Automatikus beszédfelismerés a SPICOS rendszerben

HERMANN NEY
Philips GmbH Kutató Laboratórium, Hamburg

HERMANN NEY

Fizikából szerzett diplomát 1977-ben a Göttingeni Egyetemen (NSZK) és elektromérnöki doktori fokozatot 1982-ben a Braunschweigi Műszaki Egyetemen. 1977 óta a Philips Kutató Laboratóriumában dolgozik Hamburgban telefonos beszélőfelismerés, digitális jelfeldolgozás, valamint szó- és beszédfelismerés területén. Tevékenysége olyan adott környezetbeli döntéshozatal alkalmazására koncentrálódik, mint például a nemlineáris időnormálás és nemlineáris simítás. Különösen érdeklődik a jelfeldolgozás és alakfelismerés matematikai módszereinek alkalmazása iránt. Jelenleg a folyamatos beszéd felismerése kutatási témák felelőse.

ÖSSZEFOGLALÁS:

Ebben a cikkben a **SPICOS** rendszerben használt technika kerül bemutatásra. Egy olyan integrált megközelítés az alap, ahol különböző ismeretforrások, mint részszómodellek tára, kiejtési lexikon és nyelvmodell, kombinálódnak a döntési folyamatban az akusztikai felismerés javítása érdekében. A felismerési döntés egy nagy állapot-térben való keresést jelent késleltetett döntésekkel. 5 beszélővel mentek végbe a kísérleti tesztek. Beszélőnként 376 mondatot teszteltek, ami 2584 szó felismerési tesztjét jelenti.

1. Bevezetés

A **SPICOS** szó a *'Siemens-Philips-Ipo COntinous Speech recognition and understanding'* (Siemens-Philips-*I*po folyamatos beszéd felismerése és megértése) helyett áll. A projekt célja egy olyan ember-gép párbeszédes rendszer, amely képes folyamatosan kimondott német mondatokat megérteni és így hanggal történő hozzáférést tesz lehetővé adatbázisokhoz.

A választott adatbázis információkat tartalmaz a **SPICOS** projektről magáról. Személyi dokumentációs rendszernek lehet tekinteni a belső tevékenységekről és a **SPICOS** projektbe bevont kutatócsoportok közötti kommunikációról. 200 mondatot választottak ki a rendszer által kezelendő kérdések lefedésére. Hogy egyszerűsítsük a nyelvészeti szerkezeteket, néhány megszorítást vezettünk be: nincs vonatkozói mellékmondat, szenvedő szerkezet, névmások. A szótár 917 teljes szóalakot tartalmaz, amelyek 420 szótőből származnak.

Ebben a cikkben egy, a kimondott szavak akusztikai felismerésére szolgáló módszerről van szó. Úgy sorolható be, mint integrált megközelítés és kísérlet az input beszédadatok legjobb értelmezésére olyan ismeretforrásokat használva, mint nyelvmodell, kiejtési lexikon és részszó egységek tára.

2. Statisztikai döntéselmélet

Folyamatos beszéd felismerése bonyolultabb feladat, mint izolált szavaké több tekintetből: a szókincs nagyobb, a szavakat kevésbé tisztán ejtik, és a felismerő rendszernek figyelembe kell vennie a mondatok szintaktikai-szemantikai szerkezetót. A következmény az, hogy a felismerő folyamatnak egy nagyobb keresési térben kell mozognia, azaz nagyobb a szavakhoz és mondatokhoz alkotott hipotézisek száma, és magasszintű szintaktikai és szemantikai megszorításokkal kell számolnia.

A rendszerfelépítések szisztematikus osztályozására hasznos kritérium az a mód, ahogyan a hipotézisek kétértelműségeit kezelik és a kereső eljárás szervezése [Lea, 1980; Haton, 1982; de Mori, Suen, 1985]. Gyakran a rendszerfelépítést aszerint osztályozzák, hogy a hipotéziseket alulról-fölfelé (bottom-up) vagy fölülről-lefelé (top-down) hozzák létre. Az attributumok eredetileg egy nem-probabilisztikus környezetfüggetlen nyelvtan elemzéséből származnak [Hopcroft, Ullman, 1979].

