DR. GORDOS GÉZA
BME Híradástechnikai Elektronika Intézet

Újabb eredmények a gépi beszédfeldolgozásban

ETO 534.78:681.3

A Budapesti Műszaki Egyetem Villamosmérnöki Karának Híradástechnikai Elektronika Intézetében, illetve jogelődjén — a Vezetékes Híradástechnikai Tanszéken — az 1960-as évek közepe óta folynak vizsgálatok a gépi beszédfeldolgozás tárgyában. Jelen közlemény a Villamosmérnöki Kar alapításának 30 éves évfordulója alkalmából készült, és a szerző vezetésével folytatott vizsgálatok legfontosabb irányait és — elsősorban az ikerzygozitás diagnosztizálására irányuló — néhány eredményét célozza összefoglalni.

Rövid helyzetkép

A gépi beszédfeldolgozás az emberi beszélés, hallás és megértés — röviden az emberi beszédlánc — egy vagy több funkciójának művi megvalósítása. Főbb ágai a

- beszédfelismerés,
- beszélőazonosítás és
- beszéd-előállítás.

A beszédfelismeréssel és beszélőazonosítással rokon módszereket használ a csecsemősírás osztályozása, mely a sírás okának (éhség, fájdalom stb.) megállapítását célozza.

A gépi beszédfeldolgozás néhány lényegesnek tűnő alkalmazása az alábbi:

- lehetőleg kétirányú ember—gép kommunikáció létesítése akusztikus úton,
- diktálás gépi leírása,
- beszédrészletek vizualizálása süketek beszélni tanításához,
- csökkentett csatornaigényű beszédátvitel,
- titkosítás,
- ikrek egy- vagy kétpetéjűségének (zygozitás) eldöntése akusztikus vizsgálatokkal, egyes betegségek örökletes voltának megállapítását célzó vizsgálatok elősegítésére,
- beszélő- és hallószervek egyes elváltozásainak kimutatása.

A jelen dolgozatban ismertetendő eljárások jobb megértéséhez célszerűnek tűnik a beszédfelismerés (és a vele nagyon rokon beszélőazonosítás) néhány alapelvének megkülönböztetése.

A beszédfelismerés első lényeges eredménye — a hosszan tartott magánhangzók és egyes zöngés mássalhangzók gépi detektálása — a formánsstruktúra felismerésén alapult (pl. [3]). Ez a siker lineáris szűrőkkel végrehajtott spektrális vizsgálatok irányába terelte a kutatásokat. A Helmholtztól [1] származó és Békési [2] által anatómiailag is megalapozott és továbbfejlesztett hallószervi modell lényeges eleme az egymással laza csatolásban levő rezonátorok halmaza, és ez a tény egy ideig azt sugallta, hogy a szűrős rendszerek finomításával a beszédfelismerés legdöntőbb mozzanata megoldható lesz. Noha e próbálkozások sok maradandó eredményt is hoztak, a kezdeti nehézségeket (nem hosszan tartott, hanem normális idejű hangzók bizonytalan felismerése, jelentésmódosító átmenetek kezelhetetlensége, nem-, ill. személyfüggőség stb.) megnyugtatóan soha nem sikerült leküzdeni. A 70-es években fogalmazódtak meg a módszer elvi korlátai (pl. [4]). Ekkor sikerült ugyanis kapcsolatba hozni a Heisenberg-féle határozatlansági reláció lineáris szűrőkre Gábor Dénes által levezetett [5] és Harkevics által [6] finomított következményeit és a fül frekvencia-analizáló képességére nyert kísérleti adatokat (pl. [7]). Az előbbi leszögezi, hogy egyrészt egy ΔT ideig tartó szinuszcsomag frekvenciáját csak akkor lehet sávszűrők segítségével Δf pontossággal megállapítani, ha $\Delta T.\Delta f > C$, másrészt C a színuszcsomag burkolójától függően 1 és 2 körüli. Ezzel szemben az utóbbi szerint a fül ilyen irányú teljesítőképességét — a kb. 2 kHz-ig terjedő sávban — a $\Delta T.\Delta f > 0{,}18$ reláció írja le. E két tény összevetéséből leszűrhető az a megállapítás, hogy a fül a spektrális analízis sebessége szempontjából szűrő típusú lineáris rendszerrel nem modellezhető, annál mintegy egy nagyságrenddel hatékonyabb.

