V
Автоматичне розпізнавання та синтез мовних сигналів

Automatic Recognition and Synthesis of Speech Signals

Захист мовних сигналів на основі часових переставлень

ОЛЬГА БАГРИНОВСЬКА, СВІТЛАНА ПЕЛЕХ, ІГОР ПРОЦЬКО, ЮРІЙ РАШКЕВИЧ

Університет "Львівська політехніка" *290646 Львів, вул. Степана Бандери 12*
Тел.: (0322) 39-8228 Факс: (0322) 74-7143
Електронна пошта: rvv@polynet.lviv.ua

Ol'ha Bahrynovs'ka, Svitlana Pelekh, Ihor Prots'ko, Jurij Rashkevych. Protection Speech Signals on Base the Time Transpositions.

It is described speech coding on base the time transposition for protection of speech information. Number codes and levels of security are shown. The questions of hardware implementation is considered.

Метод часового переставляння фрагментів мовного сигналу. Захист мовної інформації при оперативному обміні актуальний в багатьох ділянках людської діяльності. Поряд з кодуванням вхідної мовної інформації на базі складних нелінійних перетворень в часовій і частотній областях [1], широко застосовують метод часового переставляння фрагментів мовного сигналу.

Послідовність цифрових даних мови розбивають на фрагменти певної розмірности. Переставляння мовних сигналів накладає конкретні умови для забезпечення захисту мови. Так, наприклад, неможливе розміщення двох сусідніх фрагментів з початкової послідовності цифрових даних мовного сигналу, а через k-фрагментів. Вибір допустимих кодів буде менший n!, де n-кількість фрагментів у відрізку мовного сигналу. Менш трудомістке обчислення, ніж шляхом всіх можливих переставлень і визначення всіх допустимих кодів, можна проводити за алгоритмом побудови квадратної матриці з одиничних елементів. Побудовані квадратні матриці визначають тип переставляння стовпця відрізка фрагментів вхідного мовного сигналу. Кількість M матриць у найпростіших випадках буде (див. Табл.), де M - кількість матриць переставлень з відсутніми розміщеннями сусідніх фрагментів з вхідного мовного сигналу; M1 - кількість матриць переставлень, вибраних з M з відсутніми розміщеннями сусідніх фрагментів через k=1 фрагмент.

На основі методу часового переставлення можливе забезпечення розширення ступенів захисту мовної інформації: інверсне передавання фрагмента, що не спотворює частотну характеристику фрагмента; зміна розмірності фрагмента, відрізка, що складається з послідовності цифрових даних. Це забезпечує отримання біля мільйона кодових комбінацій для передавання мовної інформації.

n	n!	M	M1(k=1)
4	24	11	5
5	120	56	25
6	720	309	131

Особливості апаратурної реалізації. Реалізація пристроїв захисту з використанням часових переставлень не вимагає складних арифметико-логічних блоків, процесорів. Основними блоками є мовно-смуговий аналоговий інтерфейс [2] та оперативно-запам'ятовуючий пристрій, об'єм і організація якого визначає розмірність фрагментів, їхню кількість для переставляння. Формування адресів відрізків, фрагментів, зміна їхнього числа, генерація адресів запису/зчитування покладена на блок керування. Важливим питанням для пристроїв з даним методом кодування є синхронізація обміну інформації, забезпечення стабільної часової прив'язки відрізків мовної послідовності між сторонами обміну інформації.

Дані пристрої забезпечують захист мовної інформації в телефонних каналах зв'язку та в інших системах передавання інформації.

Література

1. J.Flanagan, M.Schroeder, B.Atal, R.Crochiere, N.Jayant, and J.Tribolet. "Speech Coding," IEEE Transactions on Communications, April 1979.
2. Showcase (DSP SPECIAL ISSUE), April 1995, Texas Instruments.

♦

Моделювання лінгвістичних зв'язків елементів фонетичного та морфологічного рівня української мови в системах автоматичного розпізнавання сигналів

МИХАЙЛО БОНДАРЕНКО, ВІКТОР БАВИКІН, ЗЕНОВІЙ КОНОПЛЯНКО, ОЛЕКСАНДРА СТОРОЖЕНКО, ГРИГОРІЙ ЧЕТВЕРИКОВ

Технічний університет радіоелектроніки *310726 Харків, просп. Леніна 14*
Тел.: (0572) 40-9446 Факс: (0572) 40-9113

Михаил Бондаренко, Виктор Бавыкин, Зеновий Коноплянко, Александра Стороженко, Григорий Четвериков. Моделирование лингвистических связей элементов фонетического и морфологического уровня украинского языка в системах автоматического распознавания сигналов.

Рассмотрены вопросы, связанные с исследованиями лингвистических связей, существующих между различными элементами фонетического и морфологического уровня украинского языка. Предложены адекватные математические модели на базе алгебры конечных предикатов, ориентированные на использование в системах искусственного интеллекта.

Розглядаються деякі з лінгвістичних зв'язків, що існують між різними елементами фонетичного рівня української мови. Матеріяли моделювання фонетичних відношень викладені наступним чином. На підставі вивчення даних фонетики, математично описуються відношення, що зв'язують окремі фонеми з системою їх фонетичних ознак. Дослідженню та формальному опису системи правил, що регулюють механізм акцентного зсуву під час утворення форм слів змінюваних частин мови, присвячена робота [1].

Вихідним матеріалом для моделювання режимів оброблення слів на фонетичному та морфологічному рівнях природньої мови служать лінгвістичні закономірності. Математичним апаратом моделювання є алгебра скінченних предикатів.

Алгебра скінченних предикатів (АСП) задається на множині M всіх n-місних k-значних предикатів, тобто функцій виду $y = f(x_1, x_2, \ldots, x_n)$, де $x_1, x_2, \ldots, x_n$ — літерні змінні, задані на множині A алфавіту букв A=$\{a_1, a_2, \ldots, a_k\}$, $y \in \{0,1\}$ - логічна змінна.

Роль базисних елементів множини M відіграють найрізноманітніші предикати виду

$$x_i^{a_j} = \begin{cases} 1, x_i = a_j, \\ 0, x_i \neq a_j. \end{cases}$$

Базисними операціями на множині A є кон'юнкція та диз'юнкція (можливий базис - кон'юнкція та імплікація).

В роботі [1] для задач моделювання режимів автоматизованого оброблення мови розроблена й проаналізована мінімальна система аксіом АСП. Доведено - АСП повна у розумінні того, що за допомогою її формул можна в аналітичному вигляді записати довільний скінченний предикат.

Кожному n-місному відношенню R, заданому на n-й декартовій степені A алфавіту, можна поставити у відповідність рівняння виду

$$f_R(x_1, x_2, \ldots, x_n) = 1, \tag{1}$$

що зв'язує як змінні $x_1, x_2, \ldots, x_n$, так і відношення R. Предикат f_R вибирається так

$$f_R(x_1, x_2, \ldots, x_n) = \begin{cases} 1, \text{ якщо } (x_1, x_2, \ldots, x_n) \in R, \\ 0, \text{ якщо } (x_1, x_2, \ldots, x_n) \notin R. \end{cases}$$

Записавши цей предикат у вигляді формули алгебри скінченних предикатів, отримуємо аналітичний запис відношення R в формі рівняння (1).

Комп'ютерне відтворення лінгвістичних процесів, описаних залежностями, що математично представлені у вигляді рівнянь АСП, можна здійснити шляхом розв'язання вказаних рівнянь на ЕОМ. Для цього необхідно мати практично задовільні машинні методи розв'язання рівнянь АСП. В якості основного методичного способу під час опису лінгвістичних зв'язків використаємо метод

перерізу. Згідно цього методу фрагмент тексту X, що нас цікавить, виокремлюється з контексту, а сам контекст відкидається. Дія контексту замінюється набором ознак Y. Потім вивчається й математично описується лінгвістичне відношення L: X L Y, що зв'язує фрагмент тексту X з набором його ознак. Метод перерізу дозволяє розбити складну задачу математичного опису лінгвістичних закономірностей на сукупність більш простих задач без неприпустимого спрощення й огрублення загальної задачі.

Повний математичний опис фонетичних й морфологічних зв'язків досягається простим об'єднанням у єдину систему рівнянь, що описують лінгвістичні відношення. У цій системі ознаки грають чисто службову роль: вони використовуються як проміжні змінні.

В принципі, вибір кожної ознаки й області її значень можна здійснити довільним чином. Але довільний набір ознак не може бути визнаний за оптимальний. Критерієм оптимальності служить простота отриманої системи рівнянь. Складність системи рівнянь не завжди легко оцінити й цим пояснюються ускладнення, що виникли в лінгвістів під час вирішення ними питання про те, якій з можливих систем ознак слід віддати перевагу.

АСП дозволяє описувати у вигляді рівнянь довільні відношення на скінченних множинах, а звідси й відношення, що описують мовну діяльність людини. Основна перевага такого способу опису полягає в можливості відтворити за допомогою апаратних та програмних засобів ЕОМ будь-який з множини режимів поведінки людини.

Література

1. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства. Харьков, Изд-во "Основа", 1984. - 144 с.

♦

Інтелектуальні усномовні інформаційні технології та системи

ТАРАС ВІНЦЮК

Інститут кібернетики НАН *252022 Київ, просп. Академіка Глушкова 40*

Тел.: +380 44 266-4356 Факс: +380 44 266-1570

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

Taras Vintsiuk. Intellectual Speech Technologies and Systems in Ukraine.

The state of art in the intellectual speech technologies and systems designing is discussed, especially through more than 30-years NAS Institute of Cybernetics experience in this field. The actuality of this problem for Ukraine is shown. The state technical programme for 1997-2002 period is proposed; problem solving ways, expected results and possible applications are given.

Аналіз зробленого в попередні роки. На протязі 1993—1995 років в рамках Державної науково-технічної програми 05, розділ 03 в Інституті кібернетики НАН виконувався проект 06.02.03/127-93 "Розробка методів та засобів автоматичного розпізнавання, синтезу та розуміння мовлення для створення інтелектуальних усномовних інформаційних технологій взаємодії людини та машини".

Стислий опис отриманих результатів: (1) Розроблені математичні моделі мовних сигналів, методи навчання та самонавчання автоматичному розпізнаванню мовлення, методи автоматичного розпізнавання та смислової інтерпретації мовних сигналів. (2) Розроблені методи прискорення прийняття рішень при розпізнаванні великих словників. (3) Досліджені фонетичні та просодичні характеристики мовлення та створені програмно-апаратні моделі синтезу мовлення (зокрема українського мовлення) за текстом з необмеженого словника. (4) Опрацьовані архітектури систем автоматичного розпізнавання та синтезу мовлення і систем усного діялогу; розроблені експериментальні системи усного діялогу. (5) Опрацьовані питання подальшого розвитку інтелектуальних усномовних інформаційних технологій та їх використання й впровадження в комп'ютерній телефонії, в діялогових системах, в системах взаємодії людини та машини, прийняття рішень. Відповідні пропозиції подані до державних науково-технічних програм.

Аналіз отриманих результатів. Ці результати головно є теоретичними. Вони стосуються подальшого розвитку ІКДП-технології, широко визнаної в світі. Ця технологія грунтується на економному описі (заданні, композиції(К)) мовних сигналів за допомогою ієрархічно (І) організованих стохастичних автоматних породжувальних граматик та на оптимальних рішеннях на основі динамічного програмування (ДП).

Основи ІКДП-технології. Ознаками, які використовуються при описі мовних сигналів, є миттєві передавальна характеристика мовного тракту та характеристика джерел його збурення або

різні їх еквіваленти — послідовності спостережуваних векторів-елементів або елементів-імен (символів або скалярів). Світ модельних сигналів задається ієрархічною структурою автоматних породжувальних граматик. Приклади ієрархії мовних образів: мікрофонема—дифон—склад—слово—речення (фраза)—передаваний смисл; фонема—склад—слово—речення (фраза)—передаваний смисл. Задаються стохастичні автоматні породжувальні граматики, що генерують (синтезують) допустимі модельні сигнали (послідовності елементів), які відповідають образам найнижчого рівня ієрархії (наприклад, мікрофонемам, фонемам або дифонам) й дозволяють обчислювати ймовірність спостережуваного сегменту (послідовності елементів) за умови образів найнижчого рівня.

