Петта национална конференција со меѓународно учество ЕТАИ'2000
Fifth National Conference With International Participation ETAI'2000
Охрид, Република МАКЕДОНИЈА – Ohrid, Republic of MACEDONIA
21–23. IX 2000

I1-1

ХИБРИДЕН HMM/ANN СИСТЕМ ЗА ПРЕПОЗНАВАЊЕ НА МАКЕДОНСКИ ГОВОР

Иван Краљевски[1], Драган Михајлов[2], Дејан Ѓорѓевиќ[2]

[1]Ветеринарен институт ,
ул. Лазар Поп Трајков 5-7, МК-1000 Скопје, Република Македонија
vetinst@unet.com.mk
[2]Електротехнички факултет, Универзитет Св. Кирил и Методиј,
П. Фах 574, МК-1001 Скопје, Република Македонија

Извадок – Дизајниран е систем за препознавање на македонски говор базиран врз хибридна структура со комбинирање на моделот на скриени Маркови вериги и вештачки невронски мрежи.

Клучни зборови – препознавање на говор, невронски мрежи, скриени Маркови вериги, динамичко програмирање

1. ВОВЕД

Проблемите за дизајнирање системи за препознавање на говор се предмет на голем број истражувања, каде крајна цел е создавање систем способен за препознавање на континуиран неограничен говор од различни говорници. Во овој труд е прикажан систем за препознавање на македонскиот говор, кој во оваа фаза, за препознавање користи мал лексикон ограничен на цифрите од декадниот систем. Системот е дизајниран за препознавање на говор од еден говорник при изолиран - дискретен говор. Базиран е врз хибридна структура со комбинирање на моделот на *скриени Маркови вериги (HMM -Hidden Markov Model)* и *вештачки невронски мрежи (ANN - Artificial Neural Networks)* со цел да се искористат добрите карактеристики од двата пристапа.

2. ХИБРИДЕН HMM/ANN СИСТЕМ ЗА ПРЕПОЗНАВАЊЕ НА ГОВОР

Денешните системи за препознавање на говор *(ASR)* се базирани врз принципите на статистичко препознавање на примероци. Основните методи за примена на овие принципи во проблемот на препознавање на говорот првпат се применети од *Baker, Jelinek* од *IBM* во 1970 и оттогаш претрпеле многу мали измени [1], [5].

Моделот на скриени Маркови вериги врши статистичко моделирање на процесот на создавања на говорот за дефинирање на акустичен модел. Моделот на скриени Маркови вериги нуди карактеризација на говорниот сигнал на математички прифатлив начин.

Вештачките невронски мрежи *(ANN)* се карактеризираат со: масивна паралелна структура, толеранција на шум и робустност, можноста за учење од примерите и можноста на моделирање на нелинеарности и др. Затоа тие наоѓаат примена во разни области како што се класификација, оптимизација, пресликување и апроксимирање на функции, вклучувајќи ја и областа на препознавање на говорот.

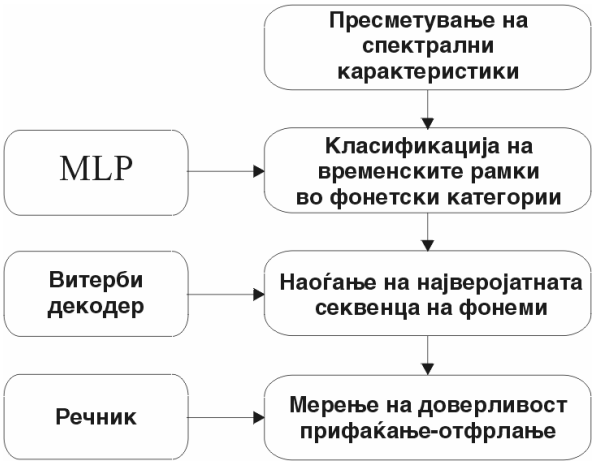
Претходно наведените принципи кои се користат во системите за препознавање на говор може да се искомбинираат во хибриден *ANN/HMM* систем за препознавање. Примената на моделот на скриени Маркови вериги е ограничена од потребата за дефинирање на строги претпоставки кои што не одговараат на стохастичката природа на говорот. Техниките кои користат повеќе нивоовски перцептрон *(MLP)* за проценка на веројатностите за појавување на одредена говорна единица-фонем ги елиминираат овие строги ограничувања [6]. Со примената на *MLP* се постигнува намалување на бројот на параметрите кои се потребни за прецизно моделирање на фонемите, кое произлегува од делењето на параметрите на моделот помеѓу фонетските класи. Пристапот за препознавање на говорот базиран врз примена само на *MLP* не се докажал како погоден за препозна-