Sok szempont — melyek közül egyet az előzőek taglaltak — játszott közre abban, hogy a 60-as évek közepétől a beszédfelismerésben egy stratégiaváltás kezdett kibontakozni. Az emberi hallásmechanizmus gépi utánzása mellett a beszédfelismerés a hanghullámot azzal a céllal is analizálni kezdte, hogy belőle visszakövetkeztessen a beszélő szervek folyamataira. Nyilvánvaló ugyanis, hogy a beszélő szervek folyamatait a mondanivaló határozza meg, s a beszédfelismerés célja éppen e mondanivaló megállapítása. A beszédfelismerést a beszédképzés oldaláról megközelíteni általában is ígér nyereséget: az ember az általa felfoghatóknál sokkal kevesebb akusztikai jelenség létrehozására képes. A beszédfelismerés tehát azonosan értékes végeredményhez lényegesen egyszerűbbnek tűnő mechanizmus — a beszédképzés — tanulmányozásával juthat.

Az alkalmazott módszerek további lényeges eltolódásához vezetett az a fokozatosan érlelődő sejtés, majd meggyőződés (Klatt, Stevens, 1971), hogy kizárólag akusztikai jelenségekből az ember sem képes tökéletes felismerésre.

Ezért a gépi beszédfelismerés ma már több szinten, és a szintek között esetenként visszacsatolással operál. A leggyakrabban alkalmazott részeljárások:

akusztikai szintű eljárás, amely

Beérkezett: 1979. X. 15.

— akusztikai lényegkiemelést végez,
— osztályoz és
— akusztikai elemet (fonéma, apel stb.) eredményez,

nyelvi szintű eljárás, amely
— szószintre,
— szintaktikai szintre és
— zemantikai szintre bontható, s végül a

pragmatikai szintű (a beszélő személyt, a közlés jelentőségét stb. figyelő) eljárás.

A magasabb szintű eljárások ellenőrzik az alacsonyabb szintűeket, az azok által nyújtott megoldási lehetőségekből válogatnak, illetve azokat új lehetőségek felkutatására utasítják.

Egy lényegkiemelő rendszer

A BME—HEI-n folyó munkálatok a gépi beszédfeldolgozás helyzetének folyamatos figyelemmel kísérésén túl a beszédfelismerés akusztikai szintjére és a teljes beszélőfelismerésre terjednek ki.

E vizsgálódások bázisa kezdetben (1974-ig) néhány cél-eszköz és egy Odra 1024, majd egy PDP8, s végül egy PDP11/40 számítógép volt, de már megkezdődött (1978-tól) a szuverén mikroprocesszoros lényegkiemelő alrendszerek kiépítése is.

Illusztratív példaként röviden felsoroljuk a PDP 11/40 gépre alapozott lényegkiemelő rendszer legfontosabb jellemzőit. A hanganyag változtatható frekvenciájú mintavételezés, majd kvantálás után kerül a háttértárba. Onnan folyamatosan mozgatható ablakkal áttekintésre grafikus vagy alfanumerikus megjelenítőre hozható. Az ablak tipikusan 512 mintát tartalmaz. A megjelenítésben szomszédos minták a tárolt mintasorozat ritkításával is származhatnak, miáltal a vizsgált hanganyagról „távlati" kép is nyerhető (egy ablak ideje pl. 40 ms és 1,3 s között változhat). Két függvényrészlet összehasonlításának elősegítésére lehetőség van a két függvényrészlet egyidejű megjelenítésére is. A megjelenített függvényrészleteken két folyamatosan mozgatható kijelölővel (cursor) tetszőleges szegmens kiválasztható, és további feldolgozásra definiálható. A két kijelölővel időmérés is végezhető (pl. zárhangidő mérése).

