Liet. matem. rink., T. 43, spec. nr., 618–624

Neuroninių tinklų panaudojimo atskirai sakomų žodžių atpažinime tyrimas

Mark FILIPOVIČ (MII)
el. paštas: m.filipovic@mail.lt

1. Įvadas

Automatinis šnekos atpažinimas, paprastai apibūdinamas kaip automatinis tariamų žodžių arba sakinių (frazių) vertimas atitinkamu tekstu, yra daugelio mokslininkų ir inžinierių sprendžiamas uždavinys. Visos šiuolaikinės šnekos atpažinimo sistemos yra grindžiamos pavyzdžių atpažinimo (angl. k. – pattern recognition) teorija [1] ir naudoja šios teorijos sukurtus statistinius metodus, kurie remiasi paprastu ir tuo pat metu galingu mokymo iš pavyzdžių principu, kuris šnekos atpažinimui sėkmingai buvo taikomas jau nuo maždaug 1975-tųjų metų.

Pagal aukščiau minėtą principą baigtinis pavyzdžių (mokymo duomenų), t.y., skaitmeninių šnekos signalų įrašų, išreikštų atitinkamomis požymių vektorių (stebėjimų) sekomis, skaičius kartu su žinomomis ir teisingomis jų transkripcijomis yra naudojami tam tikro modelio $\mathbf{M}$ apmokymui. Vėliau šis modelis gali būti naudojamas naujų, prieš tai „nematytų“ įrašų (testinių duomenų) atpažinimui. Modelis dažniausiai apibūdinamas tam tikra struktūra ir parametrais, kurie mokymo metu yra derinami taip, kad maksimizuotų tam tikrą iš anksto apibrėžtą optimalumo kriterijų. Šnekos atpažinimo atveju šis kriterijus dažniausiai yra santykinis žodžių atpažinimo tikslumas.

Beveik visose šiuolaikinėse, tapusiomis standartu šnekos atpažinimo sistemose modelis $\mathbf{M}$ yra paslėptas Markovo modelis (PMM) [2], modeliuojantis stebėjimų vektorių seką. Požymių vektorių modeliavimui esant tam tikram šnekos garsui (fonemai, žodžiui) dažniausiai naudojami standartiniai Gauso mišinio tankiai, kurie žymiai supaprastina modelio mokymą ir naudojimą, bei sudaro galimybę jo mokyme naudoti kelių šimtų valandų trukmės šnekos duomenis, būtinus efektyviems modeliams sukurti.

Nuo maždaug 1980-tųjų metų naudojant mokymo iš pavyzdžių principą dirbtiniai neuroniniai tinklai [3] sėkmingai buvo taikomi statistiniams (tame tarpe ir šnekos) pavyzdžių atpažinimo uždaviniams spręsti. Sparčiame neuroninių tinklų teorijos vystymesi atsirado didelė jų skirtingų struktūrų, modelių ir mokymo algoritmų įvairovė. Čia galima paminėti šiame straipsnyje nagrinėjamus vienus iš dažniausiai naudojamų atpažinimo ir klasifikavimo uždaviniams spręsti vadinamuosius tiesioginio sklidimo neuroninius tinklus (angl. k. – feedforward neural networks) [3, 4, 5].

Atskirai sakomų (izoliuotų) žodžių atpažinimo sistemose, kuriose naudojami PMM, dažniausiai kiekvienam žodžiui (išreikštam požymių vektoriais) yra sudaromas ir apmokomas (įvertinamas) atskiras modelis. Atpažinimo metu daroma prielaida, kad nežinomą

(pasakytą) žodį atstovaujanti tiriamų šnekos požymių vektorių seka yra gaunama iš PMM. Skaičiuojamos to žodžio atitikimo kiekvienam modeliui tikimybės, ir labiausiai tikėtinas modelis identifikuoja žodį.

Skirtingai nuo aukščiau aprašytos sistemos, šiame straipsnyje pateikiama penkiasdešimties atskirai sakomų (izoliuotų) lietuvių kalbos žodžių atpažinimo sistema, kurioje yra naudojami tiesioginio sklidimo neuroniniai tinklai apmokomi pagal vadinamąjį klaidos atbulinio sklidimo (angl. k. – error back-propagation) algoritmą [3, 4, 5]. Nagrinėjamas priklausomas nuo kalbėtojo šnekos atpažinimas. Pristatomas sukurtos atpažinimo sistemos parametrų ir jų įtakos žodžių atpažinimo tikslumui tyrimas.