вање на континуиран говор поради постоење потреба од прецизна сегментација на акустичниот сигнал, додека од друга страна *HMM* моделот обезбедува симултано сегментирање и класификација на акустичниот сигнал.

3. Систем за препознавање на говор на македонски јазик

Системот е базиран на хибридна *HMM/ANN* архитектура, при што е користен повеќе-нивовски перцептрон со едно скриено ниво. Системот на влезот прима изолирани зборови, а продуцира излез во облик на текстуална транскрипција на препознаениот збор. Системот се обучува со користење на говорен корпус кој содржи инстанци на зборовите кои ги претставуваат првите десет цифри од македонскиот говор. Со тоа е создаден систем со можност на примена секаде каде што постои потреба од внесување цифри (сигурносни системи, воени системи, системи за комуникација, гласовно бирање на телефонски броеви и др.).

Шематски приказ на процесот на препознавање е даден на сликата 1.



Сл. 1.: Процесот на препознавање во систем за препознавање на говор

3.1. Дигитализација

Системот не е врзан директно со опремата за снимање на говорот, туку влезот на системот претставува веќе снимени говорни секвенци во *WAVE*[1] аудио формат. За потребите на системот за препознавање на македонски говор е создаден мал говорен корпус. Секвенците се снимени и дигитализирани со стапка на семплирање од 22.05KHz и квантизирани со 16 бита прецизност со што е добиен сигнал со одличен квалитет. Врз снимениот сигнал е извршена нормализација и редукција на шумот поради изедначување на сигналите кои се користат во обучувачкото множество и говорните секвенци врз кои треба да се изврши препознавање.

3.2. Параметризација и предпроцесирање

За да се придушат ефектите на нелинеарниот фреквентен одзив на опремата која е користена за аквизиција и аналогна/дигитална конверзија на говорниот сигнал се врши предпроцесирање, односно спектрално обликување со нормализација и пре-емфазис филтрирање на влезниот сигнал. Со тоа се овозможува нагласување на одредени фреквентни компоненти со што сигналот станува перцептуално позначаен. За пре-емфазис филтрирање се користи *FIR* филтер од прв ред со следнава преносна функција:

$$H(z) = 1 + a \cdot z^{-1}, \qquad (1)$$

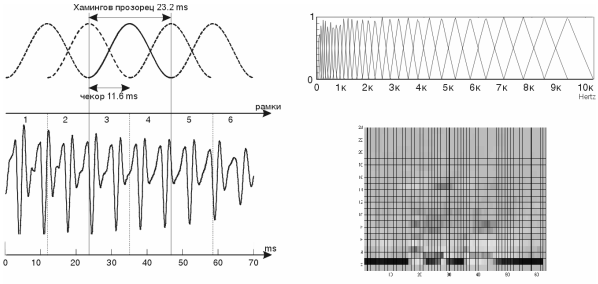
каде вредноста на a се зема дека е 0.9.