Az interpolációs opció arra ad lehetőséget, hogy a kiválasztott szegmens, a benne levő tényleges minták darabszámától függetlenül, a további feldolgozáshoz illeszkedő mintaszámúra legyen transzformálható. Ez például a különböző sebességgel kimondott, azonos fonetikai értékű részletek összehasonlítását (időillesztést), az FFT alkalmazását vagy — a később részletezett — spektrális függvények összehasonlítását (frekvenciaillesztést) segíti elő.

Lényeges szolgáltatás a kiválasztott szegmens módosításának (törlésének, áthelyettesítésének, amplitúdómodulációjának stb.) lehetősége. E szolgáltatás elsőrendű célja az indirekt lényegkutatás megvalósítása (a lényeg a hanghullám felismerést befolyásoló tulajdonsága). Ha ugyanis egy szegmens a sejtett lényeget tartalmazza, s ezt a szegmenset töröljük vagy megváltoztatjuk, akkor a mintákból visszaállított hanghullám az ember számára felismerhetetlenné válik. Ugyancsak ezzel a szolgáltatással munkáihatók tovább azok a felvetések [8], hogy egyes akusztikai elemek lefolyásának időbeli megfordítása mennyire befolyásolja a felismerhetőséget.

A spektrális vizsgálatokat tetszőleges szegmensen végrehajtható gyors Fourier-transzformáció (FFT) és cepstrum-számítás (mégpedig valamennyi változatban, nevezetesen: cepstrum, clipstrum, középrész-kivágott clipstrum) segíti elő. Egy $f(t)$ függvény

$$c(q)=\left|\mathcal{F}^{-1}\{\log|\mathcal{F}\{g[f(t)]\}|^l\}\right|^l$$

transzformáltja, ahol $\mathcal{F}$ a Fourier-transzformáció,
$l=1$ esetben „amplitúdó" jellegű,
$l=2$ esetben „teljesítmény" jellegű és
cepstrum, ha $g(x)=x$;
clipstrum, ha

$$g(x)=\begin{cases}+A, & \text{ha} \quad x\geqq 0\\ -A, & \text{ha} \quad x<0,\end{cases}$$

és középrész-kivágott clipstrum, ha

$$g(x)=\begin{cases}X-A, & \text{ha} \quad x\geqq A\\ 0, & \text{ha} \quad A>x\geqq -A\\ X+A, & \text{ha} \quad -A>x.\end{cases}$$

Újszerű szolgáltatás a dinamikus cepstrum megjelenítés, amely egy vagy két időfüggvény szabályos időközökkel eltolt szegmenseiből képezett cepstrumait jeleníti meg egymás után, ezzel a cepstrumok változását mintegy filmszerűen megjelenítve.

Mind az FFT-, mind a cepstrumszolgáltatást az alapfrekvencia, valamint a formánsfrekvenciák interaktív meghatározását elősegítő rutinok egészítik ki.

Szegmensek auto- és keresztkorrelációs vizsgálata is elvégezhető.

A fenti interaktív szolgáltatásokon túlmutat az alapfrekvencia (= hangmagasság) átlagértékének és szórásának automatikus meghatározása. Az eljárás az AMDF-en (Average Magnitudé Difference Function) alapul, melynek definíciója egy $\{x(n)\}$, $n=[0, 1+N]$ mintasorozaton:

$$y(k)=\frac{1}{L}\sum_{N=l}^{l+L-1}|x(n)-x(n+k)|,$$

ahol k az eltolás, $L<N$; $0\leqq l\leqq N-L+1$, és l a vizsgálat kezdetét jelöli ki. E függvény periodikus $\{x(n)\}$ esetén a periódusidőnek (ill. K-indexnek) megfelelő eltolásnál elvileg zérust, a gyakorlatban minimumot ad. A gyakorlatban kritikus a minimumot definiáló

$$k=K, \quad \text{ha} \quad y(k)<Y$$

összefüggésben szereplő Y megválasztása. Kísérleteink kimutatták, hogy kielégítő eredményt kaphatunk a