2. Atskirai sakomų žodžių atpažinimo sistemos aprašymas

Šnekos atpažinimo tyrimams atlikti, buvo sumodeliuota 50-ties atskirai sakomų žodžių atpažinimo sistema, susidedanti iš 50-ties tiesioginio sklidimo neuroninių tinklų (NT), kur kiekvienas tinklas atitiko vieną žodį, paimtą iš apibrėžto 50-ties skirtingų žodžių žodyno. Visi tinklai turėjo vienodą 1 pav. pavaizduotą struktūrą. Kaip matome, kiekvienas tinklas turėjo d įėjimų $x_1, x_2, \ldots, x_d$, atitinkančių įėjimo požymių vektoriaus komponentes, H paslėpto sluoksnio neuronų ir du išėjimus z_1 ir z_2. Tinklo paslėpto ir išėjimo sluoksnio neuronuose buvo panaudota loginio sigmoido (angl. k. – log-sigmoid) aktyvacijos (perdavimo) funkcija

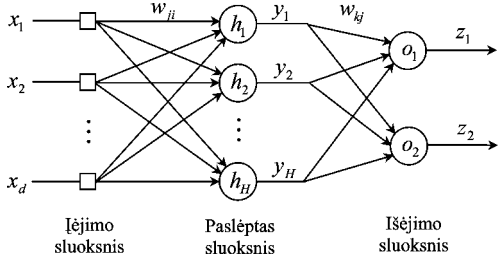
$$y = f(v) = \frac{1}{1 + \mathrm{e}^{-v}},$$

kur y – neurono išėjimas, v – neurono sužadinimo reikšmė

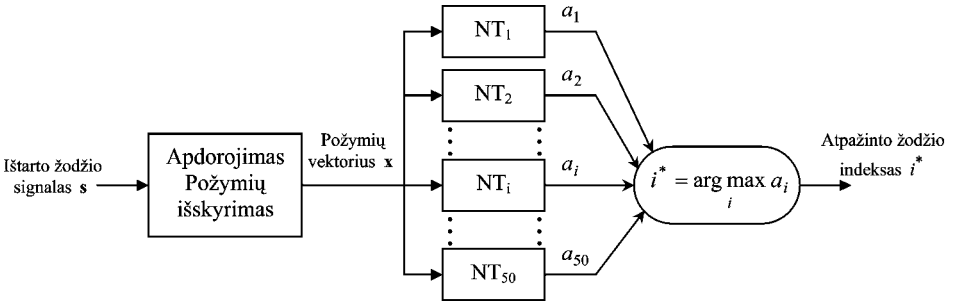
$$v = \sum_{i=0}^{d} w_i x_i,$$

o $x_0 = 1,\ x_1, x_2, \ldots, x_d$ – neurono įėjimai, $w_0, w_1, \ldots, w_d$ – neurono atitinkamų įėjimų jungčių svoriai. Tokiu būdu sukonstruotų tinklų jungčių svoriai (1 pav. pažymėti simboliais w_{ji} ir w_{kj}) sudarė sistemos apmokymo metu derinamus parametrus.

Atliekant sistemos apmokymą, kiekvienas tinklas buvo mokomas atpažinti su juo susijusį žodį. Kalbant tiksliau, jis buvo mokomas atskirti su juo susijusį žodį nuo visų kitų



1 pav. Tiesioginio sklidimo neuroninio tinklo (NT) struktūra.



2 pav. Žodžių atpažinimo sistemos schema.

žodžių. Pavyzdžiui, jeigu norime apmokyti tinklą atpažinti (atskirti) i-tąjį žodyno žodį, išreikštą požymių vektoriumi $\mathbf{x}$, tinklo svoriai, vektoriui $\mathbf{x}$ esant tinklo įėjime, yra derinami taip, kad jo išėjimas z_1 generuotų kuo artimesnę vienetui reikšmę, o išėjimas z_2 – nuliui. Jeigu tinklo įėjime yra kito žodžio požymių vektorius, jo svoriai yra derinami taip, kad išėjimas z_1 generuotų kuo artimesnę nuliui reikšmę, o išėjimas z_2 – vienetui. Tokiu būdu tinklas tarsi buvo mokomas išėjime z_1 aproksimuoti įėjime esančio žodžio tikimybę.

