Разпознаване на реч с невронни мрежи

доц.ктн.инж. Йордан Николов Колев - ТУ Варна
ас. инж. Иван Георгиев Булиев - ТУ Варна
инж. Тодор Димитров Ганчев - ТУ Варна

Разпознаването на реч отдавна не е предмет само на авангардни изследвания. Фирми - лидери в компютърния бизнес предлагат на пазара интегрални схеми и системи за разпознаване на отделни думи с обем на речника до 200000 думи. Цените, на които тези схеми се предлагат варират в широки граници. Интегрална схема на Sensory Circuits Inc. с възможност за разпознаване на речник до дванадесет думи се предлага на цена под $4. Друга схема на Kurzweil Applied Intelligence Inc. с възможност за разпознаване на тридесет до шестдесет хиляди думи се предлага на цена под $1000 [1]. Естествено това разпознаване е далеч от равнището на което го правят хората, но дори и така то е достатъчно полезно. Тенденция е бързо навлизане на речевия интерфейс в диктофоните, автомобилите, телефоните, както и като помощно средство за общуване с инвалидите [2].

Проблемът за разпознаването на реч у нас също става актуален. Простото използуване на вече разработените схеми и системи не е възможно поради естествената езикова бариера. Фирмите производителки на подобни устройства ги създават с възможност за разпознавани на думи произнесени не на български език. Настоящият доклад представя първите резултати от тези изследвания.

На този етап съществуват три основни типа алгоритми за разпознаване на реч. Това са:

- алгоритми основани на динамичното програмиране;
- алгоритми на основата на скритите модели на Марков;
- алгоритми използуващи невронни мрежи;

Съществуват и някои други, но при тях не са налице сериозни успехи.

Първите два подхода са доста добре проучени и на тяхна основа са проектирани по-голямата част от съществуващите системи за разпознаване на реч. Невронните системи са сравнително нов подход. Първите идеи за създава-

нето на подобни мрежови структури датират от шестдесетте години, но реални възможности за тяхната реализация се появяват едва през осемдесетте, когато електронноизчислителните машини са в състояние да осигурят необходимата изчислителна мощ.

Усилията на колектива бяха насочени в овладяване на този подход, като последен по хронология и може би най-перспективен в областта на разпознаването на реч. От една страна се преследваше усвояване на работата с апарата на невронните мрежи, а от друга страна решаването на задачата за разпознаването би позволило създаването на конкретни устройства, например, устройство за набиране на телефонни номера произнесени на глас (Voice Dialing).

Редица трудности усложняват решаването на задачи от типа на разпознаване на реч [3]. Те са много и разнообразни, но могат да се групират в четири групи:

- трудности при разпознаване на фонемите, дължащи се на изменчивостта на фонемите;

- трудности при отделяне на думите;

- проблеми произтичащи от разликите в дикторите и средата;

- трудности произтичащи от това, че все още не е изяснен механизмът по който хората разпознават речта;

Преодоляването на тези трудности е обемиста и сложна задача. Като се има в предвид че това е един от първите опити за използване на невронни мрежи, задачата за разпознаване беше максимално опростена, като бяха наложени някои ограничения. Мрежата трябваше да разпознава цели отделни думи. Общият брой думи в речника беше неголям - десет. Съзнателно беше отстраняван (доколкото е възможно) фоновия шум при записа на отделните реализации. Заложено беше дикторозависимо разпознаване с цел да се използуват по-малко записи на думи за обучение.

Изкуствените невронни мрежи (обикновено се наричат просто невронни мрежи) представляват структури от отделни програмни единици наречени неврони [4, 5]. Идеята на тези мрежи е изграждане на модел на мозък и осъществяване на програмна симулация на мисловната дейност на човека. Всеки един неврон може да има множество входове и само един изход. Състоянието на

този изход се определя от сигналите подадени на входовете му, чрез подбор на определен брой коефициенти и тип на предавателна функция.