$$K_1=c\cdot\left\{\frac{1}{N+1}\sum_{n=0}^{N}\left|x(n)-\frac{1}{N+1}\left[\sum_{i=0}^{N}x(i)\right]\right|\right\},$$

$$c=0{,}6\div 0{,}8,$$

vagy a

$$K_2=\frac{1}{32}\cdot x|\{(n)\}|_{max}$$

választással. A szolgáltatás kiterjed k ésszerű határok között tartására, és l automatikus, egyenlő közű változtatására, miáltal a hanghullám minden kváziperiodikus része felismerhető, és alapfrekvenciája meghatározható.

A PDP 11/40-en rendelkezésre álló további lényegkiemelő és osztályozó szolgáltatásokat, továbbá az alapperiódus meghatározásának és a zárhangok felismerésének, valamint zárhangidejük lemérésének automatikus elvégzését biztosító már elkészült autonom mikroprocesszoros alrendszereket egy későbbi közlemény fogja ismertetni. (Már itt megjegyezzük, hogy az alrendszer egészen más célú kutatásban — nevezetesen impulzuszaj-csapda létrehozásában — is hasznosnak bizonyult.)

Iker-zygozitás eldöntése akusztikus jellemzők alapján

A gépi beszédfeldolgozás területén végzett munkáink leglényegesebb motivációját az iker-zygozitás akusztikai meghatározása jelentette. Többek között bizonyos betegségek örökletes voltának tanulmányozásában van annak jelentősége, hogy az ikrek egy- vagy kétpetéjűek (mono- vagy dizygoták). Ez utóbbi megállapítása azonban a születés után már korántsem triviális. A szokásos eljárás minden ikerpárt monozygótának tételez fel, és különböző vizsgálatokkal (antropológia, vércsoport, nyomelemek stb.) vagy megállapítja a dizygozitást, vagy továbbra is feltételezi a monozygozitást. Egyrészt a diagnózis pontosítására, másrészt kényelmesebbé tételére Forrai felvetette az ikrek hangjának vizsgálatát. E célból Forrai és Lubi 117 ikerpár személyenként kb. 1 percens azonos szöveget tartalmazó hangfelvételét bocsájtotta rendelkezésünkre (1973). A beszélőazonosítás és beszédfelismerés módszerein alapuló megközelítés alapjait [9] foglalja össze.

Az első lépésben azt sikerült számszerűen kimutatni, hogy két egyén beszédének bizonyos akusztikai jellemzői egy alkalmas módon definiált „távolság" értelmében átlagosan akkor vannak közelebb egymáshoz, ha a két egyén egy ikerpár két tagja.

Az i-edik és j-edik egyén távolsága egy J jellemző mentén legyen:

$$D_{ij}(J)=\left|\frac{J_i-J_j}{J_i+J_i}\right|.$$

Méréseink az alábbi eredményekhez vezettek:

Vizsgált jellemző J_i**	Átlagos „távolság": $E\{D_{ij}(J)\}$	
	ikrek halmazán	nem ikrek halmazán
Egyén zárhangidőinek átlaga	0,12	0,26
Formánsfrekvencia	0,05*	0,07*
Alapfrekvencia átlaga	0,04	0,06

(* Az á és é hang három—három formánsából képezett halmazon is átlagolva.)
(** Az alapadatokat a PDP 11/40-en működő rendszer szolgáltatta)

Nyilvánvaló a táblázatból az ikrek hangtani rokonsága.