Найуживанішим способом задання цих стохастичних автоматних породжувальних граматик є граф з детермінованими функціями переходів зі стану в стан: виділяються початкові та кінцеві стани, визначаються правила стиковки графів образів найнижчого рівня з певними початковими станами наступного образу цього ж рівня; при переході в дискретний момент часу в певний стан з визначеною імовірністю (залежно від стану й образу) генерується спостережуваний (модельний) елемент; генерація є статистично незалежною; залежним є детермінований порядок проходження станів. Всі старші образи задаються однією чи декількома транскрипціями, які виражаються в алфавіті образів на одиницю меншого рівня ієрархії (наприклад, декілька фонетичних транскрипцій на слово). Образи найвищого рівня ієрархії—передаваний смисл—специфікуються особливими транскрипціями (орієнтованою семантичною сіткою або типами допустимих речень).

Ймовірність спостережуваного сегменту з фіксованими границями за умови образів певного рівня обчислюється як згортка (сумування або максимізація) добутків (згідно транскрипції) ймовірностей сегментів образів нижчого (на одиницю) рівня за всіма варійованими границями між цими нижчими сегментами. Згортка виконується за допомогою багатоступеневого динамічного програмування. Багатоступеневість виникає тому, що ймовірність нижчого сегменту, в свою чергу, обчислюється як згортка за границями ще нижчих сегментів, аж поки не доберемось до сегментів найнижчих рівнів. В ІКДП-технології багатоступеневе динамічне програмування зводиться до ефективного одноступеневого.

Так само використовуються багатозначні рішення на кожному рівні ієрархії.

Обмежуючи кількість рівнів ієрархії, отримуємо розв'язання задач розпізнавання окремо вимовлюваних слів, зв'язного мовлення, що складається зі слів вибраного словника, ключових слів в потоці злитого мовлення, смислової інтерпретації квазізлитого (з паузами між словами) і зв'язного мовлення. Навчання та самонавчання розпізнаванню використовується для формування автоматних породжувальних граматик образів найнижчого рівня за навчальними вибірками. Синтез мовлення є складовою частиною ієрархічної генеративної моделі розпізнавання.

Отримані теоретичні результати стали основою створення ряду експериментальних апаратно-програмних зразків систем усного діялогу лінії RECH (моделі 3, 4, 121, "Корд"), які автоматично розпізнають як окремо вимовлювані слова, так і злите мовлення, що складається зі слів вибраного словника, синтезують (озвучують) довільні українські та російські тексти, роблять усний переклад з однієї мови на іншу, зокрема з української та на українську.

На жаль, ці системи мають обмежений робочий словник при розпізнаванні — до 1000 слів, задля реального часу потребують апаратної реалізації (напрклад модель RECH-121 є паралельною спеціялізованою обчислювальною системою реального часу), потребують автоматичного настроювання (навчання розпізнаванню) на голос диктора та робочий словник. В цих системах зовсім не реалізується автоматичне розуміння, що робить майже неможливою реалізацію диктувальних машин, які редагують та друкують тексти під диктування, машин усного перекладу з однієї мови на іншу. Бажаним також є значне покращення якості синтезованого мовлення, реалізація багатомовного синтезу мовлення. Але чи не найактуальнішою проблемою є перехід на реалізацію в сучасному комп'ютерному середовищі, наприклад в середовищі MULTIMEDIA, тобто програмними методами, без використання спеціяльних апаратних засобів оброблення усномовної інформації, які потребують тривалої розробки, обмежившись тільки стандартними "добавками" до комп'ютерів: так званими картами DSP (Digital Signal Processing), звуковими картами (Sound Blaster) тощо.

Вирішення всіх цих проблем потребує подальшої теоретичної розробки ІКДП-технології стосовно автоматичного розпізнавання, синтезу та розуміння усної мови та реалізації цієї технології на персональних комп'ютерах для використання в офісі, побуті, комунікаціях.

Стан проблеми. ІКДП-технологія автоматичного розпізнавання, синтезу та розуміння мовних сигналів, розроблена в Інституті кібернетики НАН, отримала визнання та поширення у світі. Розробки систем усного діялогу лінії RECH станом на 1990 рік були одними з кращих у світі. Так, в 1986–1990 роках за контрактами з ЮНЕСКО (Париж) була розроблена й успішно апробована багатомовна система усного діалогу для персональногло комп'ютера (використовувалось 7 мов, в тому числі і українська). Отже, українські розробки як теоретичні,

так і для використання були цілком конкурентноспроможними на рівні колишнього СРСР й іншого світу. Але десь з 1991 року нові теоретичні розробки суттєво, а розробки прикладних систем зовсім призупинились.

Тому станом на сьогодні Україна відстає на років шість порівняно з передовими західними країнами. На щастя, західні розробки орієнтовані виключно на західні мови, перепрограмування цих розробок на слов'янські мови, зокрема на українську чи російську, практично вимагає нової розробки. Ринок збуту не гарантується, тому Захід такі розробки не веде. Отже Україні, яка займала найміцніші позиції в ділянці усномовних інформаційних технологій та систем, надається можливість розвинути відповідні розробки й вийти на ринки СНД та Заходу.

На Заході набули поширення технології та засоби синтезу мовлення за текстом. Такі засоби озвучують будь-які виділені тексти. Добру якість мають синтезатори для англійської, французької, шведської та інших мов.

Зокрема, на ринку продаються (за декілька сот долярів) портативні словники-перекладачі з синтезаторами мовлення. Наприклад, англо-український (російський) та українсько(російсько)-англійський текстові словники-перекладачі: при перекладі окремих слів, які набираються за допомогою клавіятури, результат перекладу на англійську не тільки висвітлюється на екрані, а й озвучується, тоді як при перекладі на українську (російську) ці портативні словники-перекладачі "мовчать". Між тим, ці засоби знайшли б широке використання в побуті, в офісах.

Інформаційні технології синтезу (озвучення) українських та російських текстів давно опрацьовані і справа стоїть через відсутність джерел фінансування розробки.

Західні фірми зараз продають програмно-апаратні засоби автоматичного розпізнавання усної мови. Це:

1. Диктувальні машини вартістю до 3 500 долярів США, які є персональними (настроюються на голос користувача), оперують зі словником до 30 тисяч слів (слова вимовляються з паузами між ними); окрім програмного забезпечення до складу цих засобів входять мікрофонна гарнітура, Sound Blaster і карта DSP, що вставляється у персональний комп'ютер.

2. Усномовні інтерфейси прикладного програмування та введення інформації, які мають вартість від декількох сот до декількох тисяч долярів та які використовуються для заповнення електронних таблиць, управління операційною системою, інтегруються з різними прикладними пакетами; в цих системах обсяг робочих словників змінюється від декількох сот до десятків тисяч слів.

Хоч засоби автоматичного розпізнавання усної мови рекламуються часто як багатодикторні (дикторонезалежні) і такі, що розпізнають зв'язне мовлення, в дійсності прийнятна точність розпізнавання (біля 95%) забезпечується в тих системах, які настроюються на голос диктора та оперують окремо вимовлюваними словами.

На Заході вкладаються значні кошти на фундаментальні дослідження з тим, щоб розробити дикторонезалежні системи розпізнавання зв'язного мовлення, системи усного перекладу з однієї мови на іншу.

До інтелектуальних усномовних інформаційних технологій відносяться не тільки методи й засоби автоматичного розпізнавання, синтезу й розуміння того, що сказано, але й автоматичне розпізнавання, ідентифікація та верифікація за голосом того, хто говорить, а також автоматичне визначення за голосом функційного стану людини.

Окремий блок складають проблеми стискання та компресованої передачі усномовної інформації засобами телекомунікації, зокрема через INTERNET. Це організація онлайнової розмови двох осіб через персональні комп'ютери, з'єднані мережею INTERNET, організація голосової електронної пошти. Ці проблеми вважаються чи не найактуальнішими зараз на Заході. Економічний ефект від впровадження цих технологій зумовлений, зокрема, тим, що вартість міжнародних розмов знижується в декілька разів.

Взагалі, на Заході давно прийшли до висновку, що впровадження інтелектуальних усномовних інформаційних технологій та систем підніме рівень життя.

Актуальність для України. Необхідність розвитку інтелектуальних усномовних інформаційних технологій та систем в Україні зумовлене багатьма чинниками.

Чи не найголовнішим зовнішнім чинником є поширення та використання різноманітних інформаційних технологій, що надходять в основному з Заходу та є майже зовсім не пристосованими до вживання поширених в Україні мов. Це пояснюється тим, що західні фірми не в змозі "побороти" слов'янські мови й поки-що не йдуть на спільні розробки.

Між тим діє сильний внутрішній чинник, зумовлений широким використанням комп'ютерної техніки, інформаційних технологій, телекомунікації в побуті, в офісах, в управлінні та на виробництві та який потребує створення простих, доступних та ефективних засобів взаємодії людини та машини національними письмовою та усною мовами.

В той же час в Україні існує достатній висококваліфікований науково-технічний потенціял розробників в ділянці інтелектуальних усномовних інформаційних технологій та систем. Зокрема, в Інституті кібернетики НАН запропонована широко визнана і поширена в світі ІКДП-технологія. На протязі 80-х років в цьому інституті були виконані базові для СРСР розробки систем усного діалогу лінії RECH, в тому числі і для ВПК. Наприкінці 80-х років за контрактами з ЮНЕСКО була розроблена багатомовна система усного діалогу для мікрокомп'ютера.

Україна все ще зберігає свій науково-технічний потенціял в ділянці інтелектуальних усномовних інформаційних технологій та систем, щоб при відповідному фінансуванні виконати розробки, що можуть оволодіти ринками країн СНД і західного світу, оскільки комп'ютерні засоби усного та текстового перекладу, зокрема з української та на українську, будуть однаково потрібні й Заходові, який обов'язково буде "спілкуватись" з Україною та країнами СНД.

Основні напрямки розробок, очікувані результати та прогноз їх використання. Розробка інтелектуальних усномовних інформаційних технологій та систем потребує значних фундаментальних досліджень, які відносяться до усномовної інформатики.

Далі наведено можливі розробки, очікувані прикладні результати, форму їх реалізації, строки виконання та прогноз використання розробок.

А. Комп'ютерний синтез українського (також російського) мовлення; пакет програмного забезпечення для персонального комп'ютера зі звуковою картою; 01.97—01.98; використовуватиметься для озвучення (синтезу) довільних українських (російських) текстів; масове використання.

Б. Портативні англо-українські (англо-російські) та українсько-англійські (російсько-англійські) текстові словники-перекладачі з озвученням українською (російською) мовою; спеціялізований комп'ютер на покупних західних чіпах та мікроелектронній технології плюс українське програмне забезпечення; 01.97—12.98 (перша черга), 01.99—12.2000 (друга черга); масове використання в побуті, в офісах.

В. Усний словник-перекладач (англо-український та українсько-англійський); пакет програмного забезпечення для персонального комп'ютера зі звуковою картою та картою DSP (є варіанти і без карти DSP); 01.97—12.98 (перша черга — 5 тисяч слів), 01.99—12.99 (друга черга — до 50-100 тисяч слів); масове використання в побуті, в офісі.

Г. Диктувальна машина — машина для автоматичного друку та редагування текстів під послівне диктування; пакет програмного забезпечення для персонального комп'ютера зі звуковою картою та картою DSP; 01.97—12.98 (перша черга — 5 тисяч слів), 01.99—12.99 (друга черга — до 50-100 тисяч слів); використовуватиметься в офісах.

