ИЗСЛЕДВАНИЯ ПО РАЗПОЗНАВАНЕ НА РЕЧ

доц.ктн.инж.Йордан Николов Колев - ВМЕИ Варна
инж.Камен Ралев Ралев - ВМЕИ Варна
инж.Тодор Димитров Ганчев - ВМЕИ Варна
инж.Ценка Лазарова Стоянова - ВМЕИ Варна

Настоящият доклад на тема "Изследвания по разпознаване на реч" се явява продължение на доклада "Изследвания и разработки по цифрова обработка на сигнали", като тук се представя развитието на изследванията по разпознаване на реч.

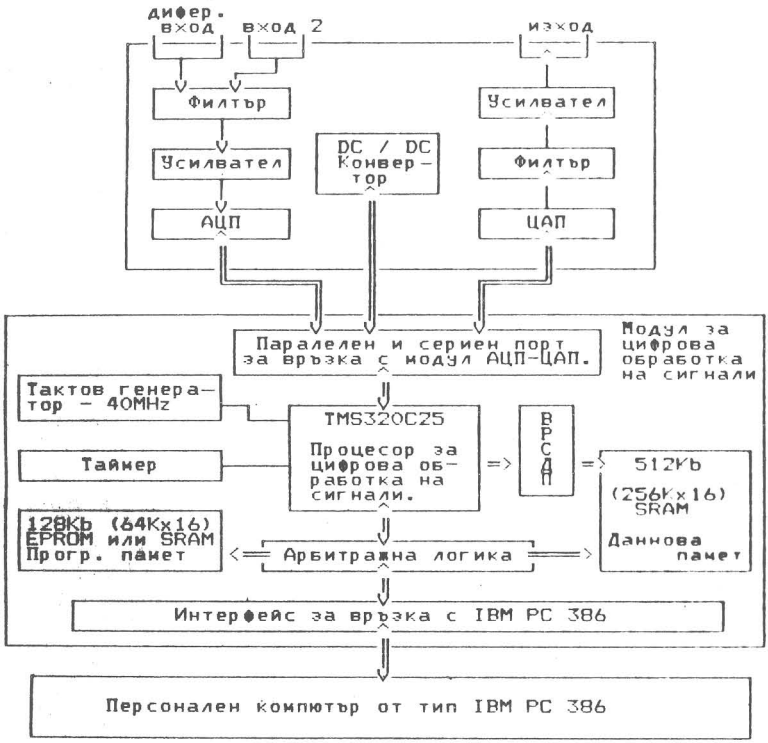
В доклада се разглежда система за разпознаване на изолирани думи от фиксиран диктор. Апаратното осигуряване на системата е показано на обобщена блокова схема фиг.1.

Трите основни съставни части на системата са персонален компютър IBM PC 386, модул за цифрова обработка на сигнали и модул за въвеждане и извеждане на реч. Чрез модула за въвеждане и извеждане се въвежда дума по праг. Върху тази дума в модула за цифрова обработка се извършва 256 точково бързо преобразуване на Фурие. Полученият по този начин спектър на думата се прехвърля в IBM 386, където се извършва обучение или разпознаване и се осъществява диалогът с оператора. Разпознаването се извършва чрез сравняване с еталон като се използва динанично програмиране и отрезово-постоянния модел на речта предложен от Винцюк. Като основни дескриптори на речта се използват амплитудите на кратковременния спектър. Програмата за разпознаване е написана на MATLAB и C, а въвеждането на дума и преобразуването на Фурие на Асемблер.

Разпознаването на изолирани думи се разви силно в последните години и вече се смята за напълно овладяно. Като най-сигурен е възприет подхода, при който входната последователност се сравнява с предварително подготвени еталони на думите от речника. Голямата точност на разпознаване тук се дължи на факта, че артикулационните ефекти в думата са заложени в еталона и се използват при разпознаването. Недостатък на метода е че при разширяване на речника времето за разпознаване нараства недопустимо.

Целта на предлаганата тук система е да разпознава надеждно 20-30 изолирани думи, с възможност за динанична смяна на речника, като предявява минимални изисквания към използваната платформа. Подхо-

дящ за това подход е предложен от Винцюк [4]. Той се основава на динамичното програмиране и отрезово-постоянният модел на речта. Преди да бъдат разгледани тези въпроси ще се спрем на първичната обработка на речта и използваните нейни дескриптори.



Блокова схема на системата за разпознаване на реч

Фиг.1.

