Magyar Pszichológiai Szemle, 2007, 62. 4. 541–563.
DOI: 10.1556/MPSzle.62.2007.4.8

BESZÉLŐFELISMERÉS – NEUROLÓGIAI HÁTTÉR ÉS PSZICHOLÓGIAI MODELLEK

BŐHM TAMÁS MIHÁLY
PhD-hallgató

Budapesti Műszaki és Gazdaságtudományi Egyetem – Távközlési és Médiainformatikai Tanszék
E-mail: bohm@tmit.bme.hu

Beérkezett: 2007. 01. 15. – *Elfogadva:* 2007. 06. 10.

Beszédük alapján fel tudjuk ismerni családtagjainkat, ismerőseinket. Ez azért lehetséges, mert a beszéd a nyelvi üzeneten kívül a beszélő személyére utaló nem nyelvi információt is tartalmaz. Ebben a tanulmányban áttekintjük az emberi beszélőfelismerés szakirodalmának egy részét, a vonatkozó definíciókat és az alkalmazott vizsgálati módszereket. A szakterület kutatásában új távlatokat nyitó legmodernebb módszert, az fMRI-t alkalmazó tanulmányokra összpontosítunk. Ezek megállapításait összehasonlítjuk a viselkedéses pszichológiai kísérletekkel elért eredményekkel, és kiemeljük a beszédtechnológiai vonatkozásaikat. Kísérletet teszünk az elszórtan publikált pszichológiai modellek összefoglalására is.

Kulcsszavak: *beszélőfelismerés, hangfelismerés, nem nyelvi információfeldolgozás, hangterület, STS, multimodális személyfelismerés*

NEM NYELVI INFORMÁCIÓ A BESZÉDBEN

Egy beszédrészlet észlelésekor agyunk többféle döntést hoz. Dekódolja a nyelvi üzenetet, azaz hogy milyen szavak, mondatok hangzottak el (beszédpercepció), valamint megítéli például a beszélő személy korát és hangulatát is (GÓSY, 2001a, hivatkozva FUJIMURÁra, 1972-ben; GOCSÁL, 1998). Az akusztikai jelben a nyelvi üzenettel párhuzamosan átvitt információkat közösen *nem nyelvi információnak* nevezzük. Gyakran ugyanaz az akusztikai paraméter hordozhat nyelvi és nem nyelvi információt is: például BATLINER, STEIDL és NÖTH (2007) szerint az érdes hang jelezheti egy bemondás végét (nyelvi), kifejezhet unalmat (nem nyelvi) vagy lehet a beszélő személy hangszalagjainak a jellegzetessége (nem nyelvi). A beszéd nem

nyelvi rétegét szokták még vokális paralingvisztikai (KREIMAN, VAN LANCKER-SIDTIS, GERRATT, 2005), indexikális (NYGAARD, PISONI, 1998) és nyelven „kívüli" információnak (GÓSY, 2005) is nevezni, de ezek jelentése nem egyértelmű, mert a nem nyelvi információnál néha tágabb, néha szűkebb értelemben használják.

A nem nyelvi információ két nagy csoportját különböztethetjük meg, úgy mint a beszélő érzelmi állapotára és a személyére vonatkozót. Ezeket tovább bonthatjuk komponensekre: az érzelem esetén például a hangulat, arousal és a stressz, míg az identitás esetén a nem, kor, megjelenés (magasság, testalkat), egészségi állapot (dohányzás, megfázás, fáradtság), képzettség és szociális státus. A fenti felosztás a legelterjedtebb, de a nem nyelvi információk más csoportosítása is ismert. Például KREIMAN, VAN LANCKER-SIDTIS és GERRATT (2005) a nem nyelvi információkat nem érzelmi állapotra és beszélőidentitásra bontják szét, hanem a beszélő fizikai, pszichológiai és szociális jellemzőit és ezek komponenseit különböztetik meg.

A mindennapi életben megfigyelhetjük, hogy az emberek milyen hatékonyan képesek a nem nyelvi információk dekódolására, de vajon miért jelennek meg ezek az információk a beszédben? Az érzelmi információ esetén a különböző állapotokhoz kapcsolódó automatikus izom-összehúzódások és az akaratlagos gesztusok megváltoztatják a beszélőszervek működését (az artikulációs konfigurációt) és így modulálják az akusztikus beszédjelet (BELIN, FECTEAU, BÉDARD, 2004). Az identitásinformáció a beszélőszervek egyéni különbségeiből (például a beszédcsatorna hossza vagy a hangszalagok rugalmassága), valamint az elsajátított artikulációs mozgások apró különbségeiből (például nyelvjárásbeli eltérések) adódnak. Ha egy szót felveszünk két bemondó kiejtésében, akkor sokkal nagyobb különbséget tapasztalunk, mintha egy bemondó kétszeri kiejtésében rögzítjük (GÓSY, NIKLÉCZY, 1999) – legalábbis bizonyos, a beszélő személyére jellemző akusztikai paraméterek tekintetében.

Nem nyelvi információ verbális tartalom nélkül is megjelenhet a beszédben: a nevetés vagy a hümmögés is hordozza az egyénre és érzelmi állapotára utaló ismertetőjegyeket. Valószínűleg a beszéden kívül tetszőleges emberi vokalizáció, azaz a hangképzőszervekkel keltett hang is tartalmazhat nem nyelvi információt. Míg a beszéd az evolúció során viszonylag későn kialakult folyamat, a nem nyelvi információkat hordozó vokalizációk több millió éve igen elterjedtek a gerincesek körében (BELIN, FECTEAU, BÉDARD, 2004). Egyes fajok számára a túléléshez létfontosságú, hogy értelmezni tudják azokat az akusztikus jeleket, amelyek azonos fajú egyedektől vagy éppen egy ragadozótól, vagy zsákmánytól származnak. A makákók például képesek társaikat felismerni a hangjuk alapján, vagy az északi bundás fóka anyák és kölykeik hang alapján megtalálják egymást a hatalmas populációban. A fókák ezt a képességet legalább 4 évig megőrzik, annak ellenére, hogy mindössze 2–5 nap alatt sajátítják el (BELIN, FECTEAU, BÉDARD, 2004; hivatkozva INSLEY-re, 2000-ben és CHARRIER-re és munkatársaira, 2001-ben). Nyitott kérdés, hogy vajon fajok között mennyire hatékony a nem nyelvi információközlés. Talán éppen ez történik, amikor egy kutyához „beszél" a gazdája: nem a nyelvi tartalmat, hanem az érzelmi töltetet dekódolja az állat.

Gyermekek beszédészlelési képessége csak hónapokkal a születés után kezd kifejlődni (HOUSTON, 2005), de mikortól képesek egyes személyeket a hangjuk alap-

ján felismerni? Pulzusszámmérésekkel kimutatták, hogy az újszülöttek diszkriminálni tudnak beszélők között, és felismerik szüleik hangját. Ez az utóbbi képesség már a születés előtt is kimutatható (BELIN, FECTEAU, BÉDARD, 2004; hivatkozva KISILEVSKY-re és munkatársaira, 2003-ban).

A nem nyelvi jelek feldolgozása bizonyos szempontból robusztusabb a beszéd nyelvi feldolgozásánál, más szempontból kevésbé az. Az előbbit jól demonstrálja, hogy távolból vagy falon át hallott érthetetlen beszéd alapján is gyakran képesek vagyunk a beszélőt azonosítani és érzelmi állapotára következtetni. Viszont ideális körülmények között a felismerés lassabb (VON KRIEGSTEIN, KLEINSCHMIDT és munkatársai, 2005) és pontatlanabb (YARMEY, YARMEY és munkatársai, 2001), mint a beszédértés.

BELIN, FECTEAU és BÉDARD (2004) megfogalmazásában az emberek hangja a „hallható arcuk". Hangok és arcok sokban hasonlítanak: agyunk számára ezek a leggyakoribb és legfontosabb auditív és vizuális objektumok. Mindkettő olyan médium, amely nyelvi, érzelmi és identitásinformációt továbbít, csak más modalitásban. Arcoknál a nem nyelvi információk eredete ugyanaz, mint beszédnél, például a személyre utaló információ egyrészt a fizikai adottságokból, másrészt az elsajátított, egyedi mozgásokból származik. Míg telefonálásnál csak az egyik modalitás áll rendelkezésre, természetes környezetben az arcot és a hangot együtt észleli a megfigyelő személy. Talán ez lehet az oka annak, hogy ismerős személyek hangját hallva az ismert arcok feldolgozására specializálódott agyi terület *(fusiform face area,* FFA) is aktiválódik (VON KRIEGSTEIN, KLEINSCHMIDT és munkatársai, 2005). A két modalitást valószínűleg képtelenek vagyunk „függetleníteni"; ha csak az egyik áll rendelkezésre, akkor hozzáképzeljük a másikat. Ezt támasztja alá az a megfigyelés is, hogy amikor ismeretlenekkel beszélünk telefonon, automatikusan egy arcot rendelünk hozzá, vagy az, amikor egy ismeretlen archoz tartozó hangot képzelünk el (például fotó alapján) – a jelenséget a foniátriában *alkati harmóniának* nevezik (GÓSY, 2001a). Akkor lesz nyilvánvaló, hogy a hang-arc asszociáció ténylegesen megtörtént, ha egy személyes találkozás alkalmával a valós és az elképzelt hang/arc ellentmond egymásnak.

A beszéd percepcióját közel egy évszázada intenzíven kutatják (például FLETCHER, 1929), de a nem nyelvi információk percepciójával csak kevés tanulmány foglalkozott. Az ilyen eredmények azonban a beszédpercepciót is tágabb környezetbe helyeznék. A beszédtechnológiában is csak az elmúlt évtizedben kezdtek komolyan foglalkozni a területtel, például a beszélő személy gépi azonosításával (beléptetőrendszerekben; CAMPBELL, 1997) és az érzelmek gépi felismerésével (telefonos dialógusrendszerekben; BLOUIN, MAFFIOLO, 2005).