Az alulról-fölfelé megközelítésben a részszó és szó szintű felismerés a magasszintű komponensektől függetlenül működik, amikor részszó vagy szó hipotéziseket hoz létre. A felülről-lefelé megközelítésben a magasszintű komponensek előre jósolnak bizonyos részszavakat vagy szavakat a már eddig feldolgozott mondatok, valamint szintaktikai, szemantikai és pragmatikai ismeretek alapján, majd pedig továbbadják ezeket az akusztikai szintnek a hasonlóság kiértékelésére.

A statisztikai döntéselmélet szempontjából a probléma a következőképpen néz ki. Azért, hogy minimalizáljuk a hiba valószínűségét, Bayes döntési szabálya alapján [Fukunaga, 1972], a beszédfelismerő feladata meghatározni azt a $w(1),.\ w(n),\ldots w(N) := w[1:N]$ (nem ismert N hosszúságú) szósorozatot, amely a legvalószínűbben előidézte az $x(1),\ldots,\ x(i),\ldots,\ x(I) := x[1:I]$ méréssorozatot. A Bayes tételt felhasználva a következő formára alakíthatjuk ezt [Jelinek, 1976]: Meghatározandó az a $w[1:N]$ szósorozat, amely maximálja a

$$Pr(w[1:N])*Pr(x[1:I]|w[1:N])\text{-t.}$$

Ez a fontos egyenlet megvilágítja a megfigyelt adatok és a rendszer ismeretforrásai közötti kölcsönhatást: a döntés a megfigyelt adatok és a

Fordította: Koutny Ilona
Elhangzott az 1987. máj 6—7-ón tartott VDE konferencián.

rendszer ismeretforrásai közötti legjobb kompromisszumra kell hogy vezessen. Az egyenlet egyidejűleg lehetővé teszi, hogy világosan definiáljuk a határt az akusztikai-fonetikai és magasabbszintű ismeretforrások között.

Az első tag $Pr(\mathrm{w}[\mathrm{i}:N])$ a w[1:N] szósorozat a-priori valószínűsége. Független az akusztikai megfigyelésektől és a magasabbszintű ismeretforrások határozzák meg egyértelműen. Más szavakkal a magasabbszintű ismeretforrások, mint a szintaxis, a szemantika és pragmatika, ekvivalensek minden w[i:N] szósorozat $Pr(\mathrm{w}[\mathrm{i}:N])$ a-priori valószínűségének a tudásával. Ezeket az ismeretforrásokat rendszerint nyelvmodellnek hívják.

A második tag $Pr(x[\mathrm{i}:I]\ \mathrm{w}[1:N])$ az x[1:I] sorozat megfigyelésének feltételes valószínűsége, ha a w[i:N] szósorozat lett kiejtve. Ennek tükröznie kell az akusztikai-fonetikai és lexikai ismeretforrásokat. Amennyiben a nyelv és az akusztikai fonetika valószínűségfüggvényei adottak, elvileg lehetséges kiértékelni minden egyes w[1:N] szósorozat esetében a $Pr(\mathrm{w}[1:N])*Pr(\mathrm{w}[1:N]|x[1:I])$-t és meghatározni közvetlenül a legvalószínűbb szósorozatot.

