A BESZÉLŐ FELISMERÉSE A BESZÉDE ALAPJÁN: ELMÉLETI HÁTTÉR ÉS MÓDSZERTANI MEGKÖZELÍTÉSEK

Gósy Mária – Nikléczy Péter
MTA Nyelvtudományi Intézete

Bevezetés

Ha hallunk egy szót, annak akusztikai hullámformája a fülön keresztül a hallóközpontba jut, majd a Wernicke-területre kerül, ahol a hangsor, majd annak szemantikai tartalma feldolgozódik. Egyúttal azonban számos más döntéssorozat is történik agyunkban. Ha a szót egy számunkra jól ismert személy ejtette ki, akkor képesek vagyunk ezt a személyt azonosítani. Ez azonban nem ilyen egyszerű és bizonyos korlátozásokkal működik. Nem mindig elegendő egyetlen szó, hogy felismerjünk valakit, ugyanakkor sokszor a telefonvonal szűk frekvenciatartománya is lehetővé teszi, hogy beazonosítsuk, ki van a vonal másik végén.

A jelen tanulmány központi kérdése az, hogy vajon a beszéde alapján felismerhető-e a beszélő személy. Az előzőekben már érintettük, hogy a felelet erre egyetlen igennél vagy nemnél lényegesen összetettebb. Az elmúlt évtizedek alatt a fonetikával, illetve beszédakusztikával foglalkozó szakemberek alapos és kimerítő vizsgálatokat végeztek a beszédelemzés legtöbb területén (vö. Stevens 1998). Sikerült akusztikai elemzéssel mintegy rekonstruálni a beszédet, sőt – bizonyos korlátokkal – beszédfelismerő rendszerek is működnek. Azzal a ténnyel azonban, hogy az emberi hang magában rejti az egyéni jellemzőket is, csak az utóbbi 10-15 évben kezdtek behatóbban foglalkozni. A kiinduló kérdés az volt, hogy a beszéd szegmentális vagy szupraszegmentális részében keresendő-e az egyéni hangra utaló összetevő, vagy esetleg mindkettő tartalmazza azt.

A beszédet hallgató ember képes a beszélő személy kizárólag akusztikus úton történő felismerésére, még akkor is, ha a hozzá eljutó beszéd sokszorosan torzul. (Természetesen csak akkor, ha a torzulás

mértéke egy bizonyos határt nem ér el.) A torzulás történhet (i) spektrális szinten (például az alsó vagy felső összetevők intenzitásának csökkenése következtében), (ii) szupraszegmentális szinten (például a beszéd dallamában vagy tempójában történik változás), (iii) valamennyi szinten (a spektrum nagy részét elfedő zaj esetén).

A cél azoknak a feltételeknek a meghatározása, amelyek a) lehetővé teszik, b) korlátozzák és c) nem teszik lehetővé/gátolják a beszélő személyének azonosítását.

A „közvetett" kommunikáció nagymértékű fejlődése – mindenekelőtt a telefonálásra gondolva – idézte elő azt, hogy az emberi beszéd középpontba került mint az egyén egyik legfőbb önazonossága. Számos területen vált szükségessé a személy biztos azonosítása; a biztonsági rendszerekben, a beléptető rendszerekben, banki azonosító rendszerek és a kriminalisztikában. A tudománynak arra a kérdésre kell mindenekelőtt válaszolnia, hogy vajon a beszéd valóban olyan mértékben jellemző-e az egyénre, mint az ujjlenyomat. Amennyiben e kérdésre igenlő választ ad a tudomány, a következő kérdéssorozat a beszélő azonosításának feltételeit, az azonosítás módszertani megoldásait és az azonosítás biztonsági fokának meghatározását érinti. Az alkalmazott fonetika központi kérdése pedig az, hogy melyek azok a paraméterek, amelyek kétséget kizáróan felidézik a beszélő személyt, azaz megtörténik a beszélő személy azonosítása. Az alkalmazott fonetika új ága, az ún. törvényszéki fonetika ('forensic phonetics'), amely önálló diszciplínaként első ízben 1995-ben jelent meg a Fonetikai Világkongresszus programjában, foglalkozik a beszélőnek a beszéde alapján történő felismerésével.

Az új analizáló eljárások következtében a nem kriminalisztikai célú alkalmazásokban (például banki rendszerekben) a beszélő felismerésének problémája – még telefonon át is – megoldottnak tűnik. A hetvenes évektől indultak meg az erre irányuló kutatások és fejlesztések (pl. Doddington et al. 1976), mára többféle, megbízhatóan működő rendszer létezik a világban. Némelyikük állítólag 99%-os biztonsággal képes a beszélő személy azonosítására. A kutatók különböző algoritmusok alkalmazásával vagy különféle többcsatornás szűrő eljárásokkal igyekeztek meghatározni a beszélő személyazonosságát. A kidolgozott eljárásokkal sikerült – technikailag jó minőségű rögzített beszéd

esetében – 90% fölötti eredményt elérni a beszélő személy azonosságának meghatározásában, de a vizsgálathoz általában 40-50 s hosszúságú hanganyagra van szükség.

Ahhoz, hogy megértsük, miért mégis megoldatlan probléma a törvényszéki beszélőfelismerés; nézzük meg, mit jelent az egyén azonosítása a biztonsági rendszerekben. A beszélő valamilyen módon azonosítja önmagát (kóddal, névvel stb.), vagyis azonnal csökkenti a lehetséges beszélő személyek számát. A beszéd alapján történő személyfelismerésnek tehát arra kell válaszolnia, hogy valóban a feltételezett személy jelentkezett be. Egy többé-kevésbé meghatározott szöveget kell a beszélőnek bemondania (például szókapcsolatot, szókapcsolatokat vagy rövid mondatot). Általában az ún. normalizált, hosszú idejű átlagos spektrumelemzéssel, amelynek során az aktuálisan bemondott szöveg különféle jegyvektorait vetik össze a beszélőtől korábban tárolt szöveg paramétereivel. Ezt követően egy hasonlósági indexet számítanak. Az egyezést a küszöbértéktől való távolság szerint határozzák meg. Ezekben az esetekben tehát a beszélő felismerését számos tényező részben megkönnyíti, részben pedig kizárólagosan lehetővé teszi. A beszélő kooperatív, ez azt jelenti, hogy azt szeretné, hogy megtörténjen a biztos azonosítása. Létezik a beszélőtől már korábban tárolt, jó akusztikai és felvételi körülmények között rögzített beszédminta. Ismert az aktuális bejelentkezés körülménye, az összevetés tehát valóban gyorsan és jó hatásfokkal elvégezhető.

A kriminalisztikai vagy törvényszéki esetekben a helyzet lényegesen bonyolultabb és bizonytalanabb. A beszélő személy ismeretlen, következésképpen nincsen „tárolt" beszédminta. Jó esetnek számít, ha van gyanúsított vagy gyanúsítottak, ez kiindulást jelenthet a személyazonosításhoz. A feltételezett eredeti beszélőnek azonban ekkor nem célja, hogy természetesen, tisztán, megfelelő hangerővel beszéljen; az akusztikai-fonetikai összevetés tehát nehezedik. Mintegy 15%-ra tehető ezekben az esetekben, hogy a beszélő akaratlagosan megváltoztatja a beszédét (Künzel 1995). A leggyakoribb ilyen torzítások a suttogás, a megemelt hangfekvés és a zárt szájjal képzett beszéd. A rögzített beszéd rendszerint zajos, egy szűk frekvenciatartományban jelentkezik, a hasznos paraméterek tehát erősen csökkentett számban

vannak jelen (nemritkán csak 15-30 másodpercnyi anyag áll az elemző rendelkezésére).

A leglényegesebb különbség a kétféle beszélőazonosítás között a lehetséges beszélők számának különbsége. Az egyik esetben tulajdonképpen a beszélő személyének igazolása történik meg; a kriminalisztikai esetekben pedig valóságos azonosítás a cél. A beszélő azonosításához rendszerint háromféle megközelítésmódot használnak:

(i) hallás alapú elemzések (általában képzett szakemberek, elsősorban fonetikusok részvételével),

(ii) akusztikai-fonetikai analízis széles sávú spektrogramok alapján,

(iii) félautomatikus, speciálisan fejlesztett számítógépes elemző rendszerek alkalmazása.

A hallás alapján történő azonosítás tulajdonképpen percepciós tesztsorozat, amikor a hallgató a rögzített beszédet igyekszik a feltételezett személlyel azonosítani (a hallgató emlékezetében tárolt minta alapján). A beszélő személyt nem ismerő lehallgatók a feltételezett egyezéseket próbálják meghatározni a rövid idejű memóriában tárolt beszédminták összevetésével. Mindkét esetben előfordulhat olyan feladat is, amikor – kizárásos alapon – azt kell megmondani, hogy melyik az a beszélő, aki biztosan nem azonosítható az eredetivel. A szakemberek olyan kérdésekre is tudnak valószínű választ adni, mint a nyelvjárás lehetősége, beszédhiba, a szociális háttér, iskolázottság, becsléssel az életkor, a beszédbeli jártasság. A fonetikus és nem fonetikus hallgatók beszélőazonosítási eredménye között nagy különbség is lehet. Köster (1987) azt találta kísérletében, hogy míg a fonetikusok 100%-ot értek el, addig a nem-fonetikusok csak 89%, 94%-ot. Kollégák beszédének 30 másodperces részletei elegendőek voltak ahhoz, hogy tökéletesen azonosítható legyen a beszéd (Ladefoged 1978).

A beszélő felismerésének humán képessége

Az anyanyelv-elsajátítás folyamán kialakulnak azok a neurális spektrogramok az agyban, amelyek lehetővé teszik, hogy a gyermek a beszélő artikulációs sajátosságaitól függetlenül képes legyen a beszédhangokat azonosítani, a szavakat felismerni. Nem tudjuk még pontosan, hogy vajon ezek a neurális spektrogramok – mint ahogy

megnevezésük felveti – valóban hasonlatosak a beszédről készült akusztikai regisztrátumokkal, a spektrogramokkal. Annál is inkább, mivel a spektrogramok mindig egyediek, a neurális spektrogramok pedig szükségszerűen valamiféle általánosított képek kell, hogy legyenek. Feltételezhetően a hangsor(ok)ra szignifikánsan jellemző invariáns jegyeket tartalmaznak, amelyek egyúttal információval szolgálnak a beszélő személyére vonatkozóan is. Minél hosszabb az ugyanazon beszélőtől származó szöveg, a hallgató annál biztosabban képes a beszélőt felismerni. Ennek alapján az is feltételezhető, hogy a beszéd hallgatásakor aktiválódó neurális spektrogram-sorozatban valamiképpen hangsúlyozottabbá válnak a beszélőt azonosító paraméterek. Ezek a feltételezések vezettek a matematikai megoldások kereséséhez, amelyek azonban nem hozták meg a várt eredményt. Pontosan az a valami hiányzott, ami a legteljesebben emberi; az emberi agynak az a képessége, amelyet emlékezésnek nevezünk.

Az emlékezést úgy határozhatjuk meg, hogy olyan folyamat, amelynek során régebben észlelt tárgyak, jelenségek és események képét/képeit és ezek összefüggéseit idézzük fel anélkül, hogy az azokat létrehozó ingerek vagy ingeregyüttesek éppen hatnának ránk. Az emlékezés az objektív valóságnak a tudatban történő visszatükröződése. Az emlékképek a múltbeli észlelések, élmények reprodukciói. A beszélő személy felismerésére vonatkoztatva két dolog alapvetően fontos: szükséges a megfelelő inger, valamint a felidézés képessége. Az észleletek, feldolgozott ingerek megjegyzéséhez az szükséges, hogy létrejöjjön az emléknyom, amely az ismétlések során bevéső dik. Minél gyakoribb az ismétlődés, annál nagyobb mértékű a bevéső dés. Ha ritkán hallunk valakit beszélni, lassabban, nehezebben azonosítjuk a beszédet a beszélővel. Minél gyakoribb a beszéd akusztikai élménye, annál gyorsabb és biztosabb lesz a beszélő személy felismerése.

Az emléknyomok felidézése többféleképpen történhet, általában valamiféle asszociáció révén. A felidézés alapja az a kapcsolat, amely bizonyos fokig már a bevéső déskor jelen van. Az asszociáció az emlékezésben azt jelenti, hogy a kialakult szinoptikus kapcsolatok működése révén az egyik emléknyom aktiválása egy vagy több hozzá kapcsolódó emléknyomot is aktivál. A beszédre vonatkozóan általános összefüggések is megfogalmazhatók. Nem véletlen például az alap-