Между два и повече неврона могат да се осъществят определени връзки. Структура от свързани по определен начин неврони се нарича невронна мрежа. Тези структури могат да бъдат най-разнообразни, като невронните мрежи се разделят основно на два типа:

- рекурентни - налице са 'затварящи' връзки, т.е. връзки, които по някакъв начин осигуряват информация от изхода на мрежата към някои от нейните входове;

- нерекурентни - всички връзки следват посоката вход-изход;

В зависимост от това колко неврона са свързани последователно между входа и изхода, мрежите биват еднослойни и многослойни. В един слой на мрежата могат да бъдат включени различен брой неврони.

Два са основните моменти, тогава когато за решаване на определен проблем се използува невронна мрежа. Първият е свързан с определянето на структурата на мрежата. Няма определени правила в това отношение. Разчита се на въображението и усета на този, който организира мрежата. В този смисъл е много важно натрупването на практически опит. Вторият момент е свързан с процеса на обучение на мрежата. Това е процес при който се извършва подбор на коефициентите на всеки един неврон, като целта е за възможните състояния по входовете на мрежата, състоянието на нейният изход максимално да съвпада с предварително зададено. Критерий за добре 'обучена' мрежа обикновено е стойността на сумарното средноквадратично отклонение на изхода на мрежата. Обикновено броят на коефициентите, които трябва да се определят е доста голям и затова този процес е доста продължителен. Съществуват различни правила за обучение, като изборът на правило отново зависи от този, който организира мрежата. След като мрежата е 'обучена', начинът за нейното използуване се свежда до изчисляване на предавателната функция за всеки един от нейните неврони.

Структурата на една система за разпознаване на отделни думи включва няколко компонента:

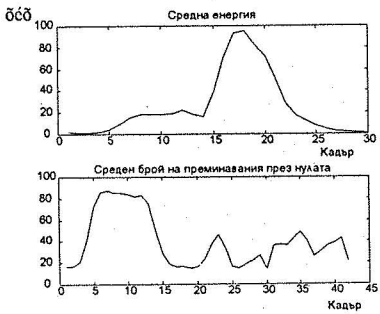
- откриване на началото и края на думата;

- оценка на параметрите на речта;

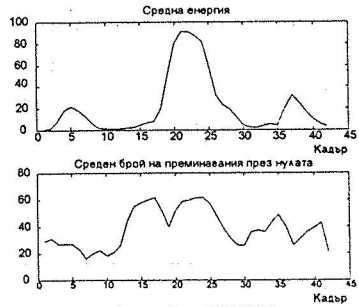
- вземане на решение от невронната мрежа за това, коя е произнесената дума;

Първите два елемента бяха реализирани върху стартовата развойна система DSP Starter Kit (DSK) на фирмата Texas Instruments за сигналния процесор TMS320C50. Изграждането на невронната мрежа беше направено на персонален компютър IBM-PC/i386 в среда MATLAB.

При подготвянето на входните данни за невронната мрежа се процедираше по следния начин. С помощта на DSK оцифровахме с честота на дискретизация 8kHz аналогов сигнал от изхода на микрофонен усилвател. За всеки 256 отчета през интервал от 128 отчета се изчисляваше средната енергия на сигнала и брой на преминаванията му през нулата. По тези два параметъра ставаше определянето на началото на думата [3]. Фиг.1а и фиг.1б пояснява алгоритъмът за определяне на началото на думите започващи със вокализиран ('ОЦЕНКА') и невокализиран звук ('СТАРТ').



фиг. 1а - 'ОЦЕНКА'



фиг. 1б - 'СТАРТ'

И в двата случая се следи дали нивото на средната енергия на сигнала и средният брой преминавания на сигнала през нулата не надвишават предварително избрани прагове. Праговите нива са обикновено функции на фоновият шум. Ако е налице такова надхвърляне се следи дали в рамките на зададен интервал от време средната енергия ще надвиши втори, по-висок праг. Ако и това условие бъде изпълнено се счита, че началото на думата е било в началото на тези 256 отчета, за които е бил надвишен някой от първите два прага.

Експериментите показаха, че с помощта на този алгоритъм действително се получава много добро определяне на началото на думата.