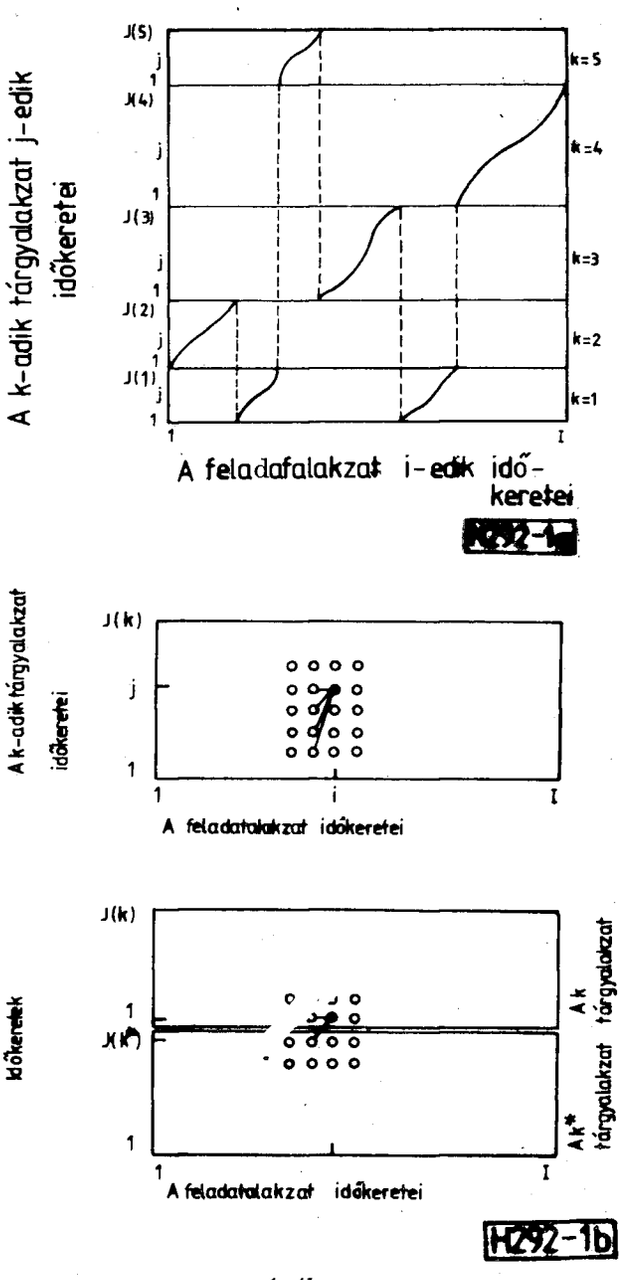
A statisztikai döntéselmélet szempontjából a megkülönböztetés top-down és bottom-up megközelítés szerint irreleváns. A kereső eljárás a lényeges a leghasonlóbb szósorozat megtalálására. A továbbiakban olyan modelleket és megközelítéseket tárgyalunk, melyek kiszámolják a nyelvi modellek feltételes valószínűségeit és a-priori valószínűségeit, valamint elvégzik az optimalizálást.

3. Egylépéses algoritmus

Ebben a részben egy dinamikus programozási algoritmus kerül bemutatásra, mely zárt alakban kínálja a megoldást nemlineáris időnormálás, szóhatár-meghatározás és szóazonosítás összefüggéseinek a kezelésére folyamatos beszéd felismerésében [Ney, 1984; Bridle, Brown, Chamberlain, 1982; Vintsyuk, 1971]. Ez az algoritmus lehetővé teszi, hogy a 2. részben bemutatott feltételes valószínűséget kiszámoljuk.

Az ismeretlen bemenő vagy tesztminta $i = 1, \ldots, I$ időablakból áll, mindegyiket egy $x(i)$ akusztikai vektor reprezentál. Azt tudjuk, hogy a bemenő minta egyedi szavakból áll, melyeket az adott szótárból választottunk. A szótár szavai megfelelnek egy K referenciamintából álló halmaznak. A szómintákat k-val indexeljük, $k = 1, \ldots K$. Minden k mintát egy véges állapotú gép modellál [Baker, 1975a, b; Jelinek, 1976], melynek $j = 1, \ldots, J(k)$ állapota lehet. Az alapötletet az 1. ábra illusztrálja.

A tesztminta i időablaka és minden k referenciaminta j állapota egy (i, j, k) rácsponthalmazt definiál. Mindegyik (i, j, k) rácsponthoz létezik egy $d(x(i)|j, k)$ lokális távolságmérték, amely az i-edik tesztablak $x(i)$ vektorának a lokális távolsága a k-adik referenciaminta j-edik állapotának a referenciavektorától. Valószínűségi keretben ez a kibocsátási valószínűség sűrűségfüggvényének a



1. ábra

negatív logaritmusát jelenti. Ezenkívül számolni kell az időtorzítás 'büntetés'-eivel (Penalties), melyeket úgy kell érteni és kiszámolni, mint az átmeneti valószínűségek negatív logaritmusát. A $T(j|j', k)$ azt a büntetést jelenti, amikor a k-adik minta j'-edik állapotából a j-edik állapotába megyünk. A kapcsoltszavas felismerés nem más, mint az (i, j, k) rácspontok halmazában annak az útnak a megtalálása, amely a legjobb illeszkedést szolgáltatja a tesztminta és a referenciaszavak ismeretlen sorozata között (1a ábra).