Ґ. Диктувальна машина — машина для автоматичного друку та редагування текстів під зв'язну мову; пакет програмного забезпечення для персонального комп'ютера зі звуковою картою та картою DSP; 01.97—12.2001 (в словнику—100 тисяч слів); використовуватиметься з демонстраційними цілями.

Д. Засоби усного діялогу для персонального комп'ютера; пакет програмного забезпечення для автоматичного розпізнавання, синтезу та розуміння усномовних сигналів для встроювання в пакети прикладного програмного забезпечення та операційні системи (персональний комп'ютер доповнюється звуковою картою та картою DSP); 01.97—12.98 (перша черга — 5 тисяч слів), 01.99—12.99 (друга черга — до 50 тисяч слів); поширюватиметься як доповнення або разом з пакетами прикладного програмного забезпечення.

Е. Машина для усного перекладу з української на англійську та навпаки — послівне наговорювання; пакет програмного забезпечення для персонального комп'ютера зі звуковою картою та картою DSP; 01.97—12.98 (перша черга — 5 тисяч слів), 01.99—12.99 (друга черга — до 50 тисяч слів); використовуватиметься в офісах.

Є. Машина для усного перекладу з української на англійську та навпаки — зв'язне мовлення; експериментальна система автоматичного розуміння, синтезу та усного перекладу зв'язного мовлення; 01.97—12.2001; використовуватиметься в демонстраційних цілях.

Ж. Засоби стискання, компресованої передачі та відтворення усномовної інформації та їх інтеграція з INTERNET для організації онлайнівської розмови двох осіб; пакет програмного забезпечення для персонального комп'ютера зі звуковою картою, картою DSP, факс-модемом; 01.97—12.97; використовуватиметься в комп'ютерній телефонії і може стати масовим засобом телекомунікації, використовуватиметься при організації голосової електронної пошти.

З. Автоматична ідентифікація та верифікація особи за її голосом, комп'ютерна фоноскопія; експериментальна система комп'ютерного аналізу фонограм; 01.97—12.99; можливе використання при аналізі фонограм в судовій та кримінальній експертизі.

◆

Усний словник-перекладач

ТАРАС ВІНЦЮК
Інститут кібернетики НАН
252022 Київ, просп. Академіка Глушкова 40
Тел.: +380 44 266-4356 Факс: +380 44 266-1570
Електронна пошта: vintsiuk@uasoiro.freenet.kiev.ua

Taras Vintsiuk. Spoken Vocabulary-Interpreter.

An information technology for spoken vocabulary-interpreter is proposed. For each speaker so called individual voice file that is both the alphabet of reference elements and the acoustical transcriptions of phonemes-threephones in this alphabet is introduced. The automatic recognition consists in the dynamic programming matching between the word signal to be recognized and word reference specified by its phoneme-threephone transcription. The problem of fast large-scale word speech recognition is discussed. The result of automatic recognition is then used for text translation, displaying and respective speech synthesis.

Актуальність проблеми. До перспективних інтелектуальних інформаційних технологій відносяться усномовні технології та системи. З-посеред останніх найважливішими є диктувальні машини, що друкують та редагують тексти під диктування, та системи усного перекладу з однієї мови на іншу, зокрема з української та на українську. В свою чергу, диктувальна машина може оперувати зв'язним чи послівним мовленням. Сьогодні реальним є створення інформаційної технології послівної диктувальної машини, коли слова вимовляються з чіткими (скажімо, тривалістю не менш ніж 0.5 с) паузами між ними. При цьому усномовне введення даних може бути коментованим чи без нього. При коментованому диктуванні, крім основної інформації, голосом подаються команди типу ІМЕННИК, ПРИЙМЕННИК, ОРУДНИЙ ВІДМІНОК, СКЛАДНЕ РЕЧЕННЯ тощо, що спрощує процес та підвищує надійність розпізнавання й сприймання даних. Першим і чи не найважливішим кроком на шляху створення послівної диктувальної машини є реалізація усного словника-перекладача, який має і самостійне значення та практичне застосування.

Суть інформаційної технології усного словника-перекладача полягає у наступному. Людина рідною мовою вимовляє слово. Результат автоматичного розпізнавання та перекладу висвітлюється на екрані монітора комп'ютера. Там же подаються варіянти вживання слова в іноземній мові. За бажанням користувача результати перекладу та варіянти вживання слова можуть озвучуватись (синтезуватись).

Структура усного словника-перекладача (УСП). УСП є програмою для персонального комп'ютера (ПК), який доповнюється звуковою картою (Sound Blaster) введення-виведення мовного сигналу, картою цифрового оброблення сигналів DSP (Digital Signal Processing) і, звичайно, мікрофоном та динаміком. Найкраще реалізується інформаційна технологія УСП в середовищі Multimedia. При достатній швидкодії ПК використання карти DSP не є обов'язковим.

Програмне забезпечення УСП складається із трьох основних функційних комплексів: автоматичного розпізнавання та перекладу, автоматичного синтезу мовлення, підготовчого (наповнення словника, формування індивідуального усномовного файлу тощо).

Лінгвістичний блок. Його основою є комп'ютерні орфографічні словники — звичайні словники, що введені в комп'ютер. Єдиною відмінністю є те, що всі слова та тексти супроводжуються знаком наголосу в словах. Це робиться вручну або автоматично під час наповнення словника. Лінгвістичний блок є відкритим — його завжди можна поповнити новими словами або вилучити окремі слова.

До лінгвістичного блоку відносимо також програмний модуль автоматичного транскрибування слів та текстів — перетворення орфографічного тексту в фонемний текст, що необхідно для автоматичного розпізнавання та синтезу мовного сигналу. Автоматичне транскрибування може виконуватись як на підготовчому етапі, так і в процесі розпізнавання або синтезу.

Індивідуальний усномовний файл диктора — акустичний блок. На підготовчому етапі, в режимі навчання розпізнаванню усномовних образів, формується індивідуальний усномовний файл диктора — сукупність прототипів (еталонів) фонем-трифонів для даного диктора. Фонемою-трифоном називатимемо звичайну фонему, яка розглядається в контексті з сусідніми фонемами: тією, що їй передує, і тією, що слідує за нею. Наведемо приклад орфографічної, фонемної та трифонної транскрипцій слова:

П'ЯТЬ — #ПьЙ<АТь# — ### ##Пь #ПьЙ ПьЙ<а Й<АТь <аТь# Ть## ###.

Так, в українській мові 65 базових фонем (розрізняємо наголошені та ненаголошені голосні, тверді та м'які приголосні, фонему-паузу тощо), відповідно буде 65^3 фонем-трифонів. Але багато

фонем-трифонів можуть бути об'єднані. Наприклад, трифони Б<АБ, Б<АД, Б<АҐ, Д<АБ, Д<АД, Д<АҐ, Ґ<АБ, Ґ<АД, Ґ<АҐ можна об'єднати в узагальнений трифон

ДЗВІНКА ВЗРИВНА <А ДЗВІНКА ВЗРИВНА .

Отже, йдеться загалом про декілька тисяч фонем-трифонів для кожної мови.

Кожна фонема-трифон ${}_uk_v$ задається індивідуальною акустичною транскрипцією $T_{{}_uk_v}$:

$$T_{{}_uk_v} = \left(j_{1,{}_uk_v}, j_{2,{}_uk_v}, \ldots, j_{q({}_uk_v),{}_uk_v}\right) - \tag{1}$$

послідовністю довжини $q({}_uk_v)$ з імен $j \in J$ еталонних елементів $e(j) \in E$ із сукупності E, що також є індивідуальною для диктора. Наприклад, в E всього $|J|=256$ еталонних елементів $e(j)$, і кожен еталонний елемент описується $(m+1)$-вимірним вектором

$$e(j) = a(j) = (1,\ a_1(j),\ a_2(j),\ \ldots\ ,\ a_m(j)). \tag{2}$$

Скажімо, $a(j)$ — 12-вимірний вектор предиктивних параметрів мовного сигналу, обчислених коваріяційним методом.

Підкреслимо ще раз, що індивідуальний усномовний файл диктора складають: 1) сукупність еталонних елементів E; 2) сукупність акустичних транскрипцій T всіх фонем-трифонів. Ці параметри обчислюються на підготовчому етапі в режимі навчання розпізнаванню образів за навчальною вибіркою — сукупністю мовних сигналів, які супроводжуються фонемними, а отже, і трифонними транскрипціями.

Блок автоматичного розпізнавання та перекладу. Мовний сигнал, одночасно з введенням в комп'ютер, піддається так званому попередньому обробленню. Поточний фрагмет сигналу із M дискрет f_n, $n = 1{:}M$ довжиною, наприклад, 15 мс ($M = 300$ при частоті дискретизації 20 кГц) характеризується коваріяційною матрицею B:

$$B(t,r) = \sum_{n=1}^{M} f_{n-t} f_{n-r}\ ; \qquad t,r = 0{:}m \tag{3}$$

та порівнюється з еталонним елементом квадратичною формою

$$g(B,e(j)) = \ln a^T(j) B a(j), \quad j \in J. \tag{4}$$

Отже, сигнал, що розпізнається, описується послідовністю матриць B_i, $i = 1{:}N$, де N — тривалість сигналу. Подальше розпізнавання виконується на підставі таблиці $g(i,j)$, $j \in J$, $i = 1{:}N$.

Будь-який сегмент $B_{\mu\nu} = (B_{\mu+1}, B_{\mu+2}, \ldots\ , B_\nu)$ розпізнаваного сигналу розглядається як реалізація фонеми-трифона. При порівнянні-розпізнаванні формується еталон фонеми-трифона за його акустичною транскрипцією $T_{{}_uk_v}$:

$$E_{{}_uk_v} = T_{{}_uk_v} E = \left(e\left(j_{1,{}_uk_v}\right), e\left(j_{2,{}_uk_v}\right), \ldots, e\left(j_{q({}_uk_v),{}_uk_v}\right)\right), \tag{5}$$

котрий піддається перетворенням $w \in W$:

$$wE_{{}_uk_v} = wT_{{}_uk_v} E = \left(\underbrace{e\left(j_{1,{}_uk_v}\right), \ldots, e\left(j_{1,{}_uk_v}\right)}_{w_1 \text{ times}}, \underbrace{e\left(j_{2,{}_uk_v}\right), \ldots, e\left(j_{2,{}_uk_v}\right)}_{w_2 \text{ times}}, \ldots, \underbrace{e\left(j_{q({}_uk_v),{}_uk_v}\right), \ldots, e\left(j_{q({}_uk_v),{}_uk_v}\right)}_{w_{q({}_uk_v)} \text{ times}}\right) =$$

$$= \left(e_1, e_2, \ldots, e_{\nu-\mu}\right), \tag{6}$$

де $w \in W = \left\{ w = \left(w_1, w_2, \ldots, w_{q({}_uk_\nu)}\right) : w_s \in \{1,2\}, s = 1{:}q({}_uk_\nu), \sum_{s=1}^{q({}_uk_\nu)} w_s = \nu - \mu \right\}$, (7)

так, щоб довжина перетвореного еталону збігалася з довжиною ν-μ сегменту $B_{\mu\nu}$.

Схожість сегменту $B_{\mu\nu}$ з трифоном ${}_uk_v$ визначаємо як найкращу суму схожостей відповідних елементів

$$G\left(B_{\mu\nu}, E_{{}_uk_v}\right) = \min_{w \in W} \sum_{s=\mu+1}^{\nu} g\left(B_s, \left(wT_{{}_uk_v} E\right)_s\right), \tag{8}$$

де $\left(wT_{{}_uk_v} E\right)_s$ — s-тий елемент в послідовності $wT_{{}_uk_v} E$.