Sistemos darbas žodžių atpažinimo metu pavaizduotas 2 pav. Matome, kad pirmiausiai ištarto žodžio signalas $\mathbf{s}$ patekdavo į apdorojimo ir požymių išskyrimo bloką, po kurio buvo gaunamas šio signalo požymių vektorius $\mathbf{x}$ (apie požymių išskyrimą žiūr. 4 sk.). Po to gautas vektorius buvo paduodamas į visų (50-ties) apmokytų tinklų įėjimus. Kiekvienas tinklas savo ruožtu generavo dviejų išėjimų reikšmes, tačiau žodžio atpažinimui buvo naudojama tik pirmojo išėjimo reikšmė. Galiausiai žodžio atpažinimas buvo atliekamas pagal didžiausią gautą i-tojo tinklo pirmojo išėjimo reikšmės a_i indeksą.

3. Sistemos apmokymas

Aukščiau aprašytos žodžių atpažinimo sistemos apmokymas susiveda į atskirų joje panaudotų neuroninių tinklų apmokymą. Čia galima priminti, kad i-tas sistemos tinklas buvo mokomas atskirti su juo susijusį i-tąjį žodį nuo visų kitų apibrėžto žodyno žodžių. Žemiau pateikiamas vieno tokio tinklo apmokymo procedūros aprašymas.

Tam, kad neuroninis tinklas galėtų atlikti žodžių, išreikštų požymių vektoriais, atpažinimą (klasifikavimą), prieš tai reikia jį tinkamai apmokyti. Tinklas apmokomas naudojant mokymo imtį. Mokyme su mokytoju mokymo imtis sudaroma iš požymių vektorių kartu su atitinkamais trokštamų išėjimų vektoriais. Trokštamo išėjimo vektorius nurodo požymių vektoriaus atitikimą vienai iš apibrėžtų klasių. Pavyzdžiui, jeigu i-tojo žodžio požymių vektorius $\mathbf{x}_i$ priklauso j-tajai klasei, tai trokštamo išėjimo vektoriaus $\mathbf{t}_i$ j-tajai komponentei priskiriamas vienetas, visoms kitoms – nuliai. Mokymo metu, naudojant mokymo algoritmą, tinklas mokosi (nežinomos) priklausomybės tarp požymių ir atitinkamų trokštamų išėjimų vektorių. Jei tinklas tinkamai apmokytas, jis gali modeliuoti (nežinomą) funkciją siejančią požymių vektorius su atitinkamais trokštamų išėjimų vektoriais ir tokiu būdu gali atlikti naujų (mokyme nenaudotų) požymių vektorių atpažinimą.

Tegul turime N skirtingų žodžių, ištartų po K kartų, požymių vektorių imtį

$$P = \{\mathbf{x}_i^k\}, \quad i = 1, 2, \ldots N, \;\; k = 1, 2, \ldots, K,$$

kur $\mathbf{x}_i^k$ – i-tojo žodyno žodžio k-tojo ištarimo požymių vektorius. Tada i-tojo sistemos tinklo mokymo imtis

$$M_i = \{P_i, T_i\},$$

kur imtis P_i sudaryta iš visų galimų požymių vektorių porų, kurių pirmas vektorius yra i-tojo žodžio k-tojo ištarimo požymių vektorius, o antras – kito (j-tojo) žodžio k-tojo ištarimo požymių vektorius, t.y.,

$$P_i = \{(\mathbf{x}_i^k, \mathbf{x}_j^k)\}, \quad j \neq i, \;\; j = 1, 2, \ldots, N, \;\; k = 1, 2, \ldots, K,$$

imtis T_i sudaryta iš atitinkamų trokštamų išėjimų vektorių porų, t.y.,

$$T_i = \{(\mathbf{t}_i^k, \mathbf{t}_j^k)\}, \quad j \neq i, \;\; j = 1, 2, \ldots, N, \;\; k = 1, 2, \ldots, K,$$

kur vektorius $\mathbf{t}_i^k = [1; 0]$ nurodo požymių vektoriaus $\mathbf{x}_i^k$ atitikimą pirmai klasei, o vektorius $\mathbf{t}_j^k = [0; 1]$ – vektoriaus $\mathbf{x}_j^k$ atitikimą antrai klasei.