Вака филтрираниот сигнал се дели на кратки временски рамки кои меѓусебно се преклопуваат. Должината на рамките се определува од вредноста на фреквенцијата на семплирање која одговара на најголем степен од 2 што одговара на должина на рамка која е пократка од 30 ms. Според тоа, при фреквенција на одбирање од 22.05 KHz, рамките имаат должина од 512 одбироци, односно во овој случај секоја рамка има траење од точно 23.3 ms. Рамките периодично се повторуваат на половина на нивната должина. Потоа, секоја рамка е времена во Хамингов прозорец за да се постигне измазнување на премините меѓу рамките. Се применува *FFT* трансформација за да се добие магнитудата на спектрумот. Магнитудата на спектрумот е претстава на сигналот во фреквентен домен, но сé уште не е погодна како влез во невронска мрежа. Големата резолуција на фреквенцијата продуцира вектори со голема должина, со што се зголемува бројот на влезните единици на невронската мрежа. Затоа се врши пресликување на векторите во друга покомпактна презентација со помош на трансформација за компресирање. Тоа се постигнува во два чекора: прво се применува филтрирање со банка на филтри, а потоа се врши косинусна трансформација врз добиените вредности од филтрирањето. Банката на филтри се состои од еквидистантни триаголни филтри кои меѓусебно се преклопуваат поставени врз Мел-фреквентна скала.

Системот користи 24 мел-цепстрални коефициенти кои го покриваат опсегот од од *0* до 11.025 KHz т.е. *fs/2*. Првите 12 параметри се

[1] WAVE - Waveform Audio File Format

мел-цепстралните коефициенти, а другите 12 се нивните деривации. Со тоа се добиваат влезни вектори во невронската мрежа со должина од 24 елементи.

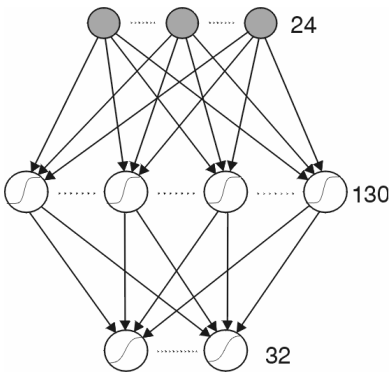


Сл. 2.: Лево: врамување на говорен сигнал со Хамингов прозорец; десно горе: банка на филтри со мел-фреквентна скала и триаголни филтри; десно доле: добиената матрица на акустични вектори по временски рамки;

3.3. Невронска мрежа, архитектура и метод на тренирање

3.3.1. Топологија на невронската мрежа

Невронската мрежа користена во системот за препознавање на македонскиот говор се состои од 24 влезни неврони, бројот на невроните во скриеното ниво изнесува 130, а бројот на излезните неврони изнесува 32 и е еднаков на бројот на фонетските категории влучувајќи ја и категоријата која означува тишина или пауза. Изборот на бројот на невроните во скриеното ниво е извршен емпириски со споредување на перформансите на мрежата во зависност од бројот на единиците во скриеното ниво. Влезното ниво ги проследува влезните вектори какви што се кон второто скриено ниво со неврони кои имаат *tansig* трансфер функција, а излезното ниво се состои од неврони кои имаат *logsig* функција.



Сл. 3.: Топологија на невронска мрежа користена за препознавање на говор

Потребно е да се нагласи дека вредностите кои се појавуваат на излезот од невронската мрежа претставуваат проценки на веројатноста на појавување на фонемите. Тие се множат со вредноста на веројатноста на појавувањето на фонемот која е добиена со статистичка обработка на лексиконот.

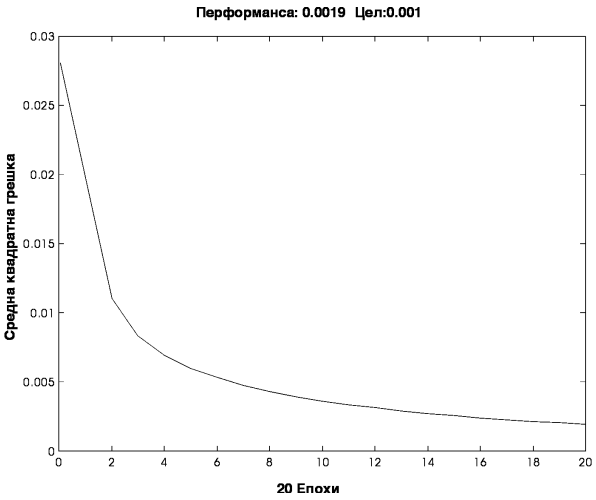
3.3.2. Обучување на невронската мрежа

При обучувањето на невронската мрежа за препознавање цифри се користи мал говорен корпус ограничен на десетте цифри во македонскиот говор. Обучувачкото множество е составено од по три примероци од секој збор кој означува една цифра. Во процесот на обуката постојат два начина на обновување на тежините на врските помеѓу невроните *on-line* и *batch*.