A végesállapotú modell következtében bizonyos folyamatossági megszorítások vagy átmeneti szabályok érvényesek az (idő, állapot) pontok hálóján keresztül vezető útra. Célszerű különbséget tenni az átmenetek két típusa között, ahogy ezt az 1b ábra mutatja: átmenetek a szómintán belül és átmenetek a szómináták határánál. Hogy a dina-

mikus programozás technikáját alkalmazhassuk [Ney, 1982], definiáljuk, a $D(i, j, k)$ minimális kumulatív távolságot az (i, j, k) rácsponthalmazban bármely útra. Szó belsejében rekurziót kell alkalmazni minden $k=1,\ldots, K$-ra és $j=1,\ldots, J(k)$-ra:

$$D(i, j, k) = d(x(i)|j, k) + \min\{D(i-1, j', k) + T(j|j', k) : j' = 0,\ldots, j\}$$

Hogy ki tudjuk számolni a szavak közötti átmenetet, bevezetünk egy további mesterséges rácspontot, $(i, 0, k)$-t, melyet egy adott i időpontban az összes (i, j, k) rácspont ($k=1,\ldots, K$ és $j=1,\ldots, J(k)$) feldolgozása után értékelünk ki:

$$D(i, 0, k) = \min\{D(i, j(k'), k') : k' = 1,\ldots, K\}$$

Általában a büntetéseket úgy modellálják, hogy csak az átviteli szélességtől függjenek: $T(j|j', k) = T(j-j')$. A rekurzív kiértékelés az összes referenciaszóra egy lépésben megy végbe. A megvalósítás három hurkot igényel; egyet a bemenő ablakoknál, egyet a referenciászavaknál és egyet minden referenciaszó ablakaira. Az utolsó input ablak feldolgozása után úgy jön létre az optimális út, hogy visszamegyünk az egyes helyi optimalizáló lépések optimális döntéseire. Így határozzuk meg, hogy a szósorozat melyik mintával illeszkedik a legjobban.

4. A nyelvmodell megszorításai

A 2. részben bevezetett nyelvmodell olyan komplex felismerési feladathoz kapcsolódik, ahol nincs közvetlen függés a szótár méretével. A továbbiakban avval egyszerűsítjük a nyelvi modellt, hogy feltesszük, hogy minden mondat egyformán valószínű. Egy lehetőség arra, hogy a felismerési feladat komplexitását, más szóval a magasszintű tudásforrások megkövetelte megszorítások fokát mérjük, az ún. elágazási faktor vagy bonyolultság, mely azon szavak átlagos számát adja meg, melyek egy megengedett mondatban valamely szó után következhetnek. Ha minden megengedett mondat egyformán valószínű és a nyelvi bonyolultság p, akkor egy adott n-hosszúságú mondatból összesen p^n különböző van. Ezeket a megszorításokat leírhatja egy véges állapotú háló [Jelinek, 1976]. A hálóban minden út legális mondathoz vezet. Bár a folyamatos beszéd felismerése esetében a szótár 1000 vagy még több szót tartalmaz, egy adott állapotnál a választható szavak átlagos száma jóval kisebb, rendszerint 10 és 100 között van. Ilyen tekintetben a folyamatos beszéd felismerése könnyebb feladat lehet, mint számsorozatok felismerése, ha a bonyolultság kisebb, mint 10. A **SPICOS** feladatra kifejlesztett nyelvmodell [Mergei, 1986] bonyolultsága 58. A háló 368 csomópontból és 3481 átmenetből áll.

A nyelvi megszorításokat a következőképpen lehet beültetni a dinamikus programozás algoritmusába. A végesállapotú hálót úgy írjuk le, hogy minden k szóátmenetre megadjuk a $b(k)$ kezdeti csomópontot és $e(k)$ végpontot:

$b(k)=$ az a csomópont, ahol k kezdődik
$e(k)=$ az a csomópont, ahol k befejeződik.

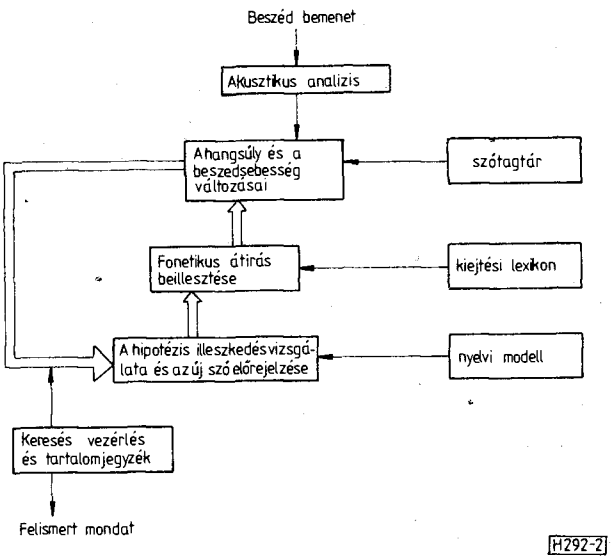
Általában ugyanannak az akusztikai szónak több példányára van szükség, hogy lefedhessük azt a változatos szintaktikai környezetet, ahol a szó előfordulhat. A dinamikus programozás rekurziója ugyanaz marad a szó belsejében, mint azt a 3. részben leírtuk. Miután egy adott i időpillanatra az összes (i, j, k) rácspont fel lett dolgozva, ki kell számolni a legjobb, szintaktikai csomóponthoz vezető utat, ezután lehet kezdeni egy k újszóátmenetet, melyet a következő egyenlet fejez ki:

$$D(i, 0, k) = \min\{D(i, J(k'), k') : k', \text{ ahol } e(k') = b(k)\}$$

5. Az integrált megközelítés kereső eljárása

A dinamikus programozási egyenletek alapvető problémája az, hogy az ismeretforrások által definiált állapottérben a teljes keresés elriasztó. A nyelvi modellnek két ellentétes hatása van. Egyrészt a szóátmenetek száma és hogy az egész kereső tér nagyobb, mint nyelvi megszorítások nélkül. Másrészt a nyelvmodell szigorú megszorításokat kényszerít a 'legális' szósorozatokra. Ezért olyan kereső algoritmust alkalmazunk, amely megpróbálja hatékonyan kihasználni a nyelvi megszorításokat és csak a keresőtér 'releváns' területein kiértékelni a fenti dinamikus programozási egyenletet.

A 2. ábra a kereső eljárás egyfajta megszervezését mutatja be. A három szintű hierarchia a nyelvmodellből, a kiejtési lexikonból és a részszavak tárából áll, melyek véges állapotú gépekként modellálhatók [Boulard et al., 1985; Ney, Mergei, Marcus, 1986]. Ez a három szint elkülönül, csak a kereső eljáráson keresztül tudnak kapcsolatba lépni. Elméletileg a keresés dinamikus programozás, ahol a keresés a leghasonlóbb hipotézisre korlátozódik. Mindezért a nagy, mintegy 200 000 állapotú kereső tér miatt speciális szervezés kell, hogy csökkenteni lehessen a költségeket ós a tároló kapacitást. Mivel a kereső eljárás mindegyik tudásforrást használja, ezt á technikát integrált meg-



2. ábra

közelítésnek hívjuk. Hasonló, globális felismerésre alapozott rendszerek az IBM-rendszer [Jelinek, 1985], a BBN-rendszer [1986], a Vintsyuk által kifejlesztett rendszer [1982]. a Bell laboratórium repülőjegy-rendelő rendszere [Rabiner, Levinson, 1981], a DRAGON-rendszer [Baker, 1975a] ós a HARPY-rendszer [Lowerre, 1976].

6. Kísérleti eredmények

Kísérleti teszteket egy 5 beszélővel létrehozott adatbázissal végeztünk. Minden beszélőnek 1—3 alkalommal be kellett mondani a 200 SPICOS mondatot folyamatosan, azaz szavak közti szünet nélkül. A felismerési teszteket a 200 SPICOS mondatból 188-on hajtottuk végre. 188 mondatra a szófelismerési tesztek száma 1292 volt.

A beszélőfüggő fonémamodellek 200 mondatból származnak, amelyeknek a szókincse nem egyezik meg a SPICOS-éval. Ezek az ún. Sotschek mondatok jellemzőek a német nyelv fonéma eloszlására és összesen 4860 fonémát tartalmaznak. Az átfedés a felismerő szótárral mindössze 51 szó volt, amelyek főként a nyelv nyelvtani szavai voltak, mint névelők és elöljárók.

A szóhiba-arány sokban függ a beszélőtől. A rendszer jelenlegi verziójában a szóhiba-arány 8 és 20% között mozog beszélőtől függően. A szóhibák eloszlása a mondatban nem egyforma, hanem inkább csoportokat alkot. A felismerési kísérletekben az algoritmus által átvizsgált terület az egész lehetséges keresőtérnek csak 2—5%-a, ami tipikusan 50—200 MIPSS-et (= másodpercenkénti millió utasítás) jelent.

Az itt leírt munkát egy kapcsolódó Siemens-Philips-IPO (Eindhoven) projekt keretében végeztük és a Német Szövetségi Kutatási és Technológiai Minisztérium (BMFT) támogatta a 413-5839-ITM 8401 sz. hozzájárulással. Csak a szerző felelős a publikáció tartalmáért.