Naudojant suformuotą mokymo imtį M_i, i-tas sistemos tinklas buvo mokomas pagal Levenberg–Marquardt klaidos atbulinio sklidimo algoritmą [5]. Tinklo apmokymo tikslumo įvertinimui buvo naudojama vidutinė kvadratinė klaida, apibrėžiama pagal formulę:

$$J = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (t_{ij} - z_{ij})^2,$$

čia N – mokymo imties požymių vektorių skaičius, M – tinklo išėjimų skaičius, z_{ij} – tinklo j-tojo išėjimo reikšmė (įėjime esant požymių vektoriui $\mathbf{x}_i$), t_{ij} – atitinkama trokštamo išėjimo reikšmė (vektoriaus $\mathbf{t}^i$ j-toji komponentė). Mokymo algoritmo veikimas buvo stabdomas, kai $J < \theta$, t.y., kai J reikšmė tapdavo mažesnė už kokią nors iš anksto parinktą nedidelę reikšmę θ (apmokymo tikslumo reikšmę).

Visi klaidos atbulinio sklidimo algoritmo variantai grindžiami gradientinio nusileidimo (gradient descent) metodu [3, 4, 5], kur tinklo svoriams $\mathbf{w}$ pirmiausiai priskiriamos nedidelės, dažniausiai atsitiktinės reikšmės, po to svoriai yra koreguojami pakeičiant juos mažu žingsniu ta kryptimi, kuria klaida (esant n-tajam požymių vektoriui)

$$J^n(\mathbf{w}) = \frac{1}{2} \sum_{j=1}^{M} (t_{nj} - z_{nj})^2$$

mažėja greičiausiai. Didžiausio mažėjimo kryptis – neigiamas gradientas

$$\Delta\mathbf{w} = -\eta \frac{\partial J^n(\mathbf{w})}{\partial \mathbf{w}},$$

kur η yra mažas teigiamas skaičius, vadinamas mokymo greičiu. Naujų, (vis geresnių) svorių vektorių seka gaunama pagal iteratyvų algoritmą, kur n-toje iteracijoje (t.y., esant n-tajam požymių vektoriui) imamas svorių vektorius $\mathbf{w}(n)$ ir keičiamas pagal formulę:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \Delta\mathbf{w}(n).$$

Esant tam tikroms sąlygoms svorių vektorius konverguos į tašką, kuriame klaida yra minimali.

4. Požymių išskyrimo procedūra

Sistemoje (2 pav.) ištarto žodžio signalas $\mathbf{s}$ buvo išreiškiamas požymių vektoriumi $\mathbf{x}$. Tam buvo naudojami melų dažnių skalės kepstriniai požymiai (MFCC) [6]. Žemiau pateikiamas požymių vektorių gavimo procedūros aprašymas.

Pirmiausiai kalbos signalas buvo skaidomas į kadrus (segmentus), kurių ilgis 25 ms, o atstumas tarp gretimų kadrų, vadinamasis kadrų žingsnis, 10 ms. Vėliau vykdomas signalo apdorojimas, kuriame kiekvienam signalo kadrui buvo atliekamas filtravimas

$$\tilde{s}(n) = s(n) - 0,97s(n-1), \quad n = 0, 1, \ldots, N-1,$$

(N – signalo kadro atskaitymų skaičius) ir padauginimas iš Hammingo lango funkcijos [2].

Toliau pagal [6] knygoje pateiktus algoritmus iš signalo kadrų buvo skaičiuojami MFCC požymiai (po 12 požymių kiekvienam kadrui) ir taip gaunami požymių vektoriai $\mathbf{p}_k = [p_{k1}, p_{k2}, \ldots, p_{k12}]$, kur $k = 1, 2, \ldots, K$ (K – kadrų skaičius).