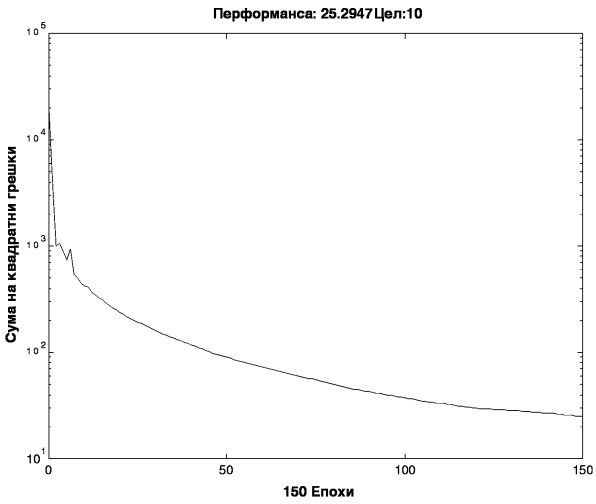
Во првиот случај, обновувањето на тежинските фактори се врши за секој влезен примерок. Критериумот за грешка кој се користи при *Back - Propagation* во овој случај е средна квадратна грешка добиена од сите рамки кои учествуваат во процесот на обуката. За обуката се користат 1457 примероци на фонетските класи кои се содржат во првите десет цифри на македонски јазик. Се покажало дека веќе по 15-25 итерации, односно епохи, средната квадратна грешка значително се намалила до тој степен кога понатаму промената има се помалку влијание врз перформансите на мрежата. Притоа, вредностите за промена на тежините на врските помеѓу невроните и офсетот се добиваат според следниве изрази:

$$\Delta w_{jk} = \frac{\gamma \cdot y_j \cdot (d_k - y_k)}{\sum_i y_i^2}, \quad \Delta \theta_k = \frac{\gamma \cdot (d_k - y_k)}{\sum_i y_i^2} \qquad (2)$$

На сликата 4 е дадена функцијата на грешка која се добива со *on-line* обучување за следниве параметри: максималниот број на итерации - епохи 20, степенот на обучување γ=0.5, целна вредност 0.001 на средната квадратна грешка. Потребно е да се ограничи бројот на епохите потребни за обучување за да се овозможи постигнување на поголем процент на точно препознаени зборови, кои не се во составот на говорниот корпус без да се случи заситување на мрежата. Другиот начин врши обновување на вредностите на тежините и офсетите по изминување на сите примероци кои спаѓаат во обучувачкото множество. На тој начин се намалува времето кое е потребно за извршување на една итерација од причина што за една итерација се врши едно обновување на вредностите.



Сл. 4.: Функција на грешка при on-line обучување на невронска мрежа



Сл. 5.: Процес на batch начин на обучување на невронската мрежа

При обучување на невронските мрежи со сигмоид функции се појавува проблемот со малите вредности на градиентите при *Back - Propagation*, а тоа иницира мали промени на тежинските фактори и офсетите, иако нивните вредности се далеку од оптималните. Затоа се користи *Resilient Back - Propagation* алгоритам за обучување за елиминирање на ефектот на малите вредности на парцијалните деривации. Големината на промената на факторите е определена од вредноста со која се врши обновувањето е еднаква на константен фактор δ_{inc} секогаш кога изводот на функцијата на грешката има ист знак во две последователни итерации.

$$\Delta w_{jk} = \delta_x \cdot sign(gw_{jk}),\ \Delta\theta_k = \delta_x \cdot sign(gw_{jk}) \qquad (3)$$

$\delta_x = \delta_{dec}$, има промена на знакот помеѓу две последователни итерации

$\delta_{ix} = \delta_{inc}$, нема промена на знакот помеѓу две последователни итерации

Алгоритмот се карактеризира со брза и стабилна конвергенција и таа не зависи од иницијалните параметри поставени за процесот на обучување [7]. При обучувањето на мрежата е користен моментум (learning momentum) со вредност 0.98, чија вредност е одредена како таква поради побрза и стабилна конвергенција, δ_o=0.07, δ_{dec}=0.5, δ_{inc}=1.2.