Задача порівняння розпізнаваного сигналу з одним словом словника полягає у виборі такої сегментації сигналу на сегменти трифонів, щоби у відповідності з фонемною (а значить, і трифонною) транскрипцією слова досягалась найкраща сумарна схожість одразу на всіх сегментах. Ця найкраща сумарна схожість знаходиться в результаті розв'язання задачі динамічного програмування (ДП). Відповіддю розпізнавання оголошується те слово, яке забезпечило абсолютно найкращу сумарну схожість [1].

Отже, в принципі при розпізнаванні необхідно розв'язувати на таблиці $g(i,j)$, $j \in J$, $i = 1:N$ стільки задач ДП, скільки слів у словнику.

Обчислення коваріяційної матриці B_i, відповідного рядка $g(B_i,j)$, $j \in J$ таблиці $g(i,j)$ і порівняння (накопичення) схожостей виконуються за рекурентними формулами ДП одночасно (паралельно) для всіх слів словника і по мірі надходження спостережуваних елементів B_i, $i = 1:N$ [1].

Далі, за відповіддю розпізнавання (номером слова) звертаємось власне до словника-перекладача і отримуємо результат перекладу, порядок вживання в іноземній мові варіянтів перекладу тощо.

Автоматичний синтез мовлення зводиться до озвучення слів та текстів, які розмічені наголосами в словах. Спершу слово або окреме речення заміщуються фонетичним і потім трифонним текстом. Останній дозволяє побудувати еталон слова або речення у вигляді послідовності еталонних елементів-векторів, якими є параметри a предиктивної моделі мовотворення. Далі викликається модуль управління просодикою (основним тоном) та модуль обчислення темпу вимовляння. Модуль темпу вимовляння визначає w-перетворення початкових еталонів трифонів, а модуль просодики вносить додаткову інформацію про частоту коливань голосових складок. Синтез (породження) мовного сигналу виконується предиктивною моделлю мовотворення [2].

Стратегія управління автоматичним розпізнаванням та перекладом. Якщо не використовувати карти DSP, то сучасні комп'ютери класу Pentium 133 в змозі розпізнавати та перекладати в реальному часі тільки до двох тисяч слів. Між тим, потрібно працювати зі словниками до 100 і більше тисяч слів. Отже, необхідні процедури прискорення прийняття рішень. До таких процедур належать такі, що грунтуються на використанні інтегральних ознак [3]. Вони відсікають слова, що гарантовано не можуть бути претендентами у відповідь розпізнавання. Недолік цих процедур у тому, що вони потребують спершу повного надходження сигналу слова. Відомі також процедури, які відсікають «неперспективні» слова за початковою частиною сигналу. Але вони не гарантують збереження оптимального рішення.

Сучасні прийоми прискорення прийняття рішень мають організаційний характер: вводяться підсловники, слова з котрих можуть слідувати за попереднім словом. Стосовно усного словника-перекладача цей прийом виливається в те, наприклад, що спершу називеться буква алфавіту, а потім слово на цю букву.

Прикінцеві положення. Розроблення інформаційної технології усного словника-перекладача пропонується до виконання в рамках однієї з державних науково-технічних програм.

Література

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. — Киев: Наукова думка, 1987, 264 с.
2. Т. Вінцюк. Комп'ютерні автоматичні розпізнавання та синтез українського мовлення. — В кн. "Проблеми українізації комп'ютерів" (Матеріяли 2-ї міжнародної конференції у Львові), Київ, 1992, с. 21–32.
3. Taras K. Vintsiuk. A New Technique for Large-Scale ASR Based on Integrated Features. — Progress and Prospects of Speech Research and Technology, Proc. of the CRIM/FORWISS Workshop, Munchen, 1994, p.p. 18–25.

◆

Гомоморфне оброблення мови з використанням зменшеного часового вікна

КОСТЯНТИН ГУСЄЄВ, ОЛЕКСАНДР НОВОСЕЛЬСЬКИЙ

Національний технічний університет *Київ, просп. Перемоги 39*

Kostiantyn Husiejev, Oleksandr Novosel's'kyj. Homomorphical Speech Preprocessing with Reduced Time Window Using.

The article describes a modification of homomorphic analysis of speech - narrowed time windowing of cepstrum. It allows to obtain more smoothed spectrum. This method may be useful for estimation of formant frequencies, poles and zeroes of spectrum of speech signal.

При аналізі мови широко використовується гомоморфне оброблення (інша назва - кепстральний аналіз). З його допомогою вирішуються питання: оцінка джерела збудження (генераторної функції мови); оцінка резонансних властивостей мовного тракту (фільтрової функції мови).

Як відомо [2], гомоморфне оброблення виконується у 4 етапи: порція відліків оцифрованої мови зважується за допомогою вагового вікна (найчастіше використовують вікно Хемінга); ця порція обробляється прямим перетворенням Фур'є (як правило, застосовують алгоритм Кулі-Тьюкі швидкого перетворення Фур'є); обчислюються логарифми амплітуд гармонік; отримані значення обробляють зворотнім перетворенням Фур'є (результат зветься кепстром).

По області кепстру, що більша 4 мс, можна визначити ознаку "тон-шум" і частоту основного тону (ОТ). Якщо ділянка вокалізована, у кепстрі буде пік в місці, що дорівнює періоду ОТ. На початку кепстру (0-4 мс) міститься інформація про огинаючу спектру мови.

Щоб її отримати, потрібно обнулити в кепстрі ділянку від 4 мс до кінця і обробити кепстр прямим перетворенням Фур'є [2]. Розмір вагового вікна 4 мс отриманий з міркувань про мінімальний період або максимальну частоту ОТ людини Тот = 1/Fот = 1/250 Гц = 4 мс.

Миттєвий спектр мови вельми порізаний. Його локальні максимуми віддалені один від одного на величину частоти ОТ. Причиною цього є вплив генераторної функції мови. В отриманій за допомогою гомоморфного оброблення огинаючій спектру мови цей вплив усунуто. Огинаюча є точною апроксимацією спектру мови методом найменших квадратів [1]. Проте, наприклад, для автоматичної оцінки формантних частот отримана огинаюча непридатна, бо має значно більше максимумів, ніж дійсно є формант.

Користь дає звуження вагового вікна для оброблення кепстру до 1,7-2 мс. Оброблений таким вікном кепстр дає більш згладжений спектр. Кількість його максимумів дорівнює кількості формант. Тобто спектр, що отриманий обробленням кепстру звуженим ваговим вікном, є придатним для оцінки формантних частот, взагалі полюсів та нулів спектру мовного сигналу. Звуження вагового вікна до деякої межі (1-1,4 мс) суттєво не змінює вид спектру. Подальше звуження значно спотворює спектр.

Цікаве порівняння спектрів, що отримані за допомогою гомоморфного оброблення та лінійного передбачення 20-го порядку (був використаний асинхронний з ОТ аналіз з застосуванням автокореляційного методу). Максимуми спектру припадають на однакові частоти.

Виникло питання: реалізація якого методу отримання згладженого спектру працює швидше? На алгоритмічній мові TurboPascal 7.0 створена програма. Результати її роботи показали, що згладжений спектр шляхом лінійного передбачення отримується у 3-4 рази швидше, ніж за допомогою гомоморфного оброблення (точне значення залежить від конкретного комп'ютера).

Але, як вище згадувалось, аналіз кепстру (результат попереднього етапу оброблення мови) дає інформацію про ознаку "тон-шум" і частоту ОТ. Тобто гомоморфне оброблення як методика отримання інформації про генераторну і фільтрову функції мови може конкурувати з методиками на основі линійного передбачення. До речі, перетворення Фур'є дозволяє обробляти одночасно два набори дійсних чисел. Тобто гомоморфне оброблення можна прискорити майже в два рази, якщо обробляти декілька порцій данних.

Для перевірки вищенаведених положень був використаний мовний матеріал: звуки "А" під наголосом, які були вилучені з тричі промовленого трьома дикторами-чоловіками слова "буран". Частота ОТ була у межах 110...160 Гц, що значно менше 250 Гц.

Література

1. Маркел Дж.Д., Грэй А.Х. Линейное предсказание речи. - Москва Связь, 1980.
2. Оппенгейм А.В., Шафер Р.В. Цифровая обработка сигналов. - Москва: Связь, 1979.

◆

On Some Approaches to Speaker-Independent Computer Recognition of Speech

OLEKSANDR ZASYPKIN, MYKHAJLO OVETS'KYJ, MYKOLA CHERVIN, VLADYSLAV SHELEPOV

Institute of Artificial Intelligence

340048 Donetsk, Artema str. 118-b
Tel.: (0622) 92-6082 Fax: (622) 92-6082
E-Mail: shel@ipii.donetsk.ua

Александр Засыпкин, Михаил Овецкий, Николай Червин, Владислав Шелепов. О некоторых подходах к дикторонезависимомому компьютерному распознаванию речи.

Описывается система компьютерного распознавания речи, которая проявляет себя с достаточно высокой надежностью как дикторонезависимая. Последнее достигается за счет соответствующей предварительной обработки сигнала и DTW-усреднения эталонов.

The problem of creating speaker independent systems of computer speech recognition still remains actual. The present report is devoted to the description of a signal processing method that has proved to provide a high level of speaker independence.

A speech signal, received from microphone during 0.5 sec is digitized with frequency of 20 kHz. Let the applying of linear sliding filter $z_n = \left(y_{n-1} + y_n + y_{n+1}\right)/3$ be called the smoothing of the signal. The further processing will be performed with the difference between the original and 10 times smoothed signal. It allows to clear signal from the individual timbre to a considerable extent.

Let l be number of readings between two local maximums. Let us define the value z:

$$z = l,\ 2 \le l < 20; \quad z = 20 + (l-20)/6,\ 20 \le l < 50; \quad z = 25 + (l-50)/10,\ 50 \le l < 90; \quad z = 29,\ l \ge 90.$$

The nearest integer not exceeding z will be called the length of the appropriate full oscillation. Thus, the length of a full oscillation is taken into account more precisely if it is smaller.

Let us choose a segment of signal. Let n be a total number of full oscillations on the segment, n_k is a number of full oscillations of length k+1 (k=1,2,...). Let us build 29-dimensional vector

$$(x_1,\ldots,x_{28},\varepsilon). \tag{1}$$

There $x_k = \dfrac{n_k}{n}$. ε is the ratio of amplitude on the examined segment to the amplitude of the full signal.

Later the full speech signal is divided into 27 segments 368 readings in each of them (this is doubled quaziperiod of the principal tone for male mid-high voice) and the vector (1) is calculated for each of these segments. Thus the full signal is interpreted as set

$$a=(a_1,a_2,\ldots,a_{27}). \tag{2}$$

of vectors $a_1,\ldots,a_{27}$, i.e. as a trajectory in 29-demensional space.

Further recognition uses well-known DTW-algorithm. Let

$$e=(e_1,e_2,\ldots,a_{27}) \tag{3}$$

be the pattern of a word from recognizing dictionary, D_{ij} is modules differences coordinates vectors a_i and e_j sum, $C_{ij} = D_{ij} + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1})$. (4)

Then DTW-distance between a and e is $C_{27,27}$.

If (2),(3) are two patterns, we can average them in the following way. Let us denote by ~ accordance between two vectors from (2) and (3), which is defined so: $a_{27} \sim e_{27}$; further, if $a_j \sim e_i$, then in case, when minimum in (4) is $C_{i-1,j-1}$, we set $e_{j-1} \sim a_{i-1}$, if minimum is $C_{i,j-1}$, we set $e_{j-1} \sim a_i$, if minimum is $C_{i-1,j}$, we set $a_{i-1} \sim e_j$.

All corresponding vectors will be averaged. In this way for example it is easy to achieve computer interpreting every word of "*Буря мглою небо кроет*" as symbol 0 and every word of "*Вихри снежные крутя*" as symbol 1.

This sketch was used for a voice-dialing phone system. For each of the basic commands:

ноль, один, ..., девять, ошибка, запрос

five speakers have recorded 3 patterns of each command. Then these patterns were averaged. The system proved to be speaker - independent for 40 speakers, not used for training, with 98% accuracy.