Požymių vektorių $\mathbf{p}_k$ negalima buvo tiesiogiai naudoti neuroninio tinklo mokymui, nes jų skaičius kiekvienam signalui (ištartam žodžiui) buvo skirtingas ir pernelyg didelis, todėl naudojant k-vidurkių klasterizavimo algoritmą [1], jie buvo klasterizuoti į 5 klasterius. Klasterizavimo algoritmas buvo inicializuojamas pradiniais centrais $\mathbf{m}_i = [m_{i1}, m_{i2}, \ldots, m_{i12}]$, paskaičiuotais iš $\mathbf{p}_k$ pagal formulę:

$$m_{ij} = \frac{1}{L} \sum_{k=(i-1)L+1}^{iL} p_{kj} \quad (j = 1, 2, \ldots, 12),$$

kur $i = 1, 2, \ldots, 5$, $L = K/5$ (K – požymių vektorių skaičius).

Po klasterizavimo algoritmo gautų klasterių centrai $\mathbf{c}_i = [c_{i1}, c_{i2}, \ldots c_{i12}]$ ($i = 1, 2, \ldots, 5$) buvo apjungti į vieną vektorių $\mathbf{x} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_5]$, kuris ir buvo naudojamas kaip ištarto žodžio požymių vektorius.

5. Eksperimentai

Eksperimentinio tyrimo tikslas buvo išnagrinėti pateiktos 50-ties atskirai tariamų (izoliuotų) žodžių atpažinimo sistemos (2 pav.), naudojančios tiesioginio sklidimo neuroninius tinklus, parametrų įtaką žodžių atpažinimo rezultatams.

Kiekvienas sistemos tinklas (1 pav.) turėjo H paslėpto sluoksnių neuronų. Tinklo inicializavimo metu jo svoriams buvo priskiriama nedidelė fiksuota inicializavimo reikšmė ω ir apmokymo tikslumo reikšmė θ.

Šnekos atpažinimo tyrimams atlikti buvo naudojamas atskirai tariamų žodžių lietuvių šnekos garsynas, apimantis vieno kalbėtojo (vyro) šnekos įrašus. Kalbėtojas buvo pateikęs 50 įrašų, kur kiekvieną įrašą sudarė 50 atskirai tariamų žodžių. Tokiu būdu garsyne iš viso buvo 2500 ištartų žodžių.

Atpažinimo sistema buvo realizuota MATLAB programavimo kalba. Melų dažnių skalės požymių išskyrimui buvo panaudotas HTK 3.2 (`http://htk.eng.cam.ac.uk`) programų paketas.

Eksperimentinio tyrimo metu buvo nagrinėjami tokie sistemos parametrai, kaip tinklų paslėpto sluoksnio neuronų skaičius H, svorių inicializavimo reikšmė ω ir apmokymo tikslumo reikšmė θ. Parametras H buvo keičiamas nuo 1 iki 10. Parametrai ω ir θ buvo keičiami nuo 0,1 iki 0,00001, kiekvieną kartą mažinant juos 10 kartų. Eksperimentai buvo atlikti su visais galimais šių parametrų deriniais. Sistemos apmokymui buvo panaudota 15 garsyno įrašų, o jos testavimui – likusieji 35 įrašai. Tokiu būdu testavimui iš viso buvo pateikta $35 \times 50 = 1750$ žodžių. Kiekvieno eksperimento metu sistemos atpažinimo tikslumas AT buvo vertinamas pagal formulę:

$$AT = \frac{T}{N} \cdot 100\%,$$

kur T – teisingai atpažintų žodžių skaičius, N – visų testinių žodžių skaičius (mūsų atveju $N = 1750$).

Atlikto eksperimentinio tyrimo rezultatai pateikiami 1 lentelėje. Kaip matome, didžiausias gautas žodžių atpažinimo tikslumas (DAT) yra lygus 95,2%, kai $H = 6$, $\omega = 0,0001$ ir $\theta = 0,1$. Iš 1a lentelės galime pastebėti, kad nepriklausomai nuo H parametro geriausi atpažinimo rezultatai gauti, kai $\omega = 0,0001$ ir $\theta = 0,1$ ir kad parametrui H didėjant nuo 1 iki 6, atpažinimo tikslumas irgi nežymiai didėja, bet parametrui H didėjant nuo 7 iki 10, atpažinimo tikslumas turi nedidelę tendenciją mažėti.