<u>Първична обработка. Дескриптори на речта</u>

Целта на първичната обработка е да премахне излишъка в речта и да определи такива нейни параметри (дескриптори), които да се използват за разпознаване. В [1,2,4] се посочват различни видове дескриптори. Те могат да бъдат групирани условно в три групи: спектрални, кепстрални параметри и коефициенти на линейно предсказване. Тук са използвани класическите спектрални параметри - изходите от лентови филтри. Филтрите са разположени линейно в началото и логаритмично в останалата част на честотната ос. Филтрацията се реализира с КБПФ, което се изчислява за правоъгълен прозорец с дължина

30ms. Той се премества във времето със стъпка 15 ms. Честотата на дискретизация на входния сигнал във времето е 8500 Hz и предварително се филтрира с филтър с характеристика 1-z-1 (отговаря на характеристиката на излъчване на речевият тракт).

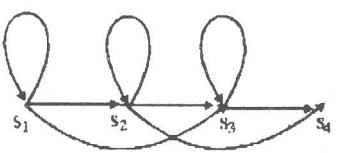
Получаването на този тип дескриптори е сравнително просто и бързо, поради което са използвани тук. В бъдеще ще бъдат използвани кепстрални параметри или коефициенти на линейно предсказване.

Процеса на съставяне на еталона и сравняването му с входната последователност е основан на отрезово-постоянния модел на речта.

Отрезово-постоянен модел на речта

В резултат на първичната обработка речта се представя с последователността X={x1,x2, ... , xl}, където xi е вектора от дескриптори на речта, асоциран с i-я момент от време. Този модел допуска, че тази последователност може да бъде представена от еталонната последователност E={e1,e2, ..., eN), N<L, като ei може да се повтаря в граници от mi до Mi. Границите на повторение образуват т.н. темпорална (временна) транскрипция на думата и заедно с ei, i=1:N образуват еталона.

Същият модел може да бъде интерпретиран чрез скрити вериги на Марков. При него всяко състояние si отговаря на генерация на р елемента ei, mi<p<M, като функцията на състоянието е равномерно разпределена в интервала (mi, Mi). Модела е показан на Фиг.2.3.



Фиг. 2.3

Сравняването на думата и еталона става чрез динамично програмиране

Динамично програмиране

Стандартната процедура на динамично програмиране описана в [5] е видоизменена в съответствие с отрезово-постоянния модел на речта и се извършва по-формулите дадени по-долу [4]:

$$Fki(1) = Fk(i-1)(1) + g(xi,ek1),$$

$$Fki(s) = \max_{mks \le u \le Mks} \{Fk(i-u)(s-1) + Gki(u,s)\}, \quad s=2:(Smk-1)$$

$$Gki(u,s) = \sum_{v=i-u+1}^{i} g(xv, eks),$$

$$Fki(Smk) = \max(Fki(Smk-1), Fk(i-1)(Smk) + g(xi, ekSmk)$$

Тук Fki(s) е интегралната разлика на първите i елемента от входната дума с първите s елемента от еталона k. g(xi,ekj) е елементарното разстояние между xi и ekj. В случая се използва евклидово разстояние. {u} са потенциално-оптималните дължини на сегментите, на които се разделя входната последователност в резултат на ДП. Mks и mks са максималния и минималния брой на повторение на eks, Smk е максималната възможна стойност на s. С k навсякъде е означен номера на еталона. Всички неопределени величини в дясната част се приравняват на $-\infty$, освен Fk0(1)=0. Изчисленията започват от момента i=0 и продължават до i=l (l-дължина на входната последователност). Fkl(Smk) е интегралната разлика между входната последователност и k-я еталон. Минимума на Fkl(Smk) по всички k определя най-близката дума в речника.

По-долу е показан процеса на сравняване на сричката на с нейния еталон. Процесът обхваща два етапа: определяне на елементарните сходства между елементите и определяне на интегралните критерии за сходства F(..i) между думата и еталона от началото до i-я елемент.

Съставяне на еталоните (обучение)

Процесът на обучение също използва динамичното програмиране. Първоначално на дадена реализация на думата се извършва самосегментация - разделя се на сегменти по такъв начин, че елементите във всеки сегмент да са максимално близки помежду си. Броя на сегментите се определя предварително и отговаря на акустичната дължина на думата. Въз основа на самосегментацията се определят първичния еталон и темпорална транскрипция на думата. Използвани са следните формули:

$$(22)\quad dki(u,s) = \max_{e} \sum_{v=i-u+1}^{i} g(xv, e),$$

потенциално оптималните дължини на сегментите могат да бъдат намерени чрез

$$(23)\quad uki(s) = \underset{mks \le u \le Mks}{\operatorname{argmax}} \{Fk(i-u)(s-1) + dki(u,s)\}$$

$$Fki(s) = Fk(i-uki(s))(s-1) + dki(uki(s),s),$$

където s=Sm:1, i+1:l.