DEFINÍCIÓK

A nem nyelvi információkat mint a beszédből a nyelvi, verbális üzeneten túl kinyerhető információk összességét definiáljuk. Az elterjedt megfogalmazás szerint *amit* mondunk, az nyelvi, *ahogy* mondjuk, az nem nyelvi információ. Még nem tisztázott azonban, hogy mennyire éles a határ a kétféle típusú információ között.

NYGAARD és PISONI (1998) például viselkedéses pszichológiai kísérletek alapján arra következtetett, hogy a kétféle információhoz közös agyi reprezentációk tartoznak. Ezzel szemben az agyi képalkotó eljárások kimutatták, hogy a nyelvi és a nem nyelvi feldolgozás legalább részben különálló agyi hálózatok segítségével történik (BELIN, FECTEAU, BÉDARD, 2005).

A beszéd nagymértékben változékony folyamat: ugyanazt a szót lehetetlen kétszer pontosan ugyanúgy kiejteni (GORDOS, TAKÁCS, 1983). Ez a változékonyság kisebb, ha a két bemondás ugyanattól a beszélőtől származik – legalábbis egyes akusztikai mértékek szempontjából. Ezek az akusztikai mértékek az *egyéni hangjellemzők* (összességében): különböző beszélők ejtései között nagy eltéréseket mutatnak, egy beszélő ejtéseiben azonban kevesebbet változnak. Ezek a mértékek lényegében invariánsak a beszélőre nézve, hasonlóan a fonetikai megkülönböztető jegyek invarianciájához egy adott fonémára nézve (STEVENS, 2005). Így az adott szó több kiejtése közötti változékonyság egyik forrása a beszélő személy egyéni hangjellemzői. További forrás lehet az érzelmi állapottal, a fizikai környezettel összefüggő hangjellemzők és véletlen vagy még nehezen magyarázható tényezők. Az egyéni hangjellemzők kialakulásában nagy a szerepe a környezeti hatásoknak, de részben genetikailag meghatározottak (GÓSY, 2001). Számos tanulmány foglalkozott az egyéni hangjellemzők feltárásával, azok artikulációs, pszichológiai, fiziológiai és szociológiai hátterével (JOHNSON, LADEFOGED, LINDAU, 1993; LADEFOGED, BROADBENT, 1957; HENTON, BLADON, 1987). Ezek a jellemzők lehetnek statikusak (időben állandóak), mint például az átlagos alapfrekvencia, az alapfrekvenciatartomány, az átlagos formánsfrekvenciák, a jellemző zöngeminőség, de lehetnek dinamikusak is, mint a felpattanó zárhangok zöngekezdési ideje *(voice onset time,* VOT), egyes magánhangzók egyedi/nyelvjárásbeli ejtéséből adódó formánseltérések (LADEFOGED, BROADBENT, 1957) vagy az akusztikai jel alakulása a bemondások kezdetén és végén.

Bár a magyar szakirodalom erre a fogalomra *egyéni beszédjellemzőkként* (GÓSY, 2004, 216) hivatkozik, ebben a tanulmányban az *egyéni hangjellemzők* megnevezést is használjuk ugyanolyan értelemben. Valószínűsíthető ugyanis, hogy a személyes sajátosságok egy része (például a szervi adottságokból adódóak) nem csak beszéd, hanem egyéb vokalizációk esetén is megjelenhetnek (BELIN, FECTEAU, BÉDARD, 2005).

A *beszélőfelismerés* (speaker recognition) vagy *hangfelismerés* (voice recognition) az embereknek az a képessége, amivel számukra ismert személyeket hangjuk alapján azonosítani tudnak (GÓSY, 1999) – tehát egy emberi kognitív folyamat. Azokat a számítógépes programokat, amelyek beszédjel alapján több-kevesebb sikerrel megállapítják a beszélő személyazonosságát, *gépi beszélőfelismerő*nek (automatic speaker identification) nevezzük, a beszédet szöveggé alakító technológia pedig a gépi beszédfelismerés (speech recognition; GORDOS, TAKÁCS, 1983).

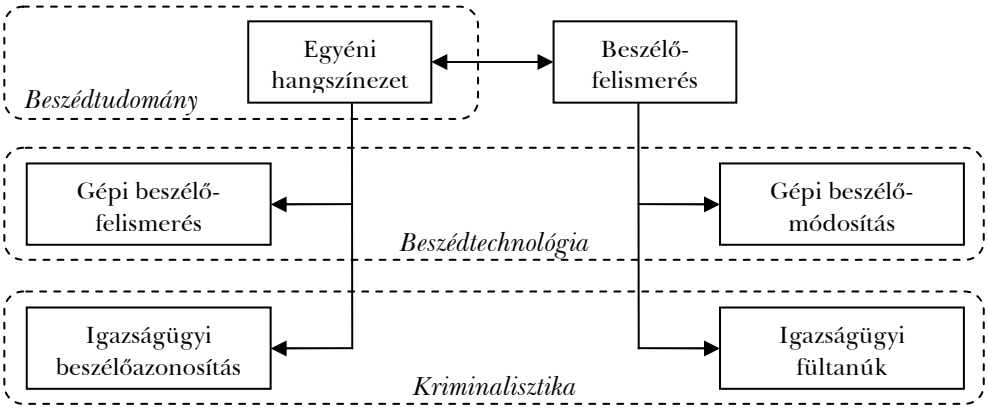
Az emberi beszélőfelismerés lényege valószínűleg az egyénre vonatkozó *érzeti ismertetőjegyek* kinyerése az akusztikai jelből, majd az ezek alapján végzett keresés az agyi személyreprezentációk között. Az *egyéni hangszínezet* ezeknek az ismertetőjegyeknek az összessége (GÓSY, 1999; GÓSY, 2004). Míg az egyéni hangjellemzők akusztikai paraméterek, az egyéni hangszínezet egy érzeti jelenség. A kettő összefügg egymással, de valószínűleg nem azonos:

– A pszichoakusztika eredményei alapján egy akusztikai mérték és a megfelelő auditív érzet között a kapcsolat általában többváltozós és nem lineáris. Például egy tiszta zenei hang hangossága nemcsak az amplitúdójától, hanem a frekvenciájától is függ, valamint az amplitúdó duplázását nem kétszeres hangosságként érzékeljük (STEVENS, 2000). Ehhez hasonlóan valószínű, hogy a beszélő hangjának érzeti ismertetőjegyeit több egyéni akusztikai hangjellemző nem lineáris összefüggésben határozza meg.
– Lehetnek olyan egyéni hangjellemzők, amelyeket beszédfeldolgozó programokkal pontosan le tudunk mérni, de hallórendszerünk nem érzékeli ezeket. Hasonlóan lehetnek olyan ismertetőjegyek, amelyeket az emberek könnyedén érzékelnek, de mérőeszközeink nem elég érzékenyek a kimutatásukra.

E két ok miatt nem sokat tudunk az ismertetőjegyekről. Az eddigi kísérletek csak azt vizsgálták, hogy egyes egyéni hangjellemzők mennyire befolyásolják a felismerés sikerét, azaz a hallgató felhasználja-e a jellemzőt az azonosítás során. Például az átlagos alapfrekvencia (ABBERTON, FOURCIN, 1978) és az irreguláris zöngeminőség (BŐHM, SHATTUCK-HUFNAGEL, 2007) megváltoztatása után a hallgatók szerint a felvétel kevésbé hasonlít az ismert beszélő hangjára.

KAPCSOLÓDÓ SZAKTERÜLETEK

A beszélőfelismerés szorosan kapcsolódik néhány egyéb területhez, az alábbiakban ezek rövid bemutatásával a tanulmány témáját pontosítjuk és kontextusba helyezzük *(1. ábra)*.



1. ábra. A beszélőfelismeréshez kapcsolódó szakterületek a beszédtudományban, a beszédtechnológiában és a kriminalisztikában

Az egyéni hangjellemzőkről rendelkezésre álló információkat felhasználják mind a beszédtechnológiában, mind a kriminalisztikában. Egyes beszédmérnökök olyan számítógépes programokon dolgoznak, amelyek egy személyt beszéde alapján azonosítani tudnak (a gépi beszélőfelismerők áttekintése olvasható CAMPBELL [1997] cikkében). Ezzel párhuzamosan a kriminalisztikában az igazságügyi szakértők jellemző feladata két beszédfelvételről megállapítani, hogy azonos bemondótól származnak-e (ROSE, 2002, 2; NOLAN, 1980; GÓSY, NIKLÉCZY, 1999). Mindkét alkalmazás számára kritikus kérdés, hogy az egyéni hangjellemzők milyen mértékben egyediek: ha nem eléggé, akkor mind a gépi beszélőfelismerés, mind a kriminalisztikai szakértői vizsgálatok elvi korlátokba ütköznek. Néhány évtizede egyes kutatók úgy gondolták, hogy az ujjlenyomathoz hasonlóan nagyon kicsi a valószínűsége, hogy van két ugyanolyan hangú ember – az ujjlenyomat angol megfelelője, a *fingerprint* nyomán *voiceprint*nek nevezték az egyéni hangjellemzőket (KERSTA, 1962). Azóta azonban számos ellenérv látott napvilágot és a kérdés jelenleg is nyitott (LADEFOGED, LADEFOGED, 1980; BONASTRE, BIMBOT és munkatársai, 2003).

A *gépi beszélőmódosító rendszerek* olyan programok, amelyek egy felvételen a beszélő hangját úgy módosítják, hogy egy másik beszélő hangjára hasonlítson. Tehát az egyéni hangszinezetet megváltoztatják, míg a nyelvi információt megőrzik. Az emberi beszélőfelismerésnek van kriminalisztikai vonatkozása is: számos országban bizonyító erejű a fültanúk vallomása, azaz ha a tanú felismeri a gyanúsított hangját (LADEFOGED, LADEFOGED, 1980; YARMEY, YARMEY és munkatársai, 2001). Mindkét terület számára kulcsfontosságú a beszélőfelismerés működésének és korlátainak megértése: a beszélőmódosítók ezt a kognitív folyamatot próbálják kijátszani, a bíróságokon pedig pontosan ismerni kell a fültanúvallomások megbízhatóságát.