A mono- és dizygoták elkülöníthetőségének esélyét az ikerpár két egyedén mért átlagos alapfrekvenciák egyszerű különbsége már megítélhetővé teszi. 49 férfi ikerpáron az alábbi különbségek adódtak:

Átl. alapfr. különbs. Hz	Korábbi osztályozás	Átl. alapfr. különbs. Hz	Korábbi osztályozás
0,8	M	7,9	M
1,1	M	8,0	D
	D	8,2	D
1,2	M	8,3	M
1,5	M	8,6	D
1,6	M	11,0	D
1,9	M	11,9	D
	M	12,0	D
2,0	M		M
2,1	M	14,5	D
	M	16,6	D
3,1	M	18,0	M
3,2	D	21,6	D
3,4	M	23,4	D
4,1	M	23,7	M
	M	23,8	D
1,5	M	27,9	M
4,6	M	30,5	D
5,1	M	33,5	M
5,7	M	33,7	M
	D	38,9	D
5,9	D	39,1	D
6,0	D	52,3	M
6,5	M	59,0	M
5,7	D		

(Az alapadatokat mikroprocesszoros mérőrendszer szolgáltatta.)

A táblázatból kitűnik, hogy van esély a szétválasztásra, de ehhez egyetlen jellemző nem elég.

Rövidesen időszerűvé válik azon reményt keltő két vizsgálat közzététele, melyek közül

— az egyik az átlagos alapfrekvenciák különbségén, szóeleji „k" hang átlagos zárhangidején és a mono-kórus kiértékelésén alapuló több változós diszkriminancia-analízissel működik,
— a másik kilenc komponensű („á" és „é" 3—3 formánsa, hangmagasság várható értéke és szórása, zárhangidők átlaga), tanuló algoritmussal optimalizált súlyozású vektorokat osztályoz a legközelebbi szomszédra való döntés alapján.

Munkatársak és köszönetnyilvánítás

A dolgozatban ismertetett eljárások és eredmények többsége egyetemi hallgatók tudományos diákköri, önálló laboratóriumi munka vagy diplomatervezés keretében kifejtett — általában hallgatónként, néhány hónapra kiterjedő — közreműködésével született. A legértékesebb segítséget adó hallgatók — azóta már végzett mérnökök — a diplomázás éve, s ezen belül az abc szerinti sorrendben: Füredi Ágnes, Nagy Péter, Takács György, Garazsi Erika, Gönci János, Molnár Géza, Rumi László, Pannuska Rudolf, Schmidt Gábor, Dankó László és Kormány

Zoltán. Köszönet illeti sok ötletet adó minőségében dr. Földváry Rudolfot, Gönci Jánost, dr. Osváth Lászlót és különösképpen Takács Györgyöt.

Az egyetemi hallgatók munkájának — és ezzel a lényegkiemelő rendszer létrejöttének — feltételeit a PDP—8 gépen dr. Pásztorniczky Lajos és Hetényi Tamás, a PDP 11/40 gépen dr. Schnell László és munkatársai, autonóm mikroprocesszoros rendszeren pedig dr. Csibi Sándor teremtették meg.

IRODALOM

[1] *Helmholtz, H.:* Die Lehre von den Tonschwingungen Braunschweig, 1913.

[2] *Békési Gy.:* Experiments in Hearing. McGraw—Hill, 1960.

[3] *Tarnóczy T.—Radnai J.:* Eine Möglichkeit automatischer Erkennung von Vokalen. Proc. VIIth ICA, Vol. III., Budapest, 1971. pp. 61—64.

[4] *Földváry R.—Gordos G.:* Az emberi hangmagasság-felismerés új hipotetikus modellje. Híradástechnika, Vol. XXV. (1974), No. 11. pp. 344—348.

[5] *Gabor, D.:* Acoustical Quanta and the Theory of Hearing. Nature, Vol. 169. (1947), May, pp. 591—602.

[6] *Harkevics, A. A.:* Spectra and analysis. Consultants Bureau. New York, 1960.

[7] *Grobben, L. M.:* Appreciation of short tones. VIIth. ICA, Budapest, 1971. Vol. 3. pp. 329—332.

[8] *Tarnóczy T.— Vicsy, K.:* Some Remarks on the Perception of Voiceless Stopconsonants. Acustica, Vol. 43. 1979. No. 2. pp. 167—173.

[9] *Forrai Gy.—Gordos, G.—Lubi, B.:* Preliminary Report on Voice-Based Discrimination between Monozygotic and Dizygotic Twins. Proc. of. Phis. Inst. ELTE (Megjelenés alatt)