DAT priklausomybė nuo tinklų apmokymo tikslumo θ pateikiama 1b lentelėje. Iš jos galime pastebėti, kad mažėjant parametrui θ, DAT irgi mažėja. Tai galima paaiškinti tuo, kad mažinant θ, tinklų apmokymo tikslumas didėja, dėl to mokymo proceso metu jie pradeda pernelyg prisitaikyti prie mokymo duomenų ir įgauna mažesnį gebėjimą atpažinti (klasifikuoti) testinius (mokyme ne naudotus) duomenis.

1c lentelėje pateikiama DAT priklausomybė nuo tinklų svorių inicializavimo reikšmės ω. Iš lentelės galime pastebėti, kad parametras ω daro didžiausią įtaką žodžių atpažinimo tikslumui. Matome, kad geriausias žodžių atpažinimo rezultatas gautas, kai $\omega = 0,0001$. Kai buvo naudojamos kitos ω reikšmės, gauti blogesni atpažinimo rezultatai.

1 lentelė. Sistemos žodžių atpažinimo tikslumo priklausomybė nuo tinklų parametrų a) paslėpto sluoksnio neuronų skaičiaus H, b) apmokymo tikslumo θ, c) svorių inicializavimo reikšmės ω; DAT – didžiausias gautas žodžių atpažinimo tikslumas, išreikštas procentais.

a)

H	DAT (%)	ω	θ
1	93,94	0,0001	0,1
2	94,29	0,0001	0,1
3	94,69	0,0001	0,1
4	94,74	0,0001	0,1
5	94,40	0,0001	0,1
6	**95,20**	**0,0001**	**0,1**
7	94,97	0,0001	0,1
8	94,97	0,0001	0,1
9	95,03	0,0001	0,1
10	94,91	0,0001	0,1

b)

θ	DAT (%)	H	ω
0,1	**95,20**	**6**	**0,0001**
0,01	94,34	5	0,0001
0,001	94,06	10	0,0001
0,0001	94,06	5	0,0001
0,00001	94,00	7	0,0001
0,000001	93,94	7	0,0001

c)

ω	DAT (%)	H	θ
0,1	89,77	3	0,000001
0,01	93,94	10	0,00001
0,001	93,66	2	0,0001
0,0001	**95,20**	**6**	**0,1**
0,00001	93,60	10	0,0001
0,000001	91,66	6	0,000001

6. Išvados

Straipsnyje buvo pateikta 50-ties atskirai sakomų (izoliuotų) lietuvių kalbos žodžių atpažinimo sistema, naudojanti tiesioginio sklidimo neuroninius tinklus. Tinklų apmokymas buvo atliekamas pagal klaidos atbulinio sklidimo algoritmą. Nagrinėjamas priklausomas nuo kalbėtojo šnekos atpažinimas. Atliktas sukurtos atpažinimo sistemos parametrų ir jų įtakos žodžių atpažinimo tikslumui tyrimas. Eksperimentinio tyrimo rezultatai leidžia teigti, kad sistema gali būti naudojama atskirai sakomų žodžių atpažinimui.

Šiame darbe šnekos atpažinimo vienetu buvo laikomas žodis. Tęsiant tyrimus, reikėtų bandyti pereiti prie šnekos atpažinimo garsų (skiemenų, fonemų) lygyje ir modeliuoti didesnio žodyno bei daugiau kalbėtojų apimančias šnekos atpažinimo sistemas.

Literatūra

[1] R. Duda, P. Hart, *Pattern Classification and Scene Analysis*, Wiley (1973).
[2] L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall (1993).
[3] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall (1999).
[4] C.M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford (1995).
[5] M. Hagan, H. Demuth, M. Beale, *Neural Network Design*, PWS Publishing (1996).
[6] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department (2002). `http://htk.eng.cam.ac.uk/docs/docs.shtml`.

Isolated word recognition using neural networks

M. Filipovič

This paper presents the speech recognition system, designed for recognition of 50 isolated Lithuanian words. The system is based on feedforward neural networks trained by error back-propagation algorithm. Speaker dependent speech recognition was investigated. The article presents analysis of the developed system parameters influence on the accuracy of word recognition. Experimental results showed that proposed system could be used for isolated word recognition.