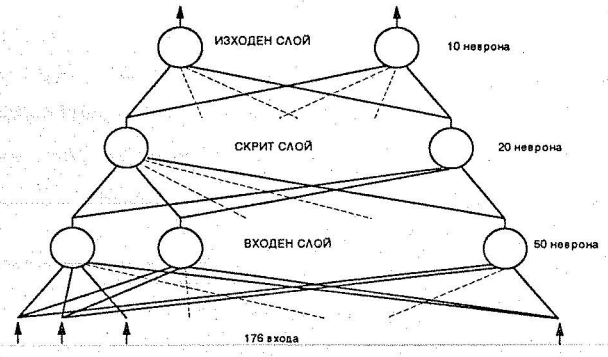
Откриване на края на думата не беше реализирано, като вместо това се извършваше запис за определено време, достатъчно за да бъде изговорена и най-дългата дума.

Като описатели на речта бяха избрани средния брой преминавания през нулата на сигнала, средната стойност на енергията на сигнала за 256 поредни отчета и девет коефициента на линейно предсказване изчислени за същия интервал.

След откриване на началото на думата за всеки 'кадър' от 256 отчета тези дескриптори се изчисляваха и запомняха. Кадрите освен това се презастъпваха с 128 отчета. Избраното време за запис осигуряваше запис на дескрипторите от шестнадесет такива кадъра. Разположени във ред тези параметри образуваха един 176-елементен вектор, който беше използуван като информация за изговорената дума. Всяка от десетте цифри беше изговорена от един и същ диктор по десет пъти и за всяка от тях бяха записани по десет такива вектора.

Невронната мрежа беше изградена в среда MATLAB. Бяха експериментирани различни структури мрежи, но най-добри резултати се получиха със структурата показана на фиг.2.

Невронната мрежа е трислойна. Входният слой се състои от 50 неврона с по 176 входа. Изходният слой е изграден от десет неврона, всеки от които активира изхода си при разпознаване на съответната дума. Връзката между

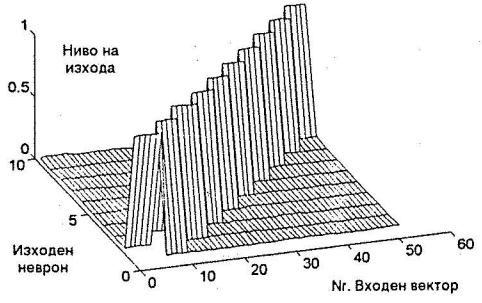


фиг. 2 Структура на невронната мрежа

входния и изходния слоеве става с трети междинен скрит слой от двадесет неврона.

Обучението на мрежата беше направено за около 5000 цикъла по правилото за обратно увеличаване на грешката (back propagation error). Като количествена мярка за процеса на обучение се използува сумарната средно-квадратичната грешка по изход на мрежата.

Беше проверено поведението на обучената мрежа върху използуваните за обучение думи. Оказа се, че всички те се разпознават напълно. Беше направен и съществения експеримент - разпознаване на други, неизползувани до този момент реализации на думи от същия диктор. И в този случай думите бяха правилно разпознати. Проблеми с разпознаването имаше когато думите бяха произнасяни по неестествен за диктора начин, или когато фоновият шум беше с по-високо ниво. На фиг.3 са представени нивата на изходните неврони в при подаване на входа на думите използувани за обучение.



фиг. 3 Резултат от обучението на мрежата

Получените резултати са обнадеждаващи. Работата в тази област продължава. Усилията са насочени към търсене на по-добри дескриптори на реч-та, подобряване на мрежовите структури, реализиране на по-голям брой цикли на обучение., Първата най-близка реална цел е пренасянето на невронната мрежа върху развойната система на Texas Instruments и пускането и в действие в реално време.

Литература

1. R.A.Quinel, Speech Recognition: No longer a dream but still a chalenge, EDN, January 19, 1995

2. R.A.Sharman, How the Technology Supports Dictation, The Computer Journal, Vol.37, No.9, 1994

3. Маркел, Грэй, Линейное предсказание речи, Москва, Связь, 1980

4. Neural Networks Toolbox, User's guide for MATLAB consumers.

5. David P. Morgan, Cristopher L. Scofield, Neural Networks an Speech Processing, Kluwert Academic Publishers, Boston, 1991