Az *1. ábrá*n felvázolt tudományterületek kutatói könnyen hasznosíthatják egymás eredményeit. Egyik oldalról a beszédtechnológiai kutatásokhoz elengedhetetlen a mögöttes beszédprodukciós és agyi feldolgozó folyamatok ismerete. Másik oldalról a műszaki alkalmazások sikerei vagy akár kudarcai pszicholingvisztikai ismeretekké fordíthatók le. Például egy hatékony gépi beszélőfelismerő működési elve új ismereteket szolgáltathat az egyéni hangszínezettel kapcsolatban. Hasonlóan egy jól vagy rosszul működő beszélőmódosító rávilágíthat az emberi beszélőfelismerés során felhasznált érzeti ismertetőjegyekre, az egyéni hangszínezet egyes aspektusaira. Ennek ellenére a szerzőnek nincs tudomása ilyen jellegű interdiszciplináris együttműködésről.

VIZSGÁLATI MÓDSZEREK

Viselkedéses pszichológiai kísérletek

Az agyi képalkotó eljárások megjelenéséig a kutatók szinte kizárólag a pszichológia viselkedéses módszereivel tudták vizsgálni a kognitív folyamatokat, így a beszélőfelismerést is. Ennek köszönhetően a szakirodalomban számos ilyen úton kapott eredmény olvasható (VAN LANCKER, KREIMAN, EMMOREY, 1985a; VAN LANCKER, KREIMAN, WICKENS, 1985b; KREIMAN, PAPCUN, 1991; GONZALEZ, OLIVER, 2005;

OWREN, CARDILLO, 2006; ALLEN, MILLER, 2004; BROWN, STRONG, RENCHER, 1973; BROWN, STRONG, RENCHER, 1974; BŐHM, SHATTUCK-HUFNAGEL, 2007). A viselkedéses kísérleti megközelítésben a kísérleti személynek mindig valamilyen választ kell adnia a hallott inger alapján (választani több lehetőség közül, skálázni vagy akár szavakkal jellemezni a hallottakat) és a következtetéseket a válaszok alapján vonhatjuk le. Ez a módszer az agyat fekete dobozként kezeli: váltogatjuk az inputot és mérjük az outputot (KANWISHER, 2006).

Bár a módszer nagyon olcsó, a kísérletek megtervezése rendkívüli körültekintést igényel: ha nem zárunk ki minden külső tényezőt, ami befolyásolhatja a kísérleti személyek válaszát, akkor alternatív magyarázatok is lehetnek az eredményekre. További nehézség, hogy kizárólag ezeknek az input-output relációknak az alapján nagyon nehéz általános következtetéseket levonni a kognitív folyamatokról. Ennek ellenére a beszélőfelismeréssel kapcsolatos tudásunk java viselkedéses kísérleti eredményekből származik.

A három legelterjedtebb kísérleti elrendezés (és egyben adatelemzési módszer) az analízis szintézissel, a faktoranalízis és a sokdimenziós skálázás. Analízis szintézissel esetén egy beszédfelvétel egyes akusztikai jellemzőit megváltoztatjuk és megfigyeljük, hogy ez hogyan befolyásolja a kísérleti személyek beszélőfelismerési vagy beszélődiszkriminációs képességét. A módszer hatékonyan alkalmazható arra, hogy kontrollált körülmények között megvizsgáljuk az egyes egyéni hangjellemzők szerepét a beszélőfelismerésben. Hátránya, hogy nem garantált a manipulált felvételek természetes hangzása: bár egy külön kísérlettel ez ellenőrizhető, nem zárható ki teljes bizonyossággal, hogy a módosítások a beszélőfelismerés szempontjából torzították a felvételt.

A kísérleti személyek gyakori problémája, hogy a rendelkezésre álló rögzített válaszlehetőségek közül egyik se vagy egyszerre több is megfelel az inger által kiváltott érzetnek. Ha azonban saját szavaikkal írják le a hangot, amit hallottak, akkor nehéz sok hallgató adatait összesíteni és statisztikai eszközökkel elemezni. A számszerű válaszokat és a válaszok szabadságát ötvözi a faktoranalízis, amikor a kísérleti személyeknek minden hangot számos szempont (például hangos, kellemes, tiszta) szerint kell értékelniük, szempontonként egy folytonos skálán. A cél, hogy a kísérleti személyek a szöveges leírást megközelítő rugalmassággal jellemezhessék a beszélőket, így akár 49 különböző szempont is lehet egy vizsgálatban (VOIERS, 1964). Egy statisztikai eljárással a válaszok alapján azonosítható az a néhány egymástól független faktor, amely a válaszok változékonyságát okozza (KREIMAN, VAN LANCKER-SIDTIS, GERRATT, 2005). Az eljárás során az egymással erősen korreláló szempontok egy faktorban egyesülnek, minden faktoron különböző súlyokkal szerepelnek az egyes változók. Ezeknek a súlyoknak az alapján az egyes faktorok jelentésére lehet következtetni és ennek megfelelően el is nevezik őket (például dallamosság, hangmagasság, férfiasság). A faktorok valószínűleg a beszélőre utaló érzeti ismertetőjegyeknek felelnek meg. A faktoranalízis hátránya – a kísérleti személyek számára fárasztó feladaton kívül – az, hogy a szempontok összeállítása és a faktorok értelmezése is a kísérletvezető szubjektív döntésein múlik. A kiszámolt faktorokat gyakran nehéz értelmezni, mert sok és merőben eltérő szempontot egyesítenek. A hallgatók csak az adott szempontokat használhatják,

így a kísérlet tervezésekor már ismerni kellene az összes lehetséges szempontot, ami a gyakorlatban kivitelezhetetlen.

Ez utóbbi problémára ad megoldást a sokdimenziós skálázás *(multidimensional scaling,* MDS), amikor a kísérleti személyeknek minden lehetséges ingerpárt össze kell hasonlítaniuk, de csak azt kell megmondaniuk, hogy a párok mennyire hasonlítanak egymásra (KREIMAN, VAN LANCKER-SIDTIS, GERRATT, 2005). A válaszok alapján egy *n*-dimenziós érzeti tér számítható ki, amiben minden egyes inger (hang) egy pont. Minél közelebb van egymáshoz két inger ebben a térben, annál hasonlóbbak. A tér dimenziói ortogonálisak, azaz a különböző dimenziók mentén mért értékek nem korrelálnak egymással. A dimenziók az egyéni hangok érzeti dimenziói, így ezek alapján az egyéni hangszínezetre lehet következtetni. A dimenziók értelmezése során korrelációt keresünk a dimenziók és az egyes ingerek más módszerrel (kísérlettel vagy akusztikai méréssekkel) kapott jellemzői között. Így az érzeti ismertetőjegyek és az egyéni akusztikai hangjellemzők közötti összefüggéseket mutathatjuk ki. Bár az értelmezés objektívebb, mint a faktoranalízis esetén, itt is nagymértékben heurisztikus és sokszor erőltetett.

Agysérülések

Szemben a viselkedéses kísérletek fekete doboz modelljével az agysérülések vizsgálata már agyi területek és funkciók kapcsolatára deríthet fényt. Ebben azt vizsgálják, hogy a balesetek, sérülések következtében kialakuló agyi léziók (szövetkárosodások) hogyan befolyásolják a kognitív funkciókat (KOVÁCS, 2006a), esetünkben a beszélőfelismerést. A módszer hátránya, hogy a léziók helyét a véletlen határozza meg és nagyon kevés kísérleti személy áll rendelkezésre.

A beszélőfelismerés szempontjából a legérdekesebbek a phonagnóziás betegek, akik nem ismerik fel családtagjaikat a hangjuk alapján (VAN LANCKER, CUMMINGS és munkatársai, 1988). A kutatók azt találták, hogy ez az állapot általában a jobb félteke egyes területeinek sérülésével jár együtt. A beszélőfelismerés és a beszédpercepció valószínűleg különböző agyi területeket vesz igénybe, mert vannak tökéletesen kommunikáló phonagnóziás és sértetlen beszélőfelismerési képességgel rendelkező afáziás (sérült nyelvi funkciókkal rendelkező) esetek is. A phonagnóziások képtelenek ismerős személyeket beszédük alapján felismerni, vannak azonban olyan betegek, akik nem tudják ismeretlen beszélők hangját megkülönböztetni egymástól. Ez a két rendellenesség együtt és külön is előfordul, ami arra utal, hogy az ismerős és ismeretlen beszélők felismerése legalább részben különálló folyamat. Az agysérüléses eredmények részletesebb áttekintése olvasható BELIN, FECTEAU és BÉDARD (2004) szemléjében. Egy *f*MRI-vel egybekötött kísérletben (VON KRIEGSTEIN, KLEINSCHMIDT, GIRAUD, 2005b) egy prosopagnóziás (arcfelismerésre képtelen, *arcvak)* embert vizsgáltak, hogy megállapítsák, milyen mértékben képes beszélőfelismerésre – ennek eredményeit a neurológiai háttérről szóló szakaszban ismertetjük.

Kiváltott válaszok

A módszer időben nagyon nagy felbontású (ms-os nagyságrendű), de téri felbontása meglehetősen durva (az elektródok cm-ekre vannak egymástól). Ennek megfelelően a kiváltott válaszokból (KV; vagy más néven eseményhez kötött potenciálok, EKP, angolul: *event-related potentials*, ERP) az agyi feldolgozási lépésekre az aktivitás időbeli lefolyásából lehet következtetni, a feldolgozóegységeket nem lehet elkülöníteni.

KAGANOVICH, FRANCIS és MELARA (2006) például egy Garner-féle szelektívfigyelem-feladat elvégzése közben rögzítették a kiváltott válaszokat. A két feladat során a kísérleti személyek ugyanazokat a magánhangzó-felvételeket hallották, de másra kellett figyelniük. Az egyik feladatban azt kellett kitalálniuk, hogy két magánhangzó közül melyiket hallották, míg a másikban a beszélő személyét kellett kiválasztaniuk a két lehetőség közül. Mindkét feladatot elvégezték úgy is, hogy az irreleváns dimenzió (magánhangzó-azonosításnál a beszélő személye, beszélőfelismerésnél a magánhangzó) végig állandó volt, és úgy is, hogy folyamatosan változott. Azt találták, hogy az utóbbi esetben a kísérleti személyek jóval lassabban válaszoltak, mint az előbbiben. Ez alapján arra következtettek, hogy a nyelvi és identitásinformáció feldolgozása nem független egymástól. A kiváltott válaszok időbeli lefutását összevetve az is kiderült, hogy a feldolgozás korai szakasza lehet közös.

