA/\SYN signal is analysis/synthesis function selector. It controls data flow direction into the I/O port . When high , data I/O are outputs and simultaneous analysis and synthesis occur. When low, data I/O are inputs and no analysis occurs.

The 4B/\3B signal specifies whether 3-bit or 4-bit ADPCM data is to be used.

The RESET is an active high input which initialises the MSM5218RS internal circuit. To be effective, must be held true for at least one V_{ck} time.

The S_1,S_2 inputs select the sampling frequency.

The SICK is a clock input for clocking in serial PCM data from an external ADC into the internal 12-bit shift register.

The ADSI signal is serial PCM data.

The \SCON output signals the start of conversion.

The SOCK is a MSM5218RS output which provides a 192 kHz signal which is synchronised with the output of the serial PCM data through the MSB/\SO pin. For all those is necessary to serial PCM data output be selected (DAS=H). Each bit of the 12-bit PCM data will be valid before the positive edge of this 192 kHz signal.

The DAS signal selected for analog signal output (DAS=L), or serial PCM data output (DAS=H).

The DAOUT is an analog output from the MSM5218RS circuit.

The MSB/\SO is the MSB of the data in the internal 10-bit DAC which will appear if analog signal output mode (DAS=L) is selected, or can be the serial PCM data clocked out when serial PCM data output mode is selected (DA=H).

In the next rows we'll refer to the internal structure of the MSM5218RS circuit with its functional features. Thus, this have a 12-bit shift register for the ADSI signal input processing, an internal oscillator coupled with a timing circuit, an ADPCM analysis stage block and an ADPCM synthesis stage block coupled with a 10-bit DAC and a 12-bit shift register in output. The ADPCM analysis stage and ADPCM synthesis stage blocks dispatches and receives the ADPCM data through a 4-bit bus.

The control logic block can be see in the figure 1. It is achieved with two flip-flop type D circuit and some logic gates NAND and NOT. It is present an asynchronous binary seven bits counter too.