References

1. Mitsevych A., Kovaliova O., Shelepov V. On system of voice induced dialing for concrete announcer//Обробка сигналів і зображень та розпізнавання образів .- Київ, 1994, p. 149.

♦

Нові підходи до розв'язання задачі автоматичної сегментації мовного сигналу

ОЛЕГ КАРПОВ, ОЛЕКСАНДР ХИЖА

Державний університет *320625 Дніпропетровськ, пров. Науковий 13*

Тел.: (05622) 44-7683

Електронна пошта: prisniak@dsu.dp.ua

Oleh Karpov, Oleksandr Khyzha. New Approaches to Automatic Speech Signal Segmentation.

A two-step speech segmentation scheme is considered in which some speech images are identified by amplitude characteristics and some by frequency ones. A transitive segment case is investigated.

Сегментація мовних сигналів, яка у загальному випадку є проблемою розпізнавання, у даній роботі реалізована такими двома способами.

1. **Розпізнавання поточного сегменту $\mathbf{a}_c$** як стану вектора **A** у класі (алфавіті) сегментів {a} в заданій системі параметрів {P} та в певному класі вирішувальних правил {R}:

$\mathbf{a}_c$ O **A**(a,P,R,t), c=1,2,...,n_a .

Вектор **A** має незчисленну кількість станів, з котрих вибираються деякі найбільш характерні {a} в залежності від параметрів {P} та роздільних можливостей вирішувальних правил {R}. Найчастіше це є: сегменти, що відповідають груповим ознакам; фонемам; частинам фонем; алофонам; транземам. Відповідно, алфавіт {a} вміщує m_a символів, прив'язаних до типу сегментів, що виділяються, а задача розпізнавання - це віднесення поточного сегменту $\mathbf{a}_c$ до одного з сегментів, заданих в {a}.

2. **Розпізнавання границі сегменту $\mathbf{b}_c$** як стану вектора **B** переходу від сегменту $\mathbf{a}_c$ до сегменту $\mathbf{a}_{c+1}$ у класі границь {b} в заданій системі динаміки параметрів {V} для деякого класу вирішувальних правил {r}: $\mathbf{b}_c$ O **B**(b,V,r,t), c=1,2,...,n_a+1.

Вектор **B**, так само, як і вектор **A**, має незчисленну кількість станів, як функція від b,V,r, і в першу чергу від t. Реально для кожного вектора також задається обмежена кількість m_b видів переходу від сегмента до сегмента в залежності від правил формування сегментів вектора **A**. При цьому можуть бути визначені додатково частини сегментів (початок, середина, кінець) та правила побудови внутрішніх границь.

В задачі сегментації одна із складностей полягає в тому, що границя, як правило, є розмита, тобто за різними параметрами (ознаками) формуються свої власні границі, тому доводиться обирати або одну з них, або серединне положення, або відносити ділянку границі до перехідного сегмента з відмінними один від одного значеннями параметрів в різних реалізаціях.

У даній роботі розглянута двоетапна схема сегментації і, відповідно, для кожного етапу вибрані свої набори мовних одиниць, котрі у деякій мірі перетинаються.

На **першому етапі** за мовні одиниці взято сегменти, які формуються за груповими ознаками: # – паузи, Ш–шумні, Т–тональні; де Т включає можливі сполучення ГП, ПГ, ПП, ГГ голосних Г та тональних приголосних П; Ш включає сполучення ШШ. Відповідно, можливі такі типи переходів (границь): #–Т, #–Ш, Ш–#, Т–#, #–Т, #–Ш. Задача розпізнавання #, Ш, Т є водночас задачею розпізнавання переходів між ними, яка розв'язується таким чином:

а) переходи Ш–#, Т–#, #–Т, #–Ш визначаються за станом амплітуди мовного сигналу в порівнянні з його пороговим рівнем;

б) переходи Т–Ш, Ш–Т, #–Ш, Ш–# визначаються за станом частоти переходу через нуль мовного сигналу в смузі 5000÷8000 Гц (при частоті переходу < 25 фіксується Т, у противному випадку - Ш).

На **другому етапі** здійснюється аналіз параметрів для розподілу станів Т і Ш на можливі складові частини Ш–Ш, Г–П, П–Г, П–П, Г–Г. Найчастіше цього виявляється достатньо для повного сегментування на ділянки, що відносяться до окремої фонеми, без визначення до якої саме.

Визначення точок сегментації або перехідних сегментів може бути виконано такими способами.

1. Інтервал тривалості і-го сегмента розбивається на таку кількість підінтервалів n, що вдвічі перевищує можливу кількість фонем; послідовно обчислюється близькість між суміжними підінтервалами сегменту від першого до останнього його підінтервалу та будується функція близькості. Далі для цієї функції обчислюються точки максимуму відстаней. Ці точки відповідають перехідним ділянкам і задають границю між фонемами всередині сегменту. Вибір

найбільш суттєвих змін забезпечується сортуванням максимальних відстаней, розміщенням цих величин в порядку зменшення і потім вибором n перших максимумів. Адреси максимумів є моментами зміни вектора станів.

2. Чисельне диференціювання функції S(F,t) двох змінних (частоти F та часу t), яку задано на прямокутній сітці матрицею відліків мовного сигналу $\{S_{i,j}\}$ за допомогою локального сплайну. Адреси відсортованих максимумів функції S(F,t) відповідають моментам сегментації мовного сигналу.

Якщо задано деякий поріг c^1, за рівнем сигналу визначається ширина піків в околі максимумів та знаходиться тривалість перехідного сегмента. І навпаки: при заданих обмеженнях на тривалість перехідного сегмента, знаходиться поріг рівня і, відповідно, можна відсікти локальні максимуми функції S(F,t).

Література

1. Карпов О.Н. Многокритериальный динамический иерархический подход к задаче распознавания речи // Вопросы прикладной математики и математического моделирования. Днепропетровск, ДГУ, 1991, с. 128-132.
2. Карпов О.Н. Способ фильтрации речевых сигналов // А.с. № 385311, "Официальный бюллетень изобретений и открытий СССР", № 25, 1973.

♦

Spoken Dialog Processing Research in Japan

RIICHIRO MIZOGUCHI

Osaka University,
Institute of Scientific and Industrial Research

8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan
Phone: +81-6-879-8415 Fax: +81-6-879-2126
miz@ei.sanken.osaka-u.ac.jp

Riichiro Mizoguchi. Spoken Dialog Processing Research in Japan.

In Japan, we have conducted a three-year project headed by Prof. Shuji Doshita [doshita@kuis.kyoto-u.ac.jp (Shuji DOSHITA)], Kyoto University, from April in 1993 to March in 1996, on spoken dialog processing by Grant-in-Aid for Scientific Research on Priority Areas by Ministry of Education, Science, Sports and Culture, Japan under the title: "Research on Understanding and Generating Dialogue by Integrated Processing of Speech, Language and Concept". Spoken language processing is related not only to signal processing but also to higher level processing such as language- and conceptual-level processing, which requires an integrated research methodologies including natural language and artificial intelligence research.

The project, therefore, consists of four groups each of which consists of five to six members to cover the wide range of research areas enabling collaboration towards to integrated research of spoken language processing. Some of the research results are summarized as follows:

A: Studies on Speech Recognition and Synthesis in Spoken Dialogue headed by Prof. Yasuhisa Niimi, Kyoto Institute of Technology [niimi@dj.kit.ac.jp]

Prof. Nakagawa of Toyohashi University [nakagawa@slp.tutics.tut.ac.jp] has studied robust recognition and interpretation methods for spontaneous speech. Many speech recognition systems employ syntactic constraints to reduce the search space of string candidates corresponding to speech. It is a good approach for read speech. In spontaneous speech, however, the syntactic constraint is much weaker than in read speech. Also, interjections, repairs and so on make speech recognition more difficult. His research is classified into three parts. (1) Comparison of some recognition methods for spontaneous speech to examine the robustness of each method. (2) Experiments in order to estimate the number of vocabulary for recognizing spontaneous speech and observation of human capability of error correction for mis-recognition results, and (3) Development of the robust spoken dialog system whose interpreter receives the recognition results that may include recognition-errors. The interpretation system is based on human's strategy of error correction in the above experiment. The performance evaluation using a travel guidance system of 500 vocabulary with perplexity 29 shows 70% understanding rate of utterance out of 232 utterances for seven speakers.

A sophisticated technology for generating response speech in advanced spoken dialogue systems has been developed by Prof. Keikichi Hirose of the Universtiy of Tokyo [hirose@gavo.t.u-tokyo.ac.jp].

B: Studies on Language Analysis and Generation in Spoken Dialogue headed by Prof. Hozumi Tanaka, Tokyo Institute of Technology [tanaka@cs.titech.ac.jp]

Prof. Tanaka of Tokyo Institute of Technology [tanaka@cs.titech.ac.jp] has investigated a robust and fast parser tailored to spoken language processing. For this purpose, a generalized LR (GLR) parser provides an exceptionally competent and flexible framework to combine linguistic information

with phonological information. The combination of a GLR parser and allophone models is considered very effective for enhancing the recognition accuracy in a large vocabulary continuous speech recognition. The main problem of integrating GLR parsing into an allophone-based recognition system is how to solve the word juncture problem, that is, how to express the phones at a word boundary with allophone models. He proposed a new method called CPM (Constraint Propagation Method) to generate an allophone-based LR table, which can effectively solve the word juncture problem. In his method, by introducing the allophone rules into the CFG and lexical rules, an LR table is generated, then the LR table is modified on the basis of an allophone connection matrix by applying the constraint propagation method. With this modified LR table, precise allophone predictions for speech recognition can be obtained.

C: Studies on Understanding and Presentation of Conceptual Information in Spoken Dialogue headed by Prof. Riichiro Mizoguchi, Osaka University [miz@ei.sanken.osaka-u.ac.jp]

Prediction of user's next utterances is an important technique to understand spoken dialog with the middle or large size vocabulary. Various kinds of dialog knowledge, such as utterance pairs, topics, and so on, are indispensable for prediction of the utterances. Prof. Mizoguchi of Osaka University [miz@ei.sanken.osaka-u.ac.jp] has engaged in developing a mechanism for utterance prediction based on a topic transition model, named TPN (topic packet network) and utterance motivation model. The topic information is very useful to predict the utterances. TPN is a static model of topic transitions represented in the form of a sort of network. In a TPN, topic packets (TP) bundle some topics and they are a priori linked each other. However, the topic of the utterance dynamically changes in a dialog as the dialog goes on. It is not easy to describe all the possible topic transitions in a static network especially for dialog tasks with the large vocabulary. A dynamic mechanism is appropriate for modeling topic transitions. Utterance motivation model has been devised to this end.

In a goal-oriented dialog, we usually have a definite motivation of making an utterance. The motivation has a close connection with the meaning of the utterance. A model of the utterance motivation is expected to flexibly predict topics in a dialog. He analyzed several dialogs concerning a task of route direction and trip reservation and came up with two-layered motivation model which is composed of communication and problem solving layers.

Motivations could be divided into two different levels: communication and problem solving levels. The motivation of communication is triggered by the state of information exchange. A state that "an attribute has multiple values" is categorized into one of the motivations of this level. The motivation of problem solving is related to how to use the derived information in the process of problem solving. "To make a comparison" is an example of the motivations of problem solving. The motivations of communication are classified into 8 categories and those of problem solving into 10 categories. The motivation of an utterance is modeled by combination of these two levels of motivation. A mechanism for predicting topics in the next utterance is described according to each utterance motivation. The former mechanism based on the TPN model had poor flexibility to dynamic change of dialogs. A model of the utterance motivation enables utterance prediction adaptive to situations in a dialog. Preliminary evaluation of the proposed mechanism showed the method is promising.