IRODALOM

[1] *J. K. Baker*, (1975a): „The DRAGON System — An Overview", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 1, pp. 24—29, February 1976.

[2] *J. K. Baker* (1975b): „Stochastic Modeling for Automatic Speech Understanding", in D. R. REDDY (ed.): 'Speech Recognition', Academic Press, New York, pp. 512—542, 1975.

[3] *H. Bourlard, Y. KAMP, H. Ney, C. J. Wellekens* (1985): „Speaker Dependent Connected Speech Recognition via Dynamic Programming and Statistical Methods", in M. R. Schroeder (ed.): 'Speech and Speaker Recognition', Karger, Basel, pp. 115—148, 1985.

[4] *J. S. Bridle, M. D. Brown, R. M. Chamberlain* (1982): „An Algorithm for Connected Word Recognition", Proc. 1982 IEEE Conf. on Acoustics, Speech and Signal Processing, Paris, France, pp. 899—902, May 1982.

[5] *Y. L. Chow, B. Schwartz, S. Roucos et al.* (1986): „The Role of Word-Dependent Coarticulatory effects in a Phoneme-Based Speech Recognition System", Proc. 1986 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Tokyo, Japan, pp. 30. 9. 1—4. April 1986.

[6] *K. Fukunaga* (1972): "Introduction to Staistical Pattern Recognition" Academic Press, New York, 1972.

[7] *J.—P. Haton (ed.)* (1982): "Automatic Speech Analysis and Recognition" Nato Advanced Study Institute Series, D. Reidel Publishing Company, Dordrecht, Holland, 1982.

[8] *J. E. Hopcroft, J. D. Ullman* (1979): "Introduction to Automata Theory, Languages and Computation" Addison-Wesley Publishing Company, Reading, Massachusetts 1979.

[9] *F. Jelinek* (1976): „Continuous Speech Recognition by Statistical Methods", Proc. of the IEEE, Vol. 64, No. 10, pp. 532—556, April 1976.

[10] *F. Jelinek* (1985): „The Development of an Experimental Discrete Dictation Recognizer", Proc. of the IEEE, Vol. 73, No. 11, pp. 1616—1624, Nov. 1985.

[11] *W. A. Lea (ed.)* (1980): "Trends in Speech Recognition" Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

[12] *B. T. Lowerre* (1976): „The Harpy Speech Recognition System", Ph. D. Thesis, Carnegie Mellon University, Dept. Computer Science, Pittsburgh, Pennsylvania, April 1976.

[13] *R. De Mori, C. Y. Suen (ed.)* (1985): 'New Systems and Architectures for Automatic Speech Recognition and Synthesis', Proc. of the NATO Advanced Science Institute held at Bonas, Gers, France July 1984, Springer-Verlag, Berlin, 1985.

[14] D. *Mergel* (1986): „A Language Model for Spoken German Data Base Queries", Int. Conf. on, 'Speech Input/Output: Techniques and Applications', London UK, pp. 9—14, March 1986.

[15] *H. Ney* (1982): „Dynamic Programming as a Technique for Pattern Recognition", Proc. 6th. Int. Conf. on Pattern Recognition, Munich, Germany, pp. 1119—1125, Oct. 1982.

[16] *H. Ney* (1984): „The Use of a One-Stage Dynamic Programming Algorithm for Conected Word Recognition", IEEE Trans. on Acoustics, Speech and Signal, Vol. ASSP-32, No. 2, pp. 263—271, April 1984.

[17] *H. Ney*, D. *Mergel, S. M. Marcus* (1986): „On the Automatic Training of Phonetic Units", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, pp. 209—213, Jan.-Feb. 1986.

[18] *L. R. Rabiner, S. E. Levinson* (1981): „Isolated and Connected Word Recognition — Theory and Selected Applications", IEEE Trans. on Communications, No. 5, Vol. COM-29, pp. 621—659, May 1981.

[19] *T. K. Vintsyuk* (1971): „Element-wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary", Kibernetika (Cybernetics), Vol. 7, No. 2, pp. 133—143, March-April 1971.

[20] *T. K. Vintsyuk* (1982): „Speech Recognition and Understanding", Kibernetika (Cybernetics), Vol 18, No. 5, pp. 101—105, Sept.-Oct. 1982.