Agyi képalkotó eljárások

A legelterjedtebb agyi képalkotó eljárás e területen is a funkcionális mágneses rezonancia szkennelés *(functional magnetic resonance imaging, f*MRI). Működésének alapja az, hogy egy agyi terület minél aktívabb, annál több oxigént igényel, így a területet behálózó hajszálerek oxigenált hemoglobinszintje megnő (KOVÁCS, 2006a). Erős mágneses térben az oxigenált hemoglobin máshogy rezonál, mint a dezoxigenált – ezt a különbséget, a BOLD *(Blood Oxigenation Level Dependent)* jelet, méri az *f*MRI. Az eredmény háromdimenziós képek sorozata a pillanatnyi agyi aktivitásokról, amit általában egy fiktív agyon színárnyalatokkal ábrázolnak. Mivel az érhálózat nagyon finoman behálózza az agyat, nagy térbeli felbontást lehet elérni (egy képpont mm^3-es nagyságrendű), viszont az oxigénszint növekedése az inger megjelenítése után csak jelentős késéssel áll be, így az időbeli felbontás alacsony (másodperces nagyságrendű).

A beszélőfelismerés *f*MRI-s vizsgálata során általában ismert vagy ismeretlen beszélők felvételeit játsszák le a kísérleti személyeknek a szkennelés közben, és megfigyelik, hogy melyik agyterület aktiválódik. Csak viszonylagos aktivitást van értelme mérni, tehát például ismert és ismeretlen beszélők hangja által kiváltott reakciók különbségét. A kísérleti személy a szkennelés alatt vagy csak passzívan hallgatja a felvételeket (BELIN, ZATORRE és munkatársai, 2000), vagy valamilyen feladatot végez, például egy célszemély hangját kell felismernie (VON KRIEGSTEIN, EGER és munkatársai, 2003). Ez utóbbi esetben az aktivitási térképeket össze lehet vetni a viselkedési mutatókkal (helyes válaszok aránya) és kideríthető, hogy a fel-

ismerési teljesítmény összefügg-e valamely terület aktivitásával (VON KRIEGSTEIN, EGER és munkatársai, 2003). További előny, hogy egy ügyesen kitalált feladat jobban ráirányítja a kísérleti személy figyelmét a felvétel vizsgált szempontjaira, mint a passzív hallgatás. Ugyanakkor fennáll a veszélye, hogy a szkennelt képeken nemcsak a beszélőfelismerés folyamata, hanem a válasz kiválasztásával és a gomb megnyomásával kapcsolatos agyi aktivitás is megjelenik.

Nancy KANWISHER, aki *f*MRI-vel elsőként határozta meg a fuziform arcterületet (FFA), egy ismeretterjesztő előadásában (2006) azt mondta, hogy „a viselkedéses kísérletekhez képest sokkal könnyebb *f*MRI-vel vizsgálni ezeket a folyamatokat, mert fekete doboz helyett bele tudok nézni az agyba – így néha úgy érzem, hogy csalok". Ez jól érzékelteti, hogy miért jelent áttörést a pszichológiai funkciók vizsgálatában az *f*MRI: amire a viselkedéses kísérletekkel csak indirekt módon lehetett következtetni (például, hogy két képességet nem ugyanaz az agyi funkció valósít meg), azt közvetlenül ki lehet mutatni. A „csalást" azonban inkább szerénysége mondatta Kanwisherrel, mint a módszer egyszerűsége: az egyes vizsgálatok megtervezésénél gyakran nagy nehézségeket okoznak a módszer korlátai.

A durva időfelbontás miatt nem megfelelőek a viselkedéses kísérletek és a kiváltott válaszok esetén alkalmazott „egyszerre egy inger" jellegű eljárások. Ha az agyi válasz a szkennelés ideje alatt nem áll fenn végig, akkor az időbeli átlagolás miatt a kapott képen elhalványodhat vagy eltűnhet. Annak érdekében, hogy ezt elkerüljék több másodpercig, akár fél percig tartó felvételsorozatot játszanak le a kísérleti személynek. Ahhoz, hogy következtetéseket vonhassunk le az előállított képekből, ezeknek a sorozatoknak valamilyen szempontból homogéneknek kell lenniük, például mindegyik felvételt ugyanaz az ismert vagy ismeretlen személy mondta be (WARREN, SCOTT és munkatársai, 2006).

Bár az ismert agyi képalkotó eljárások között az *f*MRI térbeli felbontása a legnagyobb, valószínű, hogy így is anatómiailag elkülönülő specializált központokat összemos. Például a temporális *hangterület f*MRI-vel egységesnek látszik az emberben, majmok cytoarchitekturális (sejtfelépítés) vizsgálata egymással együttműködő, de elkülönülő részterületeket mutatott ki (BELIN, FECTEAU, BÉDARD, 2004).

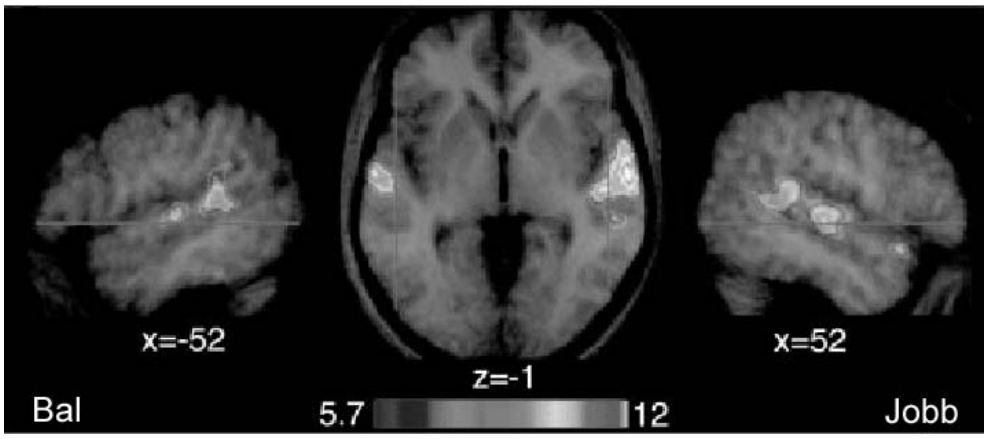
További probléma az *f*MRI-berendezés zaja. A viselkedéses kísérletekben mindent megtesznek, hogy a kísérleti személyeket semmilyen, a kísérlettől független auditív inger ne érje (általában némaszobában vagy csendes környezetben, nagyon jól szigetelő fejhallgatóval hallgattatják meg a stimulust). Ezzel szemben az *f*MRI-készülék működés közben hangos, kattogó zajt kelt. Az *f*MRI-vel végzett kísérleteknél így jóval hangosabb ingereket kell alkalmazni, és ellenőrizni kell, hogy a szkennelt személy képes-e kiszűrni a zajból a vizsgálat szempontjából érdekes hangokat (például egyszerű ellenőrző feladat elvégzésével).

Az *f*MRI talán legfontosabb korlátja az, hogy csak aktivitást lehet vele mérni: kijelöli, hogy mely területek működnek az adott pillanatban, de csak találgatni lehet, hogy azok mit csinálnak.

A BESZÉLŐFELISMERÉS NEUROLÓGIAI HÁTTERE

Ebben a részben olyan kutatások eredményeit foglaljuk össze, amelyek *f*MRI segítségével próbálták megismerni a beszélőfelismerés folyamatát és a hang-arc információ integrálását a személyfelismerés során. Egy kanadai és egy német kutatócsoport kísérletsorozatát ismertetjük.

Az FFA egy arcokra érzékeny terület az agyban: *f*MRI-vel kimutatták, hogy a fusiform agytekervénynek ez a része szelektíven erősebb aktivitást mutat arcokra, mint nem arc ingerekre (KOVÁCS, 2006b). Korábban már említettük, hogy az emberek hangja a „hallható arcuk" és – az arccal együtt – információkat hordoz gazdája kilétéről. Vajon van-e, az arcokhoz hasonlóan, a hangok feldolgozására is specializált terület az agyban? Ezt a kérdést először BELIN és ZATORRE vizsgálta munkatársaival (2000) *f*MRI segítségével, és arra jutott, hogy léteznek ilyen agyterületek: a felső halántékbarázdán *(superior temporal sulcus,* STS) helyezkednek el, többnyire a jobb féltekében *(2. ábra)*.



2. ábra. Agyi régiók, amelyek szignifikánsan nagyobb aktivitást mutattak emberi vokalizációkra, mint egyéb hangokra (csoportátlag). A legnagyobb aktivitást a jobb STS *hangterületei* mutatták *(Forrás:* BELIN, ZATORRE és munkatársai, 2000)

A kísérleti személyek mindhárom kísérletükben passzívan hallgattak emberi vokalizációkat és nem vokális hangokat – e szempont szerint voltak homogének az egy szkennelés alatt lejátszott hanganyagok. A kétféle blokk intenzitását kiegyenlítették. A vokális blokkok számos bemondótól, számos nyelven tartalmaztak beszédfelvételeket, és nem beszéd jellegű emberi vokalizációkat (például nevetés, sóhajtás, köhögés, állati hangutánzás). A nem vokális blokkok az első kísérletben vegyesen tartalmaztak a természetből és a civilizált környezetből származó hangokat, mint például állatkiáltások, vízzubogás, autók zaja vagy hangszeres zene. A második kísérletben négyféle nem vokális blokkot használtak: haranghangok felvételeit,

emberi nem vokális hangokat (például lépések, taps), emberi beszéd amplitúdóburkolójával modulált fehér zajt *(speech envelope noise,* SEN) és *összekevert beszéd*et. A SEN és az összekevert beszéd bár többnyire érthető, nem emberi eredetű hangnak halljuk. A harmadik kísérletben hasonló vokális és nem vokális felvételeket használtak, mint a másodikban, de úgy válogatták össze a hanganyagot, hogy az átlagos spektrum minden blokkra nagyjából megegyezzen. Ezenfelül minden blokkból készítettek két sávszűrt változatot is (az egyiket a beszéd szempontjából közepes, a másikat alacsony középfrekvenciával), és ezekre is lemérték az agyi aktivitást. A szkennelés után a kísérleti személyeknek egy viselkedéses kísérletben minden felvételről meg kellett állapítaniuk, hogy melyik vokális és melyik nem, valamint a beszélő nemét.