Обучувањето се врши според критериумот сума на квадратни грешки, функцијата на грешка е дадена на следнава слика.

И во двата случаја, *on-line* и *batch* обучување иницијалните вредности на тежинските фактори се поставуваат случајно, затоа во првата итерација се добива огромна грешка, но, потоа драстично паѓа за неколку реда на големина и потоа скоро монотоно опаѓа се додека промената на тежинските фактори се намали под определеното ниво, или се достигне бројот на претпоставените итерации.

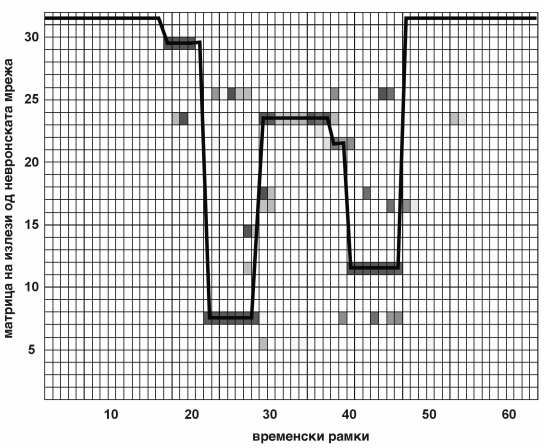
Од добиените резултати може да се воочи дека во првиот случај се потребни околу 20 епохи за да се достигне вредност 0.019 за средна квадратна грешка, додека за вториот случај се потребни околу 150 епохи на обучување, за да се добие средна квадратна грешка со вредност 0.0174. Во системот за препознавање на македонски говор се користи првиот начин на обука, од причина што процесот трае пократко од вториот начин.

Вака обучената невронска мрежа при процесот на препознавање на излезот дава матрица чии колони ги претставуваат временските, а редиците се вредностите добиени како проценка за веројатностите за појавување на одреден фонем во зависност од влезниот акустичен вектор.

3.4. Декодирање и постпроцесирање

Од добиената временска матрица на веројатностите треба да се процени најверојатната секвенца на фонеми. Во системот за препознавање на македонски говор се користи Витерби декодер. Процесот на пребарување со Витерби декодер е илустриран на следниот пример.

чччччеееееетттттттрриииииии



Сл. 6.: Пример за Витерби декодирање

На сликата 6 е прикажана патека која е избрана како патека со најголема веројатност на појавување. Добиеното решение е оптимално решение, но не треба да значи дека секогаш ќе биде и најточно. За примена на Витерби декодер потребно е да постои позната транзициона матрица *A* - матрица на веројатности на премини од една состојба во друга, односно веројатноста на премин од еден одреден фонем во друг фонем. Ова транзициона матрица може директно да се добие од говорниот корпус кој е користен при процесот на обука на невронската мрежа.

Системот е обучен да ги препознава општите карактеристики на одделните фонеми и при препознавањето на зборови кои не учествувале во обучувањето постои голема веројатност дека излезната секвенца ќе содржи и симболи на фонеми кои не се појавуваат во зборот кој се препознава. Примената на горенаведениот метод продуцира низа каде се појавуваат и други симболи со што зборот е делумно препознаен.

*********тт*ееееееввввееееее******

⇓

теве

Од примерот се гледа дека при препознавање на зборот "ДЕВЕТ" е добиен збор кој не се содржи во лексиконот но кој перцептуално е многу сличен на него. Затоа покрај добиената секвенца на фонетски симболи се вклучуваат и дополнителни информации како што се фреквенцијата на последователно појавување на фонетски симбол и фонетската категорија на која и припаѓа фонемот. За наведената секвенца може да се воочи присуство на фонем од категоријата вокали во една рамка помеѓу две низи фонеми од различна фонетска категорија. При споредувањето на секвенцата со зборовите од лексиконот, се појавува и уште еден кандидат покрај зборот "ДВА", а тоа е зборот "ДЕВЕТ", кој исто така има три заеднички фонеми во секвенцата и тоа по ист редослед и постои можност за погрешна одлука. Но, ако се вклучи и информацијата за веројатноста на појавување на фонемите, еднозначно се добива зборот "ДВА", како резултат на препознавањето.