И отново Fk0(1)=0 и uk0(1)=0, а всички неопределени F(...) в дясно са равни на $-\infty$.

След това процеса на обучение продължава със сегментиране на нови реализации на думата по текущия еталон. Итерационно се изчисляват

новите еталонни елементи и транскрипции, докато еталона престане да се променя. За целта се добавят следните формули:

(21) $uki(1)=uk(i-1)(1)+1;$

$$uki(s)= \underset{mks\le u\le Mks}{\operatorname{argmax}} \{ Fk(i-u)(s-1) + \sum_{v=i-u+1}^{i} g(xv, eks)$$

$$uki(Sm)=0, \text{ ако } Fki(Sm-1)\ge Fki(Sm) + g(xi, eksm)$$

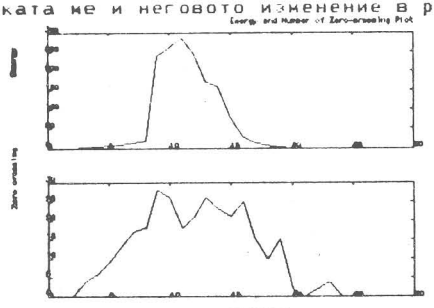
$$uki(Sm)=uk(i-1)(Sm) + 1, \text{ в обратния случай.}$$

uki(Sm) са потенциално оптималните дължини на сегментите и определят темпоралната транскрипция на думата.

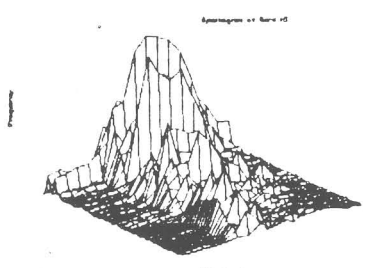
Текущите еталони се определят по формулата:

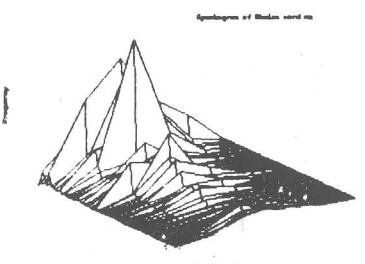
(24) $$eks= \underset{e}{\operatorname{argmax}} \sum_{r} \sum_{i} g(xir, e)$$

Когато g е евклидово разстояние eks може да се определи, като средно аритметично на определящите го елементи. Ако g е хемингово разстояние (броя на несъвпадащите елементи), eks може да се определи като вектор, чиито компоненти се срещат най-често в набора определящ eks. По-долу са даден четвъртия еталонен елемент от сричката ме и неговото изменение в резултат на две итерации.



Следват спектрограмата на сричката ме





и еталона съставен в резултат на горе описаното обучение.

Заключение

Описаният тук метод е сравнително бавен - свързан с много изчисления, особено при пресмятане на елементарните сходства. Процесът на обучение е още по-бавен и трудоемък. Въпреки това той е работоспособен и е подходящ за използване при разпознаване на ограничен брой изолирани думи.

Възможно е усъвършенстване на метода в следните посоки:

1. Избор на по-подходяща мярка за елементарно сходство, отразяваща по-добре физиологичното възприемане на речта.

2. Използване на други дескриптори, които запазвайки информацията в речта ще компресират нейния обем. При това времето за разпознаване може съществено да се снижи.

3. Преминаване към пофонемно разпознаване на речта, при което еталоните елементи се определят на базата на предварително фиксиран в резултат на обучение набор. Това ще доведе до по-бързо разпознаване, но с повече грешки.

Библиография

1. Рабинер, Шафер, Цифровая обработка речевых сигналов, Москва, Радио и Связ, 1981г.

2. Маркел, Грей, Линейное предсказание речи, Москва, Связ, 1980

3. Сапожков, Михайлов, Вокодерная связ, Москва, Радио и связ, 1991

4. Винцюк, Анализ, разпознавание и интерпретация речевых сигналов, Киев, Наукова думка, 1987

5. Computer Speech Processing, 1985