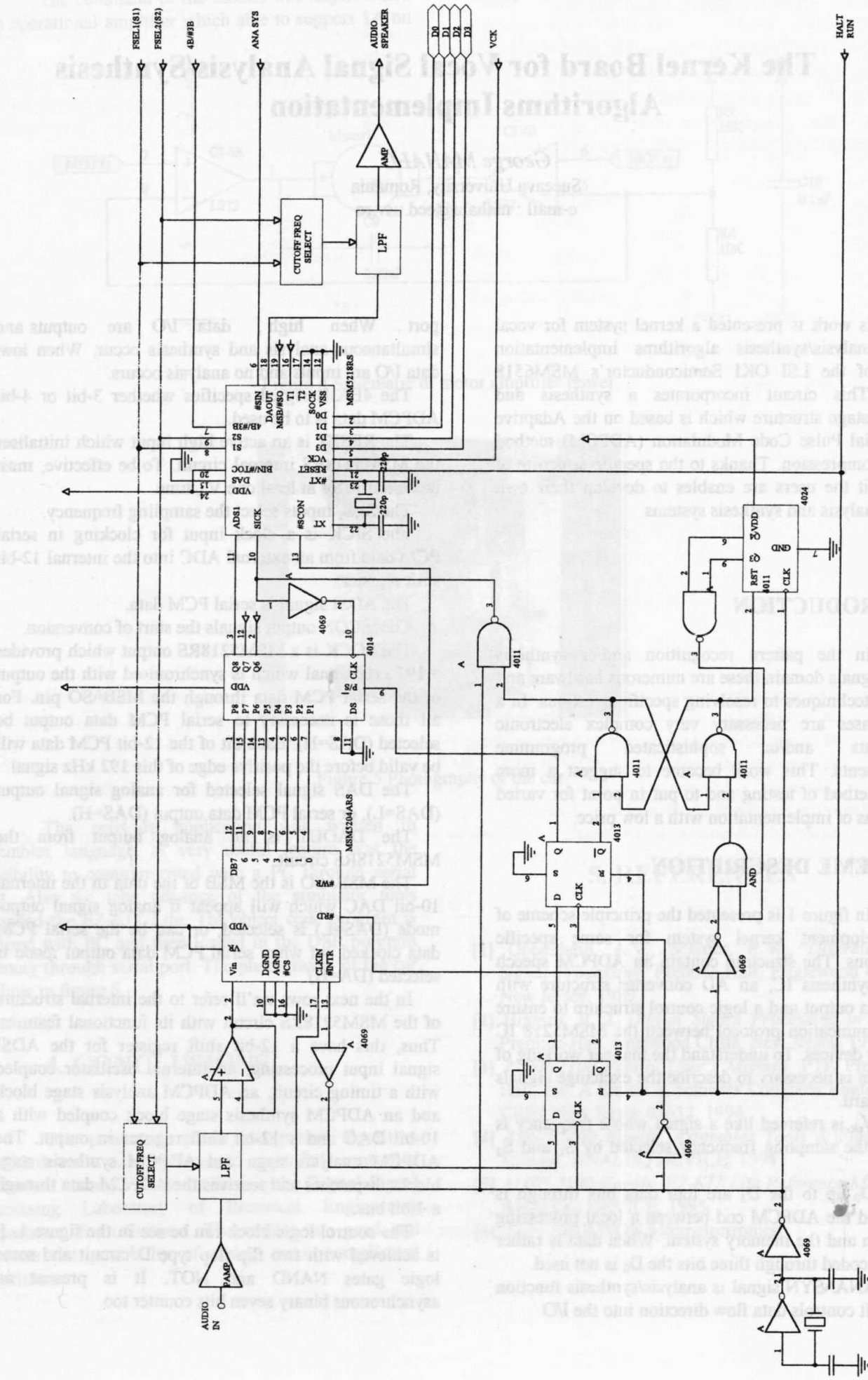

FSEL1(S1)
FSEL2(S2)
4B/#3B
ANA SYS
AUDIO SPEAKER
D0
D1
D2
D3
VCK
HALT RUN
AMP
CUTOFF FREQ SELECT
LPF
MSM5218RS
4B/#3B
S2
S1
VDD
DAS
BIN/#TOC
ADS1
SICK
#SIN
DAOUT
MSB/#SO
T1
T2
SOCK
VSS
D0
D1
D2
D3
VCK
RESET
#XT
#SCON
XT
220p
22
4069
4014
4011
4024
4013
AND
MSM5204ARS
DB7
#WR
#RD
VR
Vin
DGND
AGND
#CS
CLKIN
#INTR
PAMP
AUDIO IN

Figure 1

3. THE SYSTEM FUNCTION

Just after RESET signal applied on the HALT RUN input, a same internal RESET signal become active into the MSM5218RS circuit. The reset signal is latched within the LSI by the reset latch timing. Analysis is commenced by switching the external reset signal from H to L before a timing. The state of the reset input signal is strobed into the internal reset latch by the reset latch timing pulse. Analysis begins on the H to L transition of internal reset.

The analog speech signal is applied on the V_{in} pin of the MSM5204ARS circuit. This is an analog/digital converter with serial input/parallel output. The time of the acquisition is marked by the \SCON (Start CONversion) signal. When the ADC loaded a properly speech signal piece is generated an output signal named \INTR. This signal pass through those two flip-flop type D circuit (4013) and applies a positive pulse to the P/S input of the shift register with parallel/serial input serial output (4014). In this mode is made a parallel 8-bit load. Since that moment the data serial signal is await by the ADSI MSM5218RS input. To ensure a properly transfer of data signal between the 4014 register and the 5218 circuit will be necessary to synchronise the transfer on an appropriate clock signal. This clock signal is applied both SICK (on the MSM5218RS) and CLK (on the 4014) inputs. This clock signal is achieved like as burst of 12 pulses with 500 kHz rate by during high V_{ck} pulse. Therefore, up to 12 bits of PCM data may be strobed into the 5218 device by SICK. If more than 12 SICK pulses occur in a given V_{ck} cycle, only the last 12 are regarded valid. The cycle of SICK pulses must be completed before the next \SCON pulse.

For created that 12-bit burst is present in the structure the 4024 binary counter. After RESET signal the 4024 counts up to the 0CH code and dispatches one pulse. Thanks to this pulse the output NAND gate from the logical control subsystem is validated and supplies the SICK signal.

After all these steps, the 5218 dispatch a new \SCON signal for a new conversion command.

We can observe that first D flip-flop switches on the positive edge of the clock and thus supplies the \RD signal which command the 5204 converter to output data. The second D flip-flop switches on the next negative edge of the clock and supplies the P/S signal for the 4014 register (one positive level, therefore the active signal is P).