Mindhárom kísérletben a vokális blokkok alatt nagyobb aktivitást mértek, mint a nem vokálisok alatt. Az aktivitáskülönbség az STS-re, azon belül három helyre összpontosult az elülső, középső és hátsó részben *(2. ábra).* Megfordítva, a nem vokális ingerekre sehol sem mértek nagyobb aktivitást, mint a vokálisokra. A harmadik kísérletben az aktivitás a sávszűréstől függően csökkent, ugyanúgy ahogy a viselkedéses tesztek eredményei romlottak. Tehát az STS ezen területei nem beszédre, nem a nyelvi információra (SEN és összekevert beszéd), nem emberi eredetű hangokra (taps), nem egy kategóriára (harangok), nemcsak a magas vagy a mély frekvenciákra (sávszűrt felvételek), nem a felvétel intenzitására vagy spektrumára (mindkettő kiegyenlített a harmadik kísérletben) reagáltak, hanem specifikusan emberi vokalizációra. Az STS-régiók szelektíven nem nyelvi információt hordozó ingerekre aktivizálódtak.

Belin eredményei azonban nem zárnak ki minden kétséget az STS-területek hangspecifikusságával kapcsolatban. Ahogy az imént tárgyaltuk, számos alternatív magyarázatot kizártak a sokféle, különböző szempontokból homogén felvételblokkok használatával. Lehetséges azonban, hogy az általuk kiválasztott vokális blokkok nemcsak abban tértek el a többi felvételtől, hogy emberi hangot tartalmaztak, hanem más szempontból is. Például a harmadik kísérletben a vokális és nem vokális blokkok átlagos spektruma nagyjából megegyezik, ez azonban nem garantálja, hogy a két esetben a spektrum időbeli változásai hasonlóak. Lehet, hogy a beszéd viszonylag lassú (kb. 10 ms-os léptékű) változásához képest a nem vokális ingerek sokkal gyorsabban változnak. Számos egyéb hasonló kétely merülhet fel.

VON KRIEGSTEIN és szerzőtársai (2003; 2004; 2005a; 2005b; 2006) ezért más megközelítést választottak. Ugyanazokat a beszédfelvétel-sorozatokat kétszer is lejátszották a kísérleti személyeknek. Egy egyszerű felismerési feladattal az egyik alkalommal a nyelvi tartalomra, a másik alkalommal pedig a beszélő személyére irányították a hallgató figyelmét. Így az ugyanarra az ingerre készült két *f*MRI-képsorozat összehasonlításával úgy tudták kimutatni a nem nyelvi információfeldolgozásban részt vevő agyterületeket, hogy kiküszöbölték a Belin-féle módszerrel kapcsolatos aggályokat.

Ezt az elvet követve négy, egyre összetettebb vizsgálatot végeztek. Először csak kétféle ingert alkalmaztak: a hallgatók számára ismeretlen bemondóktól származó beszédet és kontrollként az ennek amplitúdóburkolójával generált SEN-t (VON KRIEGSTEIN, EGER és munkatársai, 2003). A beszédblokkok előtt a kísérleti szemé-

lyeknek lejátszottak egy célmondatot, majd a blokk mondatainak hallgatása alatt gombnyomással kellett jelezniük, ha ugyanazt a mondatot/beszélőt hallják (attól függően, hogy éppen a nyelvi tartalomra vagy a beszélőre szeretnék irányítani a figyelmet). A SEN-blokkok előtt egy kiválasztott mondat SEN-jét kellett hasonlóan felismerniük. Az eredmények megerősítették BELIN, ZATORRE és szerzőtársai (2000) következtetéseit: a jobb oldali STS elülső része szignifikánsan nagyobb aktivitást mutatott a beszélőfelismerési, mint a mondatfelismerési feladat során. Valószínűsíthető, hogy ez a terület fontos szerepet játszik a beszéd nem nyelvi jellemzőinek elemzésében. A fordított összehasonítás kimutatta, hogy a mondatfelismerés során pedig más, a nyelvi feldolgozással kapcsolatos területek voltak aktívabbak, mint a beszélőfelismerés során. Így a beszéd nyelvi és nem nyelvi elemzését feltételezhetően különálló agyi területek végzik.

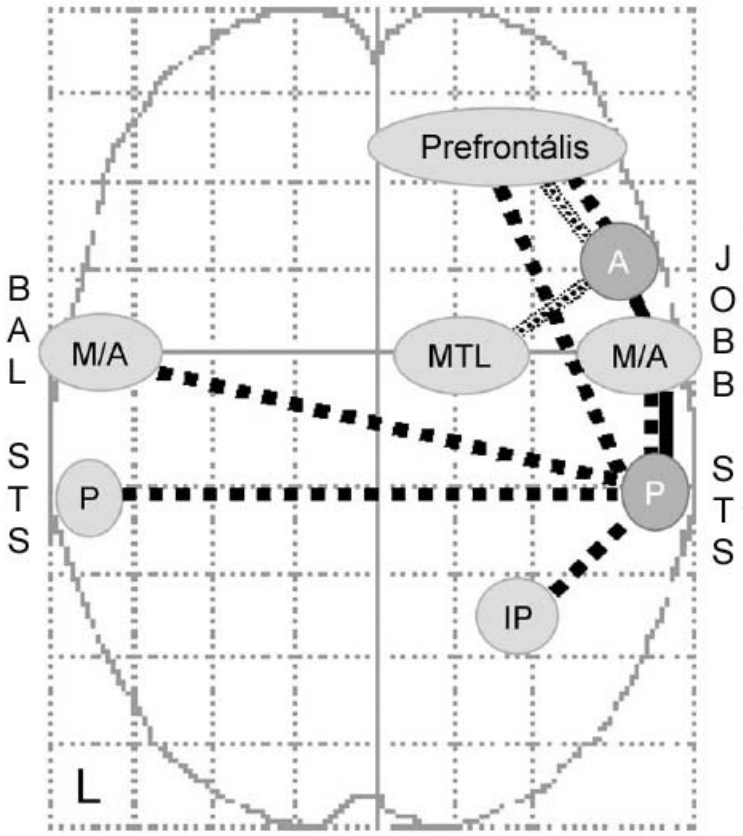
A két rendszer azonban nem független: mindkét feladat során aktívabbak voltak a szemantikai feldolgozóterületek, mint a modulált zaj (SEN) esetén. Bár ezek a területek a mondatfelismerés esetén voltak a legaktívabbak, jelentős működést mutattak ki a beszélőfelismerés esetén is. A szerzők ezt azzal magyarázzák, hogy a hallgatók a beszélőre irányított figyelem esetén is elvégezték a mondatok legalább részleges szemantikai elemzését, így a két funkció nem teljesen elválasztható. A mondatfelismerés során nem észleltek aktivitást a jobb elülső STS-ben, így ha a mondatra figyelünk, akkor nincs implicit beszélőfelismerés. Ennek egyik alternatív magyarázata lehet, hogy egyes, a beszélő személyére utaló ismertetőjegyek kinyeréséhez nem elég a pusztán akusztikai vizsgálat. A beszélő például jellegzetes szófordulatainak felismeréséhez nyelvi szintű elemzésre van szükség.

Megfigyelték továbbá, hogy a beszélőfelismerési feladat során az epizodikus memória-előhívásért felelős agyi terület *(precuneus)* erősebb BOLD-jelet adott, mint a másik két feladat során. Ismert, hogy ez a terület annál aktívabb, minél nagyobb erőfeszítést igényel az előhívás. Így a szerzők feltételezik, hogy ismeretlen hangokat nehezebb előhívni az epizodikus memóriából, mint egy mondatot vagy egy zaj jellegű hangot. Ezt támasztja alá az is, hogy ez alatt a feladat alatt volt a legalacsonyabb a helyes gombnyomások aránya. A zaj esetében a feladat egyszerűségét az adhatta, hogy magát a felismerendő felvételt játszották le a blokkok elején, míg a mondatok esetén más bemondóval, a bemondók esetén más nyelvi tartalommal felvett beszédet.

A második kísérletben ismerős bemondók hangfelvételeiből álló blokkokat is használtak (VON KRIEGSTEIN, GIRAUD, 2004). Így kimutatták, hogy a jobb STS-ben Belin által kimutatott három agyi hangterület különböző funkciókat lát el, de egymással szorosan együttműködik *(3. ábra)*. Az is nyilvánvalóvá vált, hogy az ismerős és ismeretlen beszélők felismerése részben eltérő agyi központok segítségével történik: míg az előbbi egy automatikus mechanizmus, az utóbbi részletes akusztikai elemzést igényel.

A beszélőfelismerési feladat alatt a jobb STS elülső és hátsó része aktívabb volt, mint a mondatfelismerés esetén. Az elülső rész ismert és ismeretlen személyek felismerése alatt is működött, de a másik két feladat alatt nem, így a beszélőfelismeréssel kapcsolatos általános feladata lehet. A hátsó rész mindhárom feladat alatt aktív volt, így feltételezhetően az összetett időszerkezettel rendelkező hangokra

reagál (a SEN-felismerési feladat csak az időszerkezet alapján oldható meg). A legerősebb aktivitást itt ismeretlen beszélők hangjára mérték, mert ezek feldolgozása részletesebb elemzést igényelhet, mint az ismert személyek beszéde. Ezt támasztja alá az is, hogy az ismeretlen beszélőfelismerés során hibáztak a legtöbbet a kísérleti személyek, míg ismertek esetén majdnem minden válaszuk helyes volt. Ezzel összhangban az összeköttetés-vizsgálatok azt mutatták, hogy ismeretlen hangok felismerése során az STS-területek számos más agyi területtel működtek együtt, míg ismert hangok azonosítása csak egy jóval kisebb hálózatot aktivált *(3. ábra)*. Mivel az ismert személyek felismerésére egyik STS-terület se mutatott nagyobb aktivitást, mint ismeretlenekre, viszont az STS-en kívül két terület igen, valószínűleg az ismert személyek reprezentációit az agy az STS-en kívül tárolja.



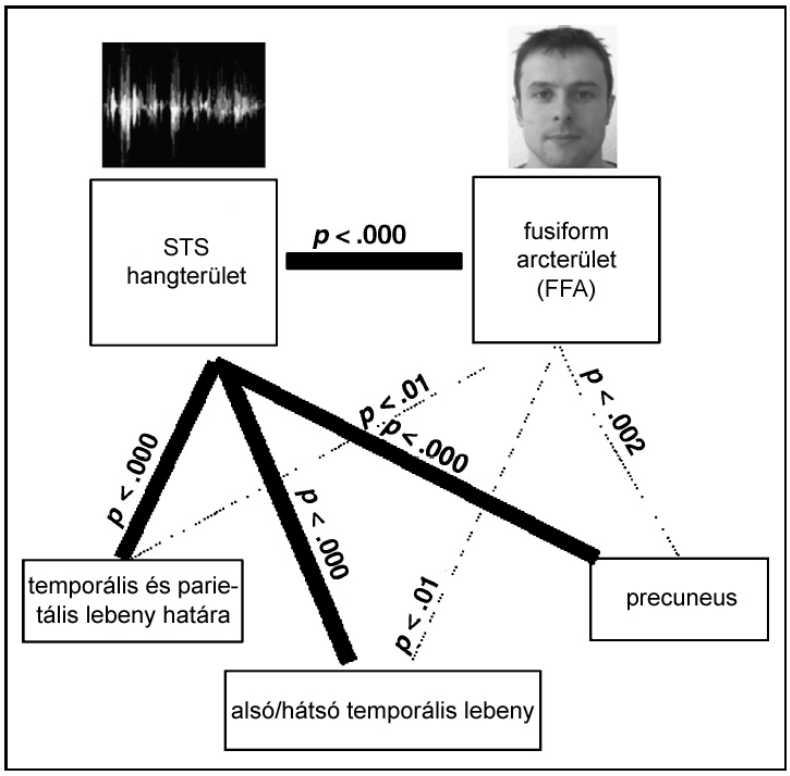
3. ábra. Az összeköttetés-vizsgálatok alapján ismeretlen beszélők felismerése (szaggatott vonal) sokkal kiterjedtebb agyi hálózatot aktivált, mint ismert beszélőké (pontvonal). A három STS-terület mindkét feladat során kapcsolatban állt egymással (folytonos vonal). *A:* elülső STS, *P:* hátsó STS, *M/A:* középső/elülső STS, *MTL:* mediális temporális lebeny, *IP:* alsó parietális. *(Forrás:* VON KRIEGSTEIN, GIRAUD, 2004)

Az összeköttetés-vizsgálatok azt is kimutatták, hogy a beszélőfelismerés alatt mindkét terület elsősorban a középső STS-sel működött együtt. Ez a terület mindegyik beszédblokkra reagált, de a SEN-blokkokra nem, így az akusztikai jelek spektrális feldolgozásában játszhat szerepet (a SEN spektruma mindig lapos, így nem hordoz információt).

Már az első kísérletükben megfigyelték, hogy egyes esetekben aktivitást mértek a látókérgen, pedig a blokkok között csak a hangfelvételekben volt különbség (egyszer se mutattak vizuális ingereket). A harmadik kísérletben ezért lokalizálták minden kísérleti személy FFA-ját, hogy a hang- és arcfeldolgozó területek együttműködését vizsgálhassák (VON KRIEGSTEIN, KLEINSCHMIDT és munkatársai, 2005a). Kimutatták, hogy az FFA nemcsak ismerős személyek arcára, hanem azok hangjára is reagál, de csak akkor, ha a beszélő személyére irányítják a figyelmet. A 2004-es cikkükben feltárt, ismerős hangok felismerését végző agyi hálózat része az FFA is. A korábbi modellekkel szemben azonban a két modalitás nem egy harmadik, szupramodális személyfelismerő központon keresztül működik együtt, hanem közvetlenül. Az összeköttetés-vizsgálatok alapján az STS hangterületei több más területtel is kommunikálnak (amelyek ilyen szupramodális személyfelismerő szerepet láthatnak el), de az FFA ezekkel nem állt kapcsolatban, kizárólag közvetlenül az STS hangterületekkel *(4. ábra)*. A szerzők szerint ez arra utal, hogy egy személy megismerésekor a hang- és arcinformáció összekapcsolódik és később a személy hangját hallva automatikusan elképzeljük az arcát is (korábban már bizonyították, hogy az FFA elképzelt arcokra is aktív).

Ugyanezt a kísérletet egy prosopagnóziás, azaz arcvak személlyel is elvégezték (VON KRIEGSTEIN, KLEINSCHMIDT, GIRAUD, 2005b). Bár a beteg FFA-ja nem válaszolt arcokra, ismerős személyek hangjára igen: ugyanolyan aktivitást mértek, mint a normál személyeknél és az STS–FFA-együttműködés is hasonló volt. A szupramodális személyfelismerő területek viszont – egészséges emberekhez képest – kevésbé voltak aktívak az ismert beszélőfelismerési feladat alatt. Ezzel összhangban a helyes gombnyomások aránya az ismerős beszélőfelismerési feladatban jóval alacsonyabb volt, míg a többi feladatban a különbség a többiek teljesítményétől nem volt szignifikáns. Ezek az adatok mind azt támasztják alá, hogy az arc- és hanginformáció integrálása a feldolgozás korai szakaszában történik (a felismerés előtt), és hozzájárul a felismerés sikeréhez. Az arcterület nem megfelelő működése esetén meghiúsulhat a személy felismerése. Ellentétben a közkeletű feltételezéssel, miszerint az arcvakok arc helyett gyakran hang alapján ismerik fel családtagjaikat, kimutatták, hogy az arcvakság a beszélőfelismerő képességet is korlátozza.

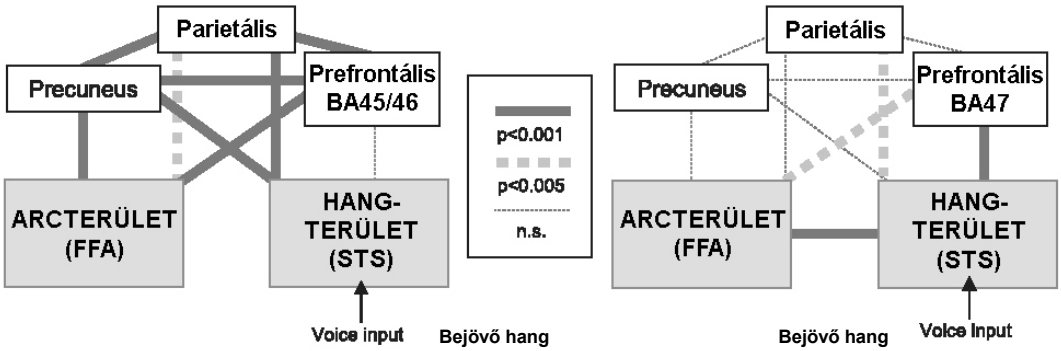
A negyedik kísérletben az előbbiekben kimutatott hang-arc asszociáció beszélőfelismerésben betöltött szerepét vizsgálták. A hallgatóknak megtanították az ismeretlen bemondók hangja és neve vagy arca közötti összerendelést, és korábbi módszertanuk szerint *f*MRI-felvételeket készítettek a tanulás előtt, alatt és után, mondat- és beszélőfelismerési feladat végzése közben (VON KRIEGSTEIN, GIRAUD, 2006). Kontrollként egyes kísérleti személyeknek csengőhangok és telefonok képe vagy márkája közötti asszociációt tanították meg, és ezekre is hasonló feladatokat kellett elvégezni.



4. ábra. Összeköttetés-vizsgálatok eredményei
(Forrás: VON KRIEGSTEIN, KLEINSCHMIDT és munkatársai, 2005a).

Azok a kísérleti személyek, akik egy személy hangját az arcával együtt tanulták, nagyobb arányban ismerték fel, mint azok, akik a nevével együtt tanulták. (SHEFFERT és OLSON [2004] viselkedéses kísérletükkel egy hasonló jelenséget már korábban kimutattak.) A hang-arc tanulók FFA-aktivitása a tanulás után jelentősen magasabb volt, mint előtte, míg a hang-név tanulóknak nem változott az arcterület-aktivitása. A csengőhangok esetén nem volt különbség a két csoport felismerési teljesítménye és FFA-aktivitása között. Tehát az arcterület beszélőfelismerés alatti működése nem magyarázható se szakértelemmel (csengőhangok), se az ismerős személyekhez kapcsolódó egyéb ismeretekkel (az „ismerős beszélőket" is a kísérletben hallották/látták először a kísérleti személyek).

A beszélőfelismerést végző rendszer első állomása az STS hangterületei. Az összeköttetés-vizsgálatok alapján azonban a rendszer további elemei jelentősen megváltoztak a hang-arc tanulás hatására. Míg a tanulás előtti beszélőfelismerési feladatban az STS számos más agyi területtel volt összeköttetésben, a tanulás után csak az FFA-val és a prefrontális kéreggel *(5. ábra)*. Az STS–FFA közvetlen kapcsolatban állt egymással, nem egy szupramodális központon keresztül kommunikált.



5. ábra. A vizsgált agyi területek közötti összeköttetések a beszélőfelismerési feladat alatt a hang-arc tanulás előtt (bal oldal) és után (jobb oldal).
(Forrás: VON KRIEGSTEIN, GIRAUD, 2006)

A negyedik kísérletből tehát látszik, hogy a beszélőfelismerés alatti FFA-aktivitás nem egyszerűen a felismerés mellékhatása, hanem annak fontos résztvevője, amely már szenzoros szinten bekapcsolódik a folyamatba. A szerzők szerint az FFA- és az STS-területek elosztott hang-arc reprezentációkat tárolnak, amelyek hatékony szenzoros predikciót tesznek lehetővé. Ezek a reprezentációk multimodális *Gestalt*ként működhetnek: még unimodális ingerre is együtt aktiválódnak. Az eredmények alapján ezeknek a multimodális reprezentációknak a személy neve nem része.

A BESZÉLŐFELISMERÉS PSZICHOLÓGIAI MODELLJEI

A tanulmányok többsége úgy ismertet eredményeket a beszélőfelismerési folyamattal kapcsolatban, hogy nem szól arról, milyen modellt feltételez. Ezek a kutatások olyan érzeti ismertetőjegyeket keresnek (mint például a jellemző hangmagasság), amelyek alapján a kísérleti személyek hatékonyan felismerik a beszélőket. Így általában – bár explicit módon nem írják le – egy jegyalapú modellt feltételeznek, ahol tetszőleges hallgató tetszőleges beszélőt fel tud ismerni egy rögzített jegyhalmaz vagy jegyköteg alapján. Ez a modell a beszédpercepció *megkülönböztető jegymodell*jére *(distinctive features;* lásd STEVENS, 2005) hasonlít.

A nyelvi információra kidolgozott jegymodell azonban ennél jóval finomabb: az összes lehetséges jegy közül egy fonéma felismeréséhez csak néhányat használ. Például egészen mások a jegyek mássalhangzók és magánhangzók esetén. Hasonlóan finomította a beszélőfelismerés fenti, merev jegymodelljét VAN LANCKER, KREIMAN és EMMOREY (1985a). Viselkedéses kísérleti eredményeik alapján úgy gondolták, hogy egy adott hallgató különböző jegyeket használhat a különböző beszélők felismerésére – például azokat, amelyek a legjellemzőbbek rá vagy ame-

lyek mentén a legegyedibb. Előfordulhat az is, hogy egy adott beszélőt különböző hallgatók különböző ismertetőjegyek alapján azonosítanak – azok szerint, amelyek alapján a hallgató által ismert beszélők halmazától leginkább eltér. Tehát a felismerés során alkalmazott ismertetőjegyek az adott beszélő-hallgató párostól függnek.

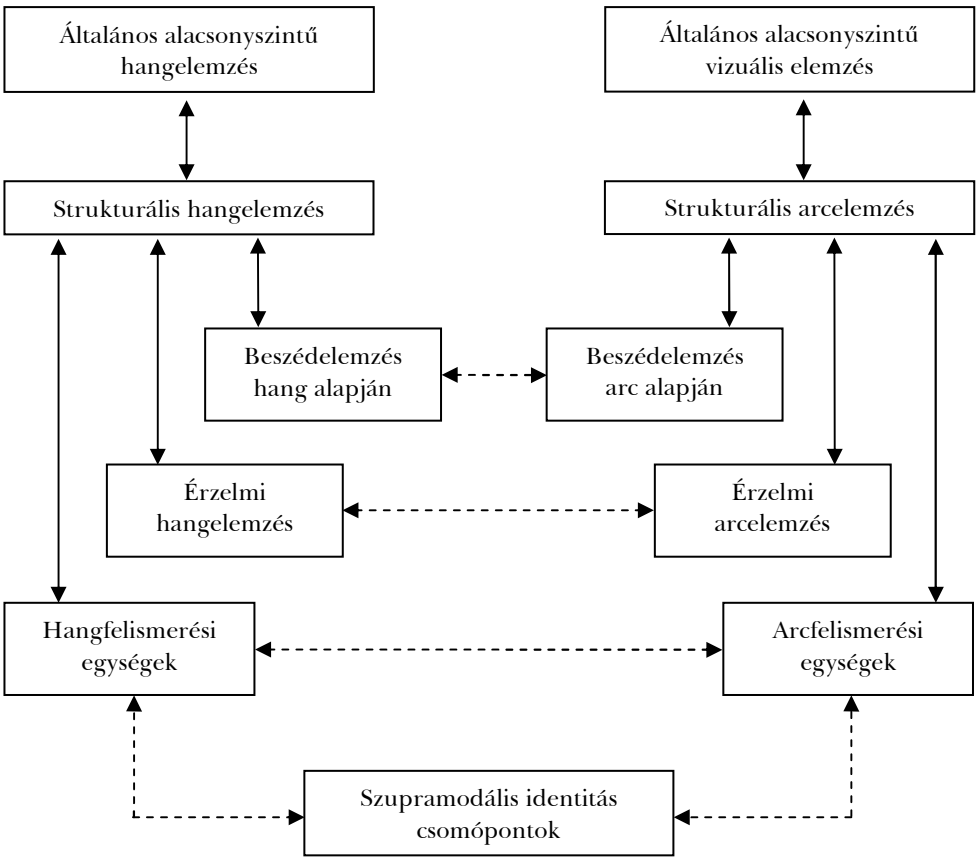
Ugyanennek a kutatócsoportnak egy későbbi cikkében (KREIMAN, PAPCUN, 1991) vezetik be a prototípus-modellt. Eszerint minden hallgatónak van egy prototípus-beszélője és az ismert hangokról csupán annyit tárol, hogy miben tér el a prototípustól. Azok a beszélők, akiknek hangja nagyon távol van a prototípustól, könnyen felismerhetőek, akiké viszont hasonlít a prototípusra, azok nehezen felismerhetőek. Idővel a hallgató az egyes eltéréseket elfelejti, így egy beszélőfelismerési feladatban válaszai a legtipikusabb beszélőkhöz konvergálnak. A prototípus lehet például az összes ismert hang átlaga vagy a hallgató saját hangja. A prototípus-modell könnyen összeegyeztethető Van Lancker rugalmas jegymodelljével.

VAN LANCKER, KREIMAN és EMMOREY (1985a) cikkükben még egy lépéssel tovább mennek: még egy adott beszélő-hallgató páros esetén se mindig ugyanazokon az ismertetőjegyeken alapul a felismerés. *Gestalt*-modelljük szerint az adott beszélő-hallgató pároshoz rendelkezésre álló jegyhalmaz redundáns, így elég azok közül csak néhányat detektálni a felismeréshez *(Gestalt*-zárás). Ez megmagyarázza a folyamat robusztusságát: még számos érzeti ismertetőjegy torzulása vagy kiesése esetén is felismerhető az ismerős személy.

A beszélőfelismerés a multimodális személyfelismerés – hang, arc, illat, név stb. alapján – egy speciális esete. Az arcfelismeréssel foglalkozó kutatók több modellt adtak multimodális felismerésre, amelyeket az *f*MRI-beszélőfelismerés szakirodalom átvett. BELIN, FECTEAU és BÉDARD (2004) például Bruce és Young arcfelismerési modelljét egészíti ki a hangmodalitással, amely egységbe foglalja a három információfolyam – nyelvi, érzelmi és identitás – feldolgozását *(6. ábra)*. A hangon az elsődleges hallókéregben végzett alacsonyszintű elemzés után magasabb szintű, strukturális elemzést végzünk. Ezután a háromféle információnak megfelelően három különböző rendszer párhuzamosan végzi a feldolgozást. A hangfelismerési egységek valószínűleg a jobb elülső STS-ben vannak, és minden ilyen egységet egy ismert beszélő hangja aktivál. A személyről összegyűjtött hang- és arcinformációt a szupramodális identitás csomópontok integrálják.

VON KRIEGSTEIN és GIRAUD (2006) Burton multimodális modelljét finomította arra az esetre, amikor pusztán a hangja alapján ismerünk fel valakit. Kísérleteikkel bebizonyították, hogy – szemben a korábbi elképzeléssel – ebben az esetben a hang- és arcfelismerő területek közvetlenül együttműködnek egymással, míg a név csak a felismerés melléktermékeként hívódik elő. Az arc és a hang mintegy multimodális *Gestalt*ként működik: csupán a hangot érzékelve agyunk lezárja a személysémát, azaz melléteszi az arcot, hogy a későbbi személyfelismerést ezzel segítse.

Bár számos tanulmányt publikáltak beszélőfelismerés témakörben (elsősorban viselkedéses kísérletek eredményeit), a felismerés alapját képező működési mechanizmus megértéséhez, modellezéséhez még sok kérdést kell megválaszolni. Ezeket a kérdéseket azonban a legtöbb szerző fel sem teszi.



6. ábra. Belin multimodális személyfelismerési modellje. A jobb oldalon látható, Bruce és Young-féle arcfelismerési modellt egészítette ki a baloldali hang-ággal. A szaggatott nyilak multimodális interakciókat jelölnek.
(Forrás: BELIN és munkatársai, 2004)

KITEKINTÉS

Az ismert személyek felismerése hangjuk (beszéd vagy más vokalizáció) alapján alapvető kognitív folyamat. A beszédpercepciónál bizonyos szempontból robusztusabb és jóval korábban alakul ki mind a törzsfejlődés, mind az egyedfejlődés során. Évtizedek óta kutatják viselkedéses kísérletekkel, azonban az utóbbi években az agyi képalkotó eljárások segítségével új fordulatot vett a terület vizsgálata. Azonosítani tudták az emberi hangra szelektíven érzékeny és a beszélőfelismerés során aktív területeket és rávilágítottak, hogy ismert és ismeretlen beszélők felismerése során más idegi hálózat lép működésbe. Sokszor hasznosnak bizonyult az *f*MRI és

viselkedéses eredmények összehasonlítása, mert így az input-output relációk összevethetők az agyi aktivitási térképekkel.

Az elért eredmények azonban még csak kezdetinek tekinthetők, mert a folyamat lényegét, működési mechanizmusát eddig nem sikerült felderíteni. Ennek elsődleges oka az, hogy mindkét vizsgálati módszer (viselkedéses, *f*MRI) alapján csak indirekt információkat kaphatunk a mögöttes kognitív folyamatról. A két megközelítés azonban jól kiegészíti egymást. Így talán a jövőben, ha a két iskola aktívabban együttműködik, eredményeik „összeérhetnek", és olyan általánosan elfogadott modellt adhatnak a beszélőfelismerésre, mint amilyen már régóta létezik a beszédpercepcióra. Egy ilyen modell hatékony gépi beszélőmódosító technológiák kifejlesztésére is alkalmazható lenne.

IRODALOM

Abberton, E., Fourcin, A. J. (1978) Intonation and speaker identification. *Language and Speech, 21,* 305–318.

Allen, J. S., Miller, J. L. (2004) Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115 (6),* 3171–3183.

Batliner, A., Steidl, S., Nöth, E. (2007) Laryngealizations and emotions: how many babushkas? *Proceedings of The International Workshop on Paralinguistic Speech*, August 3, Saarbrücken, 17–22.

Belin, P., Fecteau, S., Bédard, C. (2004) Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8 (3),* 129–135.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., Pike, B. (2000) Voice-selective areas in human auditory cortex. *Nature, 403*, 309–312.

Blouin, C., Maffiolo, V. (2005) A study on the automatic detection and characterization of emotion in a voice service context. *Proceedings of Interspeech 2005*, September 4–8, Lisbon, 469–472.

Bőhm, T., Shattuck-Hufnagel, S. (2007) Utterance-final glottalization as a cue for familiar speaker recognition. *Proceedings of Interspeech 2007*, August 27–31, Antwerpen, 2657–2660.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J. P., Reynolds, D. A., Magrin-Chagnolleau, I. (2003) Person authentication by voice: a need for caution. *Proceedings of Eurospeech 2003*, September 1–4, Geneva, 33–36.

Brown, B. L., Strong, W. J., Rencher, A. C. (1973) Perceptions of personality from speech: effects of manipulations of acoustical parameters. *Journal of the Acoustical Society of America, 54 (1),* 29–35.

Brown, B. L., Strong, W. J., Rencher, A. C. (1974) Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. *Journal of the Acoustical Society of America, 55 (2),* 313–318.

Campbell, J. P., Jr. (1997) Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85 (9)*, 1437–1462.

CHARRIER, I., MATHEVON, N., JOUVENTIN, P. (2001) Mother's voice recognition by seal pups. *Nature, 412,* 873.

FLETCHER, H. (1929) *Speech and Hearing*. D. Van Nostrand Company, New York

GOCSÁL Á. (1998) Életkorbecslés a beszélő hangja alapján. In Gósy M. (szerk.) *Beszédkutatás '98*. 122–134. MTA Nyelvtudományi Intézet, Budapest

GONZALEZ, J., OLIVER, J. C. (2005) Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *Journal of the Acoustical Society of America, 118 (1),* 461–470.

GORDOS G., TAKÁCS GY. (1983) *Digitális beszédfeldolgozás*. Műszaki Tankönyvkiadó, Budapest

GÓSY M. (1999) Az egyéni hangszínezet és a beszélő felismerésének kísérleti-fonetikai megközelítése. *Magyar Nyelvőr, 123 (4),* 424–438.

GÓSY M. (2001a) A testalkat és az életkor becslése a beszéd alapján. *Magyar Nyelvőr, 125 (4),* 478–487.

GÓSY M. (2001b) A genetikai tényező a beszélő személy felismerésében. In Andor J., Szűcs T., Terts I. (szerk.) *Színes eszmék nem alszanak: Szépe György 70. születésnapjára*. 423–431. Lingua Franca Csoport, Pécs

GÓSY M. (2004) *Fonetika, a beszéd tudománya*. Osiris Kiadó, Budapest

GÓSY M. (2005) Beszélőfelismerés: elmélet, kísérlet, gyakorlat. Előadás a *Beszédinformációs rendszerek* tárgyon, BME, Budapest. Előadásvázlat: http://speechlab.tmit.bme.hu

GÓSY M., NIKLÉCZY P. (1999) A beszélő felismerése a beszéde alapján: elméleti háttér és módszertani megközelítések. In Gósy M. (szerk.): *Beszédkutatás '99*. 1–18. MTA Nyelvtudományi Intézet, Budapest

HENTON, C. G., BLADON, A. (1987) Creak as a sociophonetic marker. In Hyman, L. M., Li, C. N. (eds) *Language, speech and mind: Studies in honour of Victoria A. Fromkin*. 3–29. Routledge, London

HOUSTON, D. M. (2005) Speech perception in infants. In Pisoni, D. B., Remez, R. E. (eds) *The Handbook of Speech Perception*. 417–448. Blackwell Publishing, Malden, MA

INSLEY, S. J. (2000) Long-term vocal recognition in the northern fur seal. *Nature, 406,* 404–405.

JOHNSON, K., LADEFOGED, P., LINDAU, M. (1993) Individual differences in vowel production. *Journal of the Acoustical Society of America, 94 (2/1),* 701–714.

KAGANOVICH, N., FRANCIS, A. L., MELARA, R. D. (2006) Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114,* 161–172.

KANWISHER, N. (2006) The brain basis of human vision. *Soap Box talk*, MIT Museum, April 23, 2006. http://mitworld.mit.edu/video/366

KERSTA, L. G. (1962) Voiceprint Identification. *Nature, 196,* 1253–1257.

KISILEVSKY, B. S., HAINS, S. M., J., HUI YE, H. (2003) Effects of experience on fetal voice recognition. *Psychological Science, 14,* 220–224.

KOVÁCS GY. (2006a) Az emberi agy és vizsgálati módszerei. In Kovács I., Szamarasz V. Z. (szerk.) *Látás, nyelv, emlékezet*. 11–25. Typotex, Budapest

KOVÁCS GY. (2006b) Halak, vadak, madarak és egyéb kategóriák az emberi agyban. In Kovács I., Szamarasz V. Z. (szerk) *Látás, nyelv, emlékezet*. 50–57. Typotex, Budapest

KREIMAN, J., PAPCUN, G. (1991) Comparing discrimination and recognition of unfamiliar voices. *Speech Communication, 10 (3),* 265–275.

KREIMAN, J., VAN LANCKER-SIDTIS, D., GERRATT, B. R. (2005) Perception of voice quality. In Pisoni, D.B., Remez, R.E. (eds) *The Handbook of Speech Perception*. 339–362. Blackwell Publishing, Malden, MA

LADEFOGED, P., BROADBENT, D. E. (1957) Information Conveyed by vowels. *Journal of the Acoustical Society of America, 29 (1),* 629–637.

LADEFOGED, P., LADEFOGED, J. (1980) The ability of listeners to identify voices. *UCLA Working Papers in Phonetics, 49*, 43–51.

NOLAN, F. (1980) *The Phonetic Bases of Speaker Recognition*. Cambridge University Press, Cambridge

NYGAARD, L. C., PISONI, D. B. (1998) Talker-specific learning in speech perception. *Perception & Psychophysics, 60 (3),* 355–376.

OWREN, M. J., CARDILLO, G. C. (2006) The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *Journal of the Acoustical Society of America, 119 (3),* 1727–1739.

ROSE, P. (2002) *Forensic Speaker Identification*. Taylor and Francis, London

SHEFFERT, S. M., OLSON, E. (2004) Audiovisual speech facilitates voice learning. *Perception and Psychophysics, 66 (2),* 352–362.

STEVENS, K. N. (2000) *Acoustic Phonetics*. MIT Press, Cambridge, MA

STEVENS, K. N. (2005) Features in speech perception and lexical access. In Pisoni, D. B., Remez, R. E. (eds) *The Handbook of Speech Perception*. 156–181. Blackwell Publishing, Malden, MA

VAN LANCKER, D. R., CUMMINGS, J. L., KREIMAN, J., DOBKIN, B. H. (1988) Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex, 24,* 195–209.

VAN LANCKER, D., KREIMAN, J., EMMOREY, K. (1985a) Familiar voice recognition: patterns and parameters; Part I. *Journal of Phonetics, 13*, 19–38.

VAN LANCKER, D., KREIMAN, J., WICKENS, T. D. (1985b) Familiar voice recognition: patterns and parameters; Part II. *Journal of Phonetics, 13*, 39–52.

VOIERS, W. D. (1964) Perceptual bases of speaker identity. *Journal of the Acoustical Society of America, 36 (6),* 1065–1073.

VON KRIEGSTEIN, K., EGER, E., KLEINSCHMIDT, A., GIRAUD, A.-L. (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*, 48–55.

VON KRIEGSTEIN, K., GIRAUD, A.-L. (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948–955.

VON KRIEGSTEIN, K., GIRAUD, A.-L. (2006) Implicit multisensory associations influence voice recognition. *PLoS Biology, 4 (10),* 1809–1820.

VON KRIEGSTEIN, K., KLEINSCHMIDT, A., GIRAUD, A.-L. (2005b) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex, 16 (9),* 1314–1322.

VON KRIEGSTEIN, K., KLEINSCHMIDT, A., STERZER, P., GIRAUD, A.-L. (2005a) Interaction of voice and face areas during speaker recognition. *Journal of Cognitive Neuroscience, 17 (3),* 367–376.

WARREN, J. D., SCOTT, S. K., PRICE, C. J., GRIFFITHS, T. D. (2006) Human brain mechanisms for the early analysis of voices. *NeuroImage, 31*, 1389–1397.

YARMEY, A. D., YARMEY, A. L., YARMEY, M. J., PARLIAMENT, L. (2001) Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*, 283–299.

SPEAKER RECOGNITION – NEUROLOGICAL BASIS AND PSYCHOLOGICAL MODELS

BŐHM, TAMÁS MIHÁLY

We are able to recognize our family members and friends solely by their voice. This is possible due to the speaker cues in nonlinguistic information that speech carries simultaneously with the linguistic message. This paper gives a partial review of the human speaker recognition literature in terms of the definitions, the applied methods, and the results. The latest experimental method, fMRI, opened a new avenue in the research of this cognitive process, thus we focus on studies employing this technique. We compare their results with those of behavioral studies and highlight their implications for speech technology. We also make an attempt to summarize the sporadically published psychological models of speaker recognition.

Key words: *speaker recognition, voice recognition, nonlinguistic information processing, voice area, STS, multimodal person recognition*