D: Studies on Modeling of Spoken Dialogue headed by Prof. Katsuhiko Shirai, Waseda University [shirai@shirai.info.waseda.ac.jp]

Prof. Itahashi of Tsukuba University [itahashi@milab.is.tsukuba.ac.jp] has conducted Corpora development projects. He took telephone shopping dialogues as an example of spoken dialogue recording. Forty seven telephone shopping dialogues were recorded by 13 male and six female speakers. The speakers were university students in their twenties. Out of 47 dialogues, 35 dialogues were considered favorable with the shortest utterance time of 3 minutes 2 seconds, the longest of 8 minutes 36 seconds and 4 minutes 30 seconds on average. So far, we have transcribed 10 dialogues into text. The smallest number of utterances was 89, the largest 216 and the average 123. The smallest number of characters of utterance was two, the largest 71 and average 11. Most of the dialogues has been converted into a CD-ROM.

Prof. Okada [okada@pluto.ai.kyutech.ac.jp] of Kyushu Institute of Technology has conducted his research "AESOPWORLD" in which two foxes' life are simulated with monologues and dialogues. The technologies employed include: (1) Language association with mind: Non-linguistic mental activities are associated with the deep structures of language. (2) Symbol grounding in perception or motion: Abstract symbols for representation of mental activities ground in concrete analogue signals for control of sensors or actuators. (3) Integration of multi-media: Vision, speech, and even motion are integrated through mind, and (4) Integration of intellect and emotion: Subjective knowledge processing is connected with objective emotion arousal. The foxes are realized as human-like agents who has a desire, makes a plan to satisfy it, takes actions to execute them, recognizes his

surroundings, gets emotional and utters those processes in natural language (Japanese). The demonstration has been successfully done.

◆

Аналіз методів часового масштабування мовних сигналів

ЮРІЙ РАШКЕВИЧ

Університет "Львівська політехніка" *290646 Львів, вул. Степана Бандери 12*
Тел./Факс: (380-322) 74-4143
Електронна пошта: rashkev@polynet.lviv.ua

Jurij Rashkevych. The Analysis of the Methods for Time-Scaling of Speech Signals.

This paper presents the analysis of the modern methods of time-scaling of speech signals. Two new time-domain algorithms for speech tempo changing using nonlinear transformation of speech signals time structure are developed. The results of simulation are also presented.

Вступ. Часова модифікація мовних сигналів (ЧММС) являє собою процес, що дозволяє стискувати або розтягувати мовний сигнал в часі без втрати його натуральності, розбірливості та тембру. Можливість модифікації швидкості відтворення мовної інформації широко використовується в різноманітних застосуваннях: передача мовлення каналами зв'язку, нормалізація сигналу в задачах розпізнаваня, читаючі машини для сліпих, логопедія тощо.

Історія розробки методів ЧММС сягає середини 40-х років, коли спочатку Gabor, а пізніше - Fairbanks запропонували надзвичайно простий підхід: ділити сигнал на рівномірні короткі відрізки і пізніше за допомогою повторення або виключення певних відрізків змінювати тривалість відтворення мовної інформації. Незважаючи на те, що якість мовного сигналу при цих процедурах суттєво погіршувалася, при невеликих коефіцієнтах зміни темпу, а також враховуючи просту технічну реалізацію, даний метод широко застосовується вже протягом понад півстоліття.

Протягом 70—80-х років було розроблено цілий ряд більш досконалих методів, найбільш відомими серед яких є: алгоритм гармонічного часового масштабування Малаха (D.Malah, 1979), аналіз-синтез за допомогою короткочасного перетворення Фур'є (КПФ) (M.Portnoff, 1981), оцінка мовного сигналу на основі модифікованого КПФ (D.Griffin, J.Lim, 1984), метод сумування з перекриттям (S.Roucos, A.Wilgus, 1985), синхронне з основним тоном сумування з перекриттям (E.Moulines, F.Charpentier, 1990).

Сучасні методи ЧММС. Майже всі сучасні методи часового масштабування мовних сигналів використовують ідею фільтрації початкового сигналу набором фільтрів з наступною модифікацією часового масштабу кожного із проміжних сигналів і синтезом вихідного сигналу зміненої тривалості.

Malah [1] розробив метод, при використанні якого мовний сигнал представляється у вигляді декомпозиції комплексних експонент з наступною зміною лише частоти кожної із експонент пропорційно коефіцієнту зміни темпу β. Амплітуда і тривалість кожної із складових залишаються без змін.

Розглянемо набір із L рівномірно віддалених смугових фільтрів з імпульсними характеристиками:

$$h_k(nT) = 2h(nT)\cos(\omega_k nT), \tag{1}$$

де $\omega_k = k\,\Delta\omega$ - центральні частоти, $k=1,L$, T - період дискретизації.

Вихідний сигнал k-го фільтру має вигляд:

$$y_k(nT)=2\sum_{r=-\infty}^{\infty} x(rT)h(nT-rT)\cos(\omega_k(nT-rT))=2\mathrm{Re}\{\exp(j\omega_k nT)X(\omega_k,nT)\}, \tag{2}$$

де $$X(\omega_k,nT)=\sum_{r=-\infty}^{\infty} X(rT)h(nT-rT)\exp(-j\omega_k rT) \tag{3}$$

являє собою дискретне короткочасне перетворення Фур'є фрагменту початкового сигналу. Якщо частота основного тону даного фрагменту мовного сигналу є відомою з достатньою точністю, то промасштабований в частотній області сигнал отримується із (2) у вигляді:

$$Y_k^{\beta}(nT) = 2\,\mathrm{Re}\{\exp(j\beta\omega_k nT)\,X(\omega_k,nT)\}. \tag{4}$$

Після синтезу модифікованих експоненціяльних компонент результуючий сигнал має тривалість початкового сигналу, але його частотні компоненти є відповідно промасштабовані. Відновити частотний масштаб дуже просто шляхом відповідної зміни швидкості відтворення. При цьому відновлюється частотний спектр і відбувається бажана модифікація часового масштабу.

Portnoff [2] розробив метод аналізу-синтезу мовного сигналу на базі комплексної модифікації проміжного представлення сигналу у вигляді КПФ.

Нехай $t(m,n)$ являє собою лінійний фільтр із змінними в часі параметрами, який моделює зміни голосового тракту в процесі мовлення, $T(n,\omega)$ - його Фур'є-перетворення, $P(n)$ - період основного тону. Гармонічне представлення вокалізованого мовного сигналу має вигляд:

$$X(n) = \sum_{k=0}^{P(n)-1} c_k(n)\exp(jk\phi(n)), \text{ де } c_k(n) = \frac{1}{P(n)} T(n,k\Omega(n))\exp(jk\phi 0) .$$

Модифікований в часовій області мовний сигнал зображується наступним чином:

$$X^{\beta}(n) = \sum_{k=0}^{P(\beta n)-1} c_k(\beta n)\exp(jk\phi(\beta n)/\beta) . \quad (5)$$

Як видно із виразу (5), як амплітуда, так і фаза початкового сигналу є модифіковані. Для досягнення високої якості вихідного сигналу необхідно дуже акуратно виконувати операції модифікації фази. Взагалі, даний метод є надзвичайно чутливим до похибок оцінки параметрів.

Griffin і Lim [3] розробили метод оцінки сигналу на основі модифікованого КПФ, який дозволяє уникнути проблем, пов"язаних з модифікацією фази.

Нехай $Y(\omega_k,nT)$ являє собою модифіковане КПФ. Ітераційний алгоритм, оснований на методі найменших квадратів, мінімізує середньоквадратичну похибку між $|X(\omega_k,nT)|$ і $|Y(\omega_k,nT)|$ в кожній із ітерацій. Часова модифікація початкового сигналу може бути досягнута шляхом отримання оцінки сигналу з амплітудою КПФ, близькою до масштабованої $|Y(\omega_k,nT)|$. Наприклад, модифікація $T_1 : T_2$ може буди досягнута за рахунок обчислення $|Y(\omega_k,nT)|$ із зсувом T_1 і $|X^i(\omega_k,nT)|$ із зсувом T_2 .

Roucos і Wilgus [4] запропонували алгоритм сумування з перекриттям (SOLA), який дозволяє синтезувати високоякісний сигнал, не використовуючи при цьому ітераційних процедур.

Moulines і Charpentier [5] розробили процедуру сумування з перекриттям, синхронну до періоду основного тону. Відмінність від попереднього алгоритму полягає в тому, що SOLA-алгоритм працює на стадії аналізу асинхронно і використовує автокореляційну техніку для ресинхронізації синтезованих коротких ділянок з основним тоном.

Необхідно відзначити, що всі розглянуті методи забезпечують високу якість промасштабованого в часовій області мовного сигналу лише в діяпазоні зміни коефіцієнту масштабування не більше 2. Це пояснюється невідповідністю використовуваних алгоритмів лінійного перетворення часової осі мовного сигналу до реальних процесів зміни темпу мовотворення, які за своєю суттю є очевидно нелінійними.

Адаптивні алгоритми. У відповідності із фонетичною теорією мовотворення основна інформація в мовному сигналі зосереджена в транзитних ділянках, довгі стаціонарні фрагменти мовного сигналу є інформаційно набагато біднішими. В той же час прискорення або сповільнення мови досягається в основному шляхом зміни тривалості стаціонарних ділянок.

Таким чином, ми можемо підвищити якість часового масштабування мовного сигналу, використовуючи адаптивний підхід: в значній мірі змінюючи тривалість стаціонарних сегментів сигналу та пауз, в той же час обережно відносячись до перетворення тривалості коротких перехідних ділянок.

В 1992 році Moulines і Charpentier запропонували алгоритм адаптивної трансформації часового масштабу мовного сигналу шляхом введення нового параметру - ймовірності вокалізованості, яка приймає значення 0 або 1, і лише ділянки з ймовірністю вокалізованості 1 підлягають модифікації їх тривалості. Незважаючи на надзвичайну спрощеність, цей підхід за словами авторів дозволив істотно підвищити якість та діяпазон регулювання темпу мовлення.

В Україні розроблення методів та алгоритмів нелінійної трансформації часового масштабу мовних сигналів ведеться з середини 80-х років [6] і базується на запропонованій Т.К.Вінцюком кусочно-лінійній моделі мовного сигналу. Основна ідея полягає на розбитті початкового сигналу на квазістаціонарні ділянки (КД) і подальшому перетворенні тривалості кожної із ділянок у відповідності до її початкової довжини.

Зобразимо мовний сигнал у вигляді послідовності його елементарних сегментів (ЕС) одинакової довжини: $X=(X_1, X_2, \ldots X_i, \ldots)$. Квазістаціонарна ділянка тепер буде являти собою повторений l_i разів сегмент X_i .

Алгоритм 1. Відкидається l_{i-1} елементарних сегментів в кожній КД. Тепер кожна ділянка буде представлена лише одним ЕС, і коефіцієнт зміни темпу залежитиме від вибраного порогу сегментації початкового сигналу на КД.

Алгоритм 2. Перетворення тривалості кожної із КД відбувається на основі формули:

l_o = entier (l_i/H) , якщо $l_i/H > 1$; $l_o = 1$, якщо $l_i/H < 1$; де l_i і l_o — тривалості вхідної та вихідної квазістаціонарних ділянок; H - експериментально визначений коефіцієнт для забезпечення необхідного коефіцієнту зміни темпу β.

Моделювання. Дослідження якості часового мастабування - регулювання темпу мовного сигналу проводилося на комп'ютері з використанням 20 мс тривалості ЕС. Сегментація проводилась на основі алгоритму, в якому мірою відмінності між сусідніми сегментами служить енергія похибки передбачення наступного сегменту на основі параметрів поточного:

$$q(l,n) = \sum_{i=0}^{C} \sum_{j=0}^{C} a_n(i)\, g_l(i,j)\, a_n(j) , \qquad (9)$$

де $g_l(i,j)$ - коефіцієнти автокореляції l-го сегменту, $a_n(i)$ - коефіцієнти апроксимуючого фільтру, $C = 12$ - порядок фільтру.