4. POSSIBLE ANALYSIS/SYNTHESIS SPEECH SIGNAL ALGORITHMS

After conversion of analog data signal to ADPCM code, this information can be stored into local system memory or in the PC memory through an acquisition on the parallel or serial port. In that moment can be made a properly processing after that the system can says if the speech signal is or not is recognised. Can be imagined varied appropriated algorithms. Thus, if in one memory array is stored the witness code (this code is stored at one previous moment) and in other memory array is stored the command code, the decision if this last code is or not is the correct code is done after some comparisons between the contents of those two memory arrays. Because those two codes not have often same phase and same "weight centre" will be necessary to launch more comparison session and will keep those results which exceed a defined level. The problem to establish this level is not simple but it can be resolved, in the first approximation by empirical mode. An other algorithm which will manage to resolve the decision of affiliation the command code to the witness class can analyses and compares the bit packets. For that the algorithm must does a properly choice of the bit packets following varied procedures. In this case can be involve some statistics methods for choice of the bit groups and for analyse those.

In the synthesis of the speech signal is often necessary to be used the filter rebuilding signal through the Fourier transform method. But on other hand, in used of the time analog windows methods can be utilise the Hilbert transform. This fact put the problem of use the digital filtering in rebuilding of the speech signal.

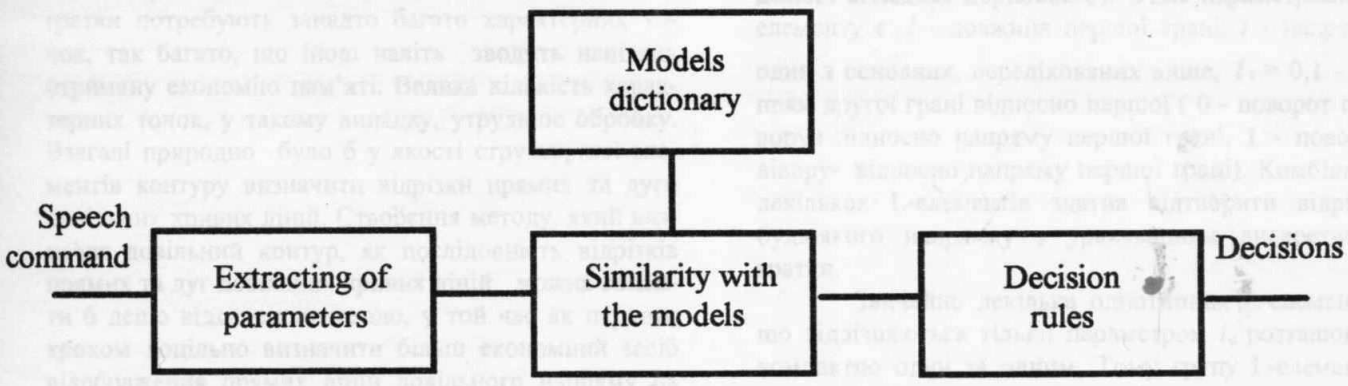


Figure 2

In the analysis domain can be involved numerous processes in understanding and recognition speech signal such as :

- the acoustic analysis, through are extracted acoustic parameters;
- the phonetic analysis, through are extracted speech sounds features;
- the prosodic analyse, through which to append the intonation information, the rhythm and accent, for purpose of the great linguistic units identification;
- the syntactic analysis, through which is testing the syntactic consistence for one hypothetically speech word comparative with the previous speech words;
- the semantic analysis, through which is check up of the understanding hypothetically sequence;
-the pragmatic analysis, through which are predicted most probable future words.

A simple structure for a determinist system of recognition can be that from the figure 2.

5. CONCLUSION

The kernel board just here presented have the advantage that has a very low cost and gives the possibility to develop a lot of software applications oriented to the pattern recognition. This system although is go like a distinct and independent structure however it can be interfacing with the PC computer and in this case it can be have all specific features. The required memory system is not so large because the achieved dates are converted on three or four bits through an ADPCM algorithm compression. On other hand this system has a very simple structure, though a high reliability. Because the work system frequency is high sufficiently we have the possibility to use it even for the vocal signal recognition, but we can use it same into the many other application such as the signs and finger print recognition (or other same). We can remark that for some applications have wrote the specific programs in C programming language with achieved and processing signal vocal purposes. Evidently, in these cases has used a PC interface structure type which running the program into the computer memory. Is possible in other cases to make a hardware link between this kernel board and an other microcontroller data processing system. In this last case the ensemble obtains a real autonomy.

REFERENCES

1. *The OKI Semiconductor Data Book - 1994/1995.*

2. D. E. Knuth - *Tratat de programare a calculatoarelor. Algoritmi seminumerici.* Editura Tehnică, Bucureşti, 1983.

3. *Data Book* - Microelectronica S.A. - 1991-1992.

4. Viorica Mărâi, Gheorghe Mărâi - *Comanda vocală a sistemelor tehnice* Editura militară, Bucureşti, 1991.