дддддддевввваааааааааааа

девет два

Бидејќи системот е дизајниран за препознавање на изолиран говор се применува паралелна организација на лексиконот, каде што сите зборови последователно се споредуваат со резултантната секвенца и се избира зборот од лексиконот кој има најголема веројатност на појавување во зависност од достапните информации од секвенцата.

4. ЗАКЛУЧОК

Практично е реализиран систем за препознавање на говор на македонски јазик, при што за негова реализација се користи хибридна *HMM/ANN*. Прецизноста на системот при наведените услови изнесува 85% точно препознаени зборови и се покажало дека е оправдано користењето на пристапот базиран врз хибридна архитектура. За да се постигне поголема прецизност, потребно е, пред сé да се создаде говорен корпус на македонски јазик кој би содржел поголем број реченици од поголем број говорници. Со создавање на овој корпус, се отвора можност за користење мрежни топологии со поголем број на процесирачки единици со што се постигнува подобрена генерализација.

5. ЛИТЕРАТУРА

[1] J.K. Baker, "The Dragon System-an Overview", IEEE Trans. ASSP, Vol. 23, p.p. 24-29, 1975.

[2] Tina-Louise Burrows, "Speech Processing with Linear and Neural Network Models",Cambridge University Engineering Department, 1996.

[3] Bourlard & Wellekens, "Links Between Markov Model and Multi-Layers Perceptrons", In Touretzky, D.S. (Ed), Advances in Neural Information Processing Systems, Vol.1, p.p. 502-510. Morgan Kaufman, San Mateo CA. 1989.

[4] P.O. Husoy, "Forvard Connected Artificial Neural Networks applied to Automatic Speech Recognition", Ph.D. thesis, Norvegian Institute of technology, 1991.

[5] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proc. IEEE, Vol. 64, No.4, p.p. 532-556, 1976.

[6] N. Morgan, H. Bourlard, "Continious Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models", ICASSP 90, pp. 414-416, Albuquerque, New Mexico 1990.

[7] M. Riedmiller, H. Braun, "A Direct Adaptive Method For Faster Backpropagation Learning: The RPROP Algorithm," Proceedings of the IEEE International Conference on Neural Networks, 1993.

[8] L.R. Rabiner, B.H. Juang "Fundamentals of Speech Recognition", Prentice Hall Inc., 1993

[9] M.D. Richard, R.P. Lippmann, "Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities", Neural Computation, 3, pp. 461-483, 1991.

[10]

- -

Summary

HYBRID HMM/ANN SYSTEM FOR SPEECH RECOGNITION OF MACEDONIAN LANGUAGE

Ivan Kraljevski[1], Dragan Mihajlov[2] , Dejan Gorgjevik[2]

[1]Veterinary Institute Skopje,
Lazar Pop-Trajkov 5-7, MK-1000 Skopje,Republic of Macedonia
vetinst@unet.com.mk
[2]Faculty of Electrical Engineering, Ss. Cyril and Methodius University,
P.O.Box 574, MK-1001 Skopje, Republic of Macedonia

Key words - artificial speech recognition, neural networks, hidden Markov model

This work presents development of a system for artificial speech recognition of Macedonian language. This system uses small vocabulary, speaker dependent and for isolated speech, limited on recognition of digits. The system is based on hybrid architecture combining the Hidden Markov Model with Artificial Neural Networks in order to exploit its advantages. The system transforms the digitalized speech signal into parameters thus obtaining a sequence of acoustical vectors containing information about spectral characteristics. Acoustical vectors are input of the neural network probability classifier. With Dynamical Programming methods the systems chooses the most probable phonetic categories sequence. Then the system using specific criterion chooses a word from the vocabulary which best mach the phonetic sequence. This system can be used in bank-automates for remote financial transactions, military communications, security systems, mathematical applications using voice-input etc.