Закінчення сегменту фіксувалося при виконанні умови: $q(l,n) / q(l,l) > P$, $(P = 1.4)$.

Для Алгоритму 1 взаємозалежність між коефіцієнтом зміни темпу та порогом сегментації в діяпазоні $2 < \beta < 4$ визначається відношенням: $P = (\beta + 4.6)/6$.

Для Алгоритму 2 взаємозв'язок між H та β визначається формулою: $H = \exp((\beta-1)/1.35)$ для $1 < \beta < 3$.

Розбірливість фраз тривалістю 3 - 4 секунди при початковому темпі 80—90 слів в хвилину у випадку прискорення $\beta = 2.5$ складала 95% , у випадку прискорення $\beta = 3$ — понад 88%. При цьому аудитори відзначали деяку зашумленість вихідного мовного сигналу внаслідок відсутності використання спеціяльних методів спряження сусідніх КД у вихідному сигналі.

Література

1. D. Malah. "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals", IEEE Trans. on ASSP, vol. 27, pp. 121-133, April 1979.
2. M.R. Portnoff. "Short-Time Fourier Analysis of Sampled Speech", IEEE Trans. on ASSP, vol. 29, pp. 364-373, June 1981.
3. D.W.Griffin, J.S.Lim. "Signal Estimation from Modified Short-Time Fourier Transform", IEEE Trans. on ASSP, vol. 32, pp. 236-243.
4. S. Roucos, A. Wilgus. "High Quality Time- Scale Modification for Speech", IEEE Int. Conf. ASSP Proc., Tampa, pp. 493-496.
5. E. Moulines, F. Charpentier. "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones", Speech Communication, vol. 9, pp. 453-467, Dec. 1990.
6. A.C. 1406636 (СССР). Способ и устройство ускорения темпа речевой информации /Ю.Рашкевич и др., 1988.

♦

Особливості зміни темпоральної структури мовних сигналів в різних темпах мовлення

ЮРІЙ РАШКЕВИЧ, РОМАН МАРЦИШИН, ЗОРЕСЛАВА ШПАК

Університет "Львівська політехніка" *290646 Львів, вул. Степана Бандери 12*

Тел./Факс: (380-322) 74-4143

Електронна пошта: rashkev@polynet.lviv.ua

Jurij Rashkevych, Roman Martsyshyn, Zoreslava Shpak. The Peculiarities of the Speech Signals' Temporal Structure Changing while Different Rates.

Statistical analysis of the temporal structure of speech is done using 3 different rates: slow, normal and fast. The peculiarities of the process of different speech fragments' longitude changing are analyzed. The statistical model of speech signals' rate changing is developed.

Необхідність розроблення засобів високоякісної зміни темпу мовної інформації для задач логопедії, навчання іноземних мов, телебачення та радіомовлення ставить завдання пошуку нових

методів та алгоритмів регулювання темпу мовлення, що використовують процедури перетворення темпоральної структури мовних сигналів, максимально наближені до тих, які існують в процесі природньої зміни темпу мовлення людиною.

В доповіді на основі статистичного матеріялу розглянута класифікація елементів усної мови з точки зору ступеня зміни їх тривалості при різних темпах мовлення. Підкреслюється очевидна нелінійність зміни темпоральної структури мовних сигналів та її відмінність від прийнятих процедур в задачах розпізнавання мовних образів та їх часового масштабування.

Досліджується вплив на розбірливість та натуральність мови зміни тривалості окремих ділянок мовного сигналу, на основі чого пропонується алгоритм високоякісного регулювання темпу мови при невисоких коефіцієнтах зміни темпу, що характерно для вказаних областей застосування.

♦

Ітераційний алгоритм автоматичного визначення квазіперіодичних і неперіодичних ділянок мовного сигналу

НАДІЯ ТИМОФІЄВА

Інститут кібернетики НАН *252022 Київ, просп. Академіка Глушкова 40*

Тел.: (044) 267-6949

Nadija Tymofijeva. The Iterative Algorithm of the Automatic Determination of the Quasiperiodic and Aperiodic Speech Signal Subsegments.

For the determination of the quasiperiodic and aperiodic subsegments of the speech signal the k iterations are used. During each iteration the analyzed segment of the signal is divided in subsegments with the length $L \in \{L_{\min}, L_{\min}+\Delta, L_{\min}+2\Delta, \ldots, L_{\max}\}$ and the following determination of the quasiperiodicity of the neighbouring subsegment. The objective function for the valuation of the results of the decision of the problem is used which consider several criterions; k quantity of the iterations, $L_{\max}$ - the maximally possible length of the quasiperiod, $L_{\min}$ - the minimally possible length of the quasiperiod, Δ - the value of the increase of the quasiperiod.

Пропонується алгоритм автоматичної сегментації мовного сигналу, який дозволяє на заданому відрізку виділити квазіперіодичні і неперіодичні ділянки, а в квазіперіодичних - визначити довжину поточного квазіперіоду.

В процесі роботи алгоритму проводиться k ітерацій, на кожній із яких відрізок сигналу, що досліджується, розбивається на ділянки довжиною $L \in \{L_{\min}, L_{\min}+\Delta, L_{\min}+2\Delta, \ldots, L_{\max}\}$ з наступним визначенням квазіперіодичності сусідніх ділянок; $L_{\min}$ - мінімально можлива довжина квазіперіоду, $L_{\max}$ - максимально можлива довжина квазіперіоду, Δ - значення приросту квазіперіоду (визначається експериментально). Міра подібності, що використовується для визначення квазіперіодичності двох сусідніх ділянок на k-й ітерації, обчислюється з урахуванням кількох критеріїв. Цільова функція, що використовується для знаходження оптимальної траєкторії із усіх сформованих в процесі роботи алгоритму, також враховує кілька критеріїв. (Траєкторія - вектор $\tau = (\tau_1^k, \ldots, \tau_\eta^k)$, елементи τ_j^k якого відповідають відліку сигналу, який задає початок можливого квазіперіоду; $\tau_j^k \neq \tau_l^k$, $j, l \in \{1, \ldots, \eta\}$, $k \in \{1, \ldots, q\}$, q - кількість ітерацій, відповідно і кількість можливих траєкторій).

Нехай задано відрізок мовного сигналу $[n, n+N]$, який подамо у вигляді числової функції $f(x)$, $x \in \{n, \ldots, n+N\}$; $n \in \{1, 2, \ldots, N^*\}$ (N - довжина відрізка сигналу, в дискретах, N^* - довжина всього сигналу, в дискретах). Розіб'ємо його на ділянки довжиною $L \in \{L_{\min}, L_{\min}+\Delta, L_{\min}+2\Delta, \ldots, L_{\max}\}$. Для визначення квазіперіодичності відрізка сигналу використовуються різні міри подібності, наприклад [1]. В даному алгоритмі вважатимемо, що ділянка функції квазіперіодична, якщо

$$p_l = \frac{\min(a_j - a_{j-1}, a_{j-1} - a_{j-2})}{\max(a_j - a_{j-1}, a_{j-1} - a_{j-2})} > \varepsilon \quad , \tag{1}$$

$$s_l = \frac{\min(b_j - b_{j-1}, b_{j-1} - b_{j-2})}{\max(b_j - b_{j-1}, b_{j-1} - b_{j-2})} > \varepsilon , \tag{2}$$

$$\beta_l = |p_l - s_l| < \varepsilon' , \tag{3}$$

$j \in \{3,\dots,m+1\}$, $l \in \{1,\dots,m\}$, де a_j - відлік, для якого значення функції $f(a_j)$ на ділянці j довжиною L — найбільше; або $\left| |f(a_j)| - |f(x_t^*)| \right| < \alpha$, якщо $\dfrac{\min(a_j - a_{j-1}, L)}{\max(a_j - a_{j-1}, L)} > \varepsilon$ (x_t^* - відлік, для якого значення функції $f(x_t^*)$ - найбільше), b_j - відлік, для якого значення функції $f(b_j)$ на цій же ділянці — найменше; або $\left| |f(b_j)| - |f(\xi_t^*)| \right| < \alpha$, якщо $\dfrac{\min(b_j - b_{j-1}, L)}{\max(b_j - b_{j-1}, L)} > \varepsilon$ (ξ_t^* - відлік, для якого значення $f(\xi_t^*)$ — найменше); α, ε, ε' - коефіцієнти міри подібності, що визначаються експериментально; $j \in \{1\dots m\}$, $\in \{n,\dots n+N\}$, m - кількість ділянок довжиною L, на які розбивається відрізок функції $f(x)$ на k-й ітерації, $L \in \{L_{min}, L_{min}+\Delta, L_{min}+2\Delta,\dots,L_{max}\}$, $k \in \{1,\dots,q\}$. Відповідно введемо впорядковані множини $A = (a_1,\dots,a_{m+1})$, $B = (b_1,\dots,b_{m+1})$, $P = (p_1,\dots,p_m)$, $S = (s_1,\dots,s_m)$, де p_l, s_l - елементарні міри подібності, які дозволяють визначати квазіперіодичність j-ї і $(j-1)$-ї ділянок.

Для першої ділянки довжиною L оцінка квазіперіодичности виконується за виразами

$$p_j = \frac{\min(c, a_j - a_{j-1})}{\max(c, a_j - a_{j-1})} > \varepsilon, \tag{4}$$

$$S_l = \frac{\min(c, b_j - b_{j-1})}{\max(c, b_j - b_{j-1})} > \varepsilon, \tag{5}$$

$$\beta_l = |p_l - s_l| < \varepsilon', \tag{6}$$

$j = 2$, $l \in \{1,\dots,m\}$, де $c = L$, якщо обчислення проводяться для першого відрізка функції, і $c = \tau_\eta^k - \tau_{\eta-}^k$, де $^k = (\tau_1^k,\dots,\tau_\eta^k)$ - траєкторія, сформована на k-й ітерації попереднього відрізка функції $f(x)$, $L \in \{L_{min}, L_{min}+\Delta, L_{min}+2\Delta,\dots,L_{max}\}$.

Для k-ї ітерації, $k \in \{1,\dots,q\}$, використовуючи міри подібності (1)-(6), будуємо траєкторію τ^k, значення τ_j^k якої обмежують квазіперіодичні і неперіодичні ділянки відрізка функції $f(x)$.

Зауваження. Визначення квазіперіодичности j-ї ділянки здійснюється шляхом порівняння j-ї і $(j-1)$-ї ділянок та j-ї і $(j+1)$ - ї ділянок. Якщо дві ділянки $[j, j+1]$ чи $[j, j-1]$ задовольняють умови (1)-(6), то вважаємо, що j - ій ділянці відповідає квазіперіод.

Оцінка і вибір оптимальної траєкторії із усіх можливих, яка дає розв'язок задачі, проводиться за цільовою функцією, що враховує кілька критеріїв:

$$F(\tau^k) = \frac{\sum_{j=1}^{m(\tau^k)} \left(1 - \gamma_j(\tau^k)\right)}{\sum_{j=1}^{m(\tau^k)} \gamma_j(\tau^k)} + \sum_{j=1}^{m(\tau^k)} \gamma_j(\tau^k)\beta_j(\tau^k) + m'(\tau^k) , \tag{7}$$

де $\gamma_j(\tau^k) \in \{0,1\}$ і визначає квазіперіодичність j -ї ділянки на k -й ітерації. Якщо $\gamma_j(\tau^k) =$, то ділянці відповідає квазіперіод, якщо $\gamma_j(\tau^k) =$, то ділянка - неперіодична; $\sum_{j=1}^{m(\tau^k)} \gamma_j(\tau^k)\beta_j(\tau^k)$ - інтегральна міра подібності k-ї ітерації; $m'(\tau^k) = \dfrac{m(\tau^1)}{m(\tau^k)}$ - коефіцієнт, який встановлює зв'язок між кількістю ділянок першої і k - ітерацій; $m(\tau^1)$ - кількість ділянок (довжина кожної з яких - L_{min}), на які розбивається відрізок функції $f(x)$ на першій ітерації, $m(\tau^k)$ - кількість ділянок, на які розбивається відрізок функції $f(x)$ на k-й ітерації.

Задача полягає в знаходженні траєкторії τ^k, $k \in \{1 \ldots q\}$, для якої цільова функція (7) досягає найменшого значення при виконанні умов (1)—(6).

Алгоритм виділення квазіперіодичних і неперіодичних ділянок відрізка функції $f(x)$ виглядає так.

1. Задаємо L_{min}, L_{max}, Δ, N, n. На заданій функції $f(x)$ виділяємо відрізок $[n, n+N]$, який досліджується на квазіперіодичність і неперіодичність .

2. Обчислюємо $m(\tau^1) = N/L_{min}$ (враховується ціла частина від ділення).

3. Вважаємо $k = 1$.

4. Введемо впорядковані множини $\tau^k = (\tau_1^k, \ldots, \tau_\eta^k)$, $A = (a_1, \ldots, a_{m(\tau^k)+1})$, $B = (b_1, \ldots, b_{m(\tau^k)+1})$, $P = (p_1, \ldots p_{m(\tau^k)})$, $S = (s_1, \ldots, s_{m(\tau^k)})$, $\tau_l^k = 0, a_j = 0, b_j = 0, p_i = 0, s_i = 0$, $j = \overline{1, m(\tau^k)+1}$, $l = \overline{1, \eta}$, $i = \overline{1, m(\tau^k)}$.

5. Розбиваємо відрізок сигналу на ділянки довжиною L. Обчислюємо $m(\tau^k) = N/L$ (береться ціла частина від ділення).

6. За вказаним правилом для j-ї ділянки знаходимо значення $a_j \in A$ і $b_j \in B$, $j = \overline{1, m(\tau^k)}$.

7. За виразами (1)-(2), (4)-(5) обчислюємо значення $p_i \in P$, $s_i \in S$, $i = \overline{1, m(\tau^k)}$. Якщо $p_i < \varepsilon$ вважаємо $p_i = d$, якщо $s_j < \varepsilon$, вважаємо $s = d/2$. (Коефіцієнт d вибирається таким чином, щоб $d - d/2 > \varepsilon'$).

8. За виразами (3), (6) визначаємо квазіперіодичність j-ї ділянки. Якщо j-а ділянка задовольняє умовам періодичності (1)-(6), то присвоюємо $\overset{k}{l}$ відлік сигналу, який відповідає можливому початку поточного квазіперіоду.

9. Обчислюємо $\beta(\tau^k), \gamma(\tau^k), m'(\tau^k)$.

10. Вважаємо $k = k+1, L = L+1$.

11. Якщо $L <= L_{max}$, переходимо до пункту 4, в противному разі - перехід до пункту 12.

12. Вважаємо $q = k - 1$. За виразом (7) для k-ї траєкторії обчислюємо цільову функцію; $k = \overline{1, q}$.

13. Визначаємо величину k^*, для якої значення функції $F(\tau^{k^*})$ — найменше.

14. За розв'язок задачі беремо траєкторію $\tau^{k^*} = (\tau_1^{k^*}, \ldots, \tau_\eta^{k^*})$.

15. Кінець роботи алгоритму.

При обробленні мовних сигналів великих розмірів вибираємо наступний відрізок $[(n + (j-1)N + 1), n + jN]$, $j \in \{1, \ldots, t\}$, і для нього повторюємо п.п. 2-15 описаного алгоритму; t - кількість відрізків, на які ділиться мовний сигнал, що досліджується.

Алгоритм реалізовано мовою СІ для операційної системи MS WINDOWS 3.1 для ПЕОМ. Результати обчислень, що проводились описаним алгоритмом, збігаються із сегментацією мовного сигналу, проведеного візуально. Час обчислень для мовного сигналу довжиною 10 000—15 000 дискрет (0,5—0,75 с), виконаних на ПЕОМ IBM PC 486/33, — 4-5 с.

Література

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

◆

Оцінка методів розв'язання задачі навчання розпізнаванню сигналів мовлення

ОЛЕКСАНДР ЮХИМЕНКО

Інститут кібернетики НАН *252022 Київ, просп. Академіка Глушкова 40*

Тел.: (044) 267-6962

Александр Юхименко. Оценка методов решения задачи обучения распознаванию сигналов речи.

Приводятся две постановки задачи обучения. Приводятся экспериментальные данные, иллюстрирующие эффективность некоторых методов решения поставленной задачи.

Більш успішне розв'язання задачі навчання розпізнаванню сигналів мовлення може привести до підвищення надійності розпізнавання цих сигналів. При застосуванні ІКДП-технології розпізнавання та синтезу сигналів мовлення задача навчання полягає в виборі моделей фонем й оцінки їх параметрів згідно з якимись критеріями [1].

Розглянемо дві постановки задачі навчання. Нехай дана навчальна вибірка (НВ) сегментів $J^r_{\mu^r\nu^r}, r = 1:U_{k^1}$, з U_{k^1} реалізацій фонеми k^1 [1]. Тоді необхідно знайти такий розподіл $p(j/k^1), j \in J$, щоб $\prod_{r=1}^{U_{k^1}} P(J^r_{\mu^r\nu^r}/k^1) \to \max$ за умови $\sum_{j\in J} p(j/k^1) = 1$.

Цей максимум досягається при $p(j/k^1) = \frac{n(j/k^1)}{n(k^1)}$, де $n(j/k^1)$- кількість повторень елемента j в сегментах фонеми k^1, $n(k^1) = \sum_{j\in J} n(j/k^1)$ - загальна кількість елементів в сегментах фонеми k^1.

Отже, розв'язком задачі навчання в цій постановці є частоти появ елементів $j \in J$ в реалізаціях фонеми k^1, $k^1 \in K^1, K^1$ - множина фонем.

Більш робастна постановка задачі навчання. Нехай $J^r_{\mu^r\nu^r}, r = 1:U$ - загальна сукупність з U сегментів з НВ; $k^1(r)$ - фонема, до якої належить сегмент з номером r. Тоді для всіх фонем

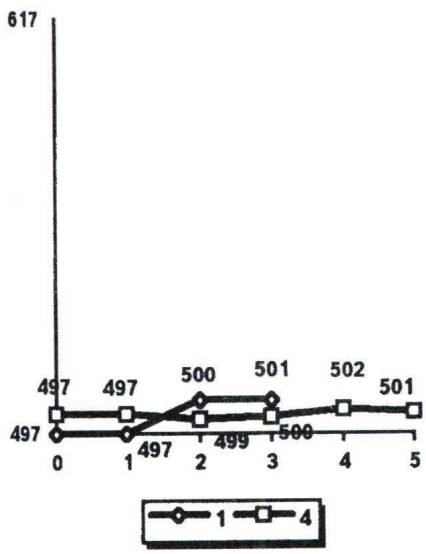

Рис.1

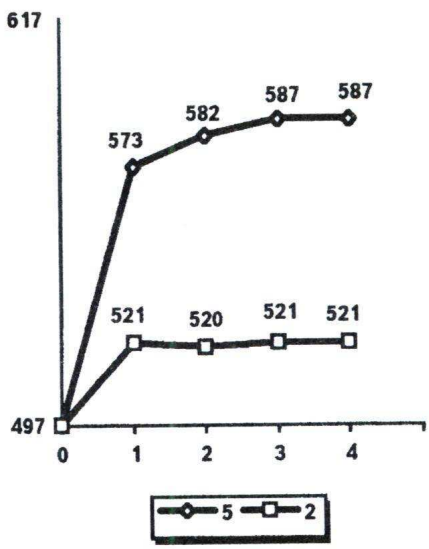

Рис.2

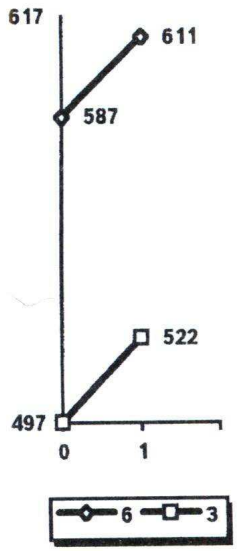

Рис.3

треба знайти такі моделі й відповідні їм розподіли, щоб виконувалось якомога більше нерівностей[1]:

$$P(J^r_{\mu^r\nu^r}/k^1(r)) > P(J^r_{\mu^r\nu^r}/t), \forall t \in K^1, t \neq k^1(r), \forall r = 1:U, \qquad (*)$$

де ймовірність сегмента $P(J^r_{\mu^r\nu^r}/t), t \in K^1$, обчислюється для кожної фонеми за своєю формулою відповідно до моделі [1]. Тобто, ймовірність кожного сегмента повинна бути на "своїй" фонемі більша, ніж на "чужих". Це забезпечить якомога більшу надійність розпізнавання.

Для розв'язання задачі навчання в другій постановці було застосовано декілька методів, що розподіляються на дві групи: моделі всіх фонем були найпростішими, тобто мали всього один стан; для фонем заводили більш складні моделі.

До першої групи відносяться: (1) Зменшення помилки розпізнавання через зменшення максимальної помилки розпізнавання окремих фонем шляхом збільшення критерія $\frac{m}{\sigma}(k^1)$ [2]. (2) Максимізація критеріїв $\frac{m}{\sigma}(k^1)$ всіх фонем $k^1 \in K^1$ [2]. (3) Максимізація загального критерію $\frac{m}{\sigma}$ для всієї системи (*) в цілому.

До другої групи відносяться: (4) Зменшення помилки розпізнавання через зменшення максимальної помилки розпізнавання окремих фонем шляхом вибору більш оптимальних моделей [3]. (5) Вибір для кожної фонеми найбільш оптимальних моделей. (6) Максимізація загального критерію $\frac{m}{\sigma}$ для всієї системи (*) після відбору найбільш оптимальних моделей кожної фонеми [3].

Експериментальні дослідження проводились на НВ з 1000 сегментів, що є реалізаціями 57 фонем. Замість системи (*), що складалася з 56000 нерівностей, розглядалась спрощена система з 1000 нерівностей [2].

Результати експериментів зображені на малюнках. Слід зауважити, що розв'язок задачі навчання в першій постановці забезпечив виконання 497 нерівностей з 1000.

Вісь ординат - кількість нерівностей, що виконуються в спрощеній системі. На рис. 1 зображені результати методів 1 й 4 (вісь абсцис - кількість кроків в алгоритмі). Рис. представляє результати методів 2 й 5 після чотирьох кроків (один крок - перебір всіх фонем). Рис. 3 - результати застосування методів 3 й 6 (вісь абсцис - початок й кінець алгоритму).

Отже, з представлених результатів випливає, що найкраще використовувати метод 6. Слід зазначити, що максимізація загального критерію $\frac{m}{\sigma}$ (метод 3 й в методі 6) дає майже однакове збільшення кількості нерівностей, що виконуються в системі, рівно ж як і максимізація критеріїв $\frac{m}{\sigma}(k^1)$ всіх фонем окремо (метод 2).

Література

1. Вінцюк Т.К., Юхименко О.А. Робастні постановки задачі навчання розпізнаванню сигналів мовлення. - Обробка сигналів і зображень та розпізнавання образів: Перша всеукраїнська конференція, 1992, с. 78-80.
2. Вінцюк Т.К., Юхименко О.А. Оцінювання ймовірнісних параметрів найпростіших моделей фонем. - Ймовірнісні моделі та обробка випадкових сигналів і полів: міжнародний симпозіум. - Тернопіль, 1993, т. 3, ч. 2, с. 11-16.
3. Юхименко О.А. Методи розв'язання задачі навчання розпізнаванню сигналів мовлення на основі використання моделей фонем різної складності. - Обробка сигналів і зображень та розпізнавання образів: Друга Всеукраїнська міжнародна конференція, 1994, с. 139-142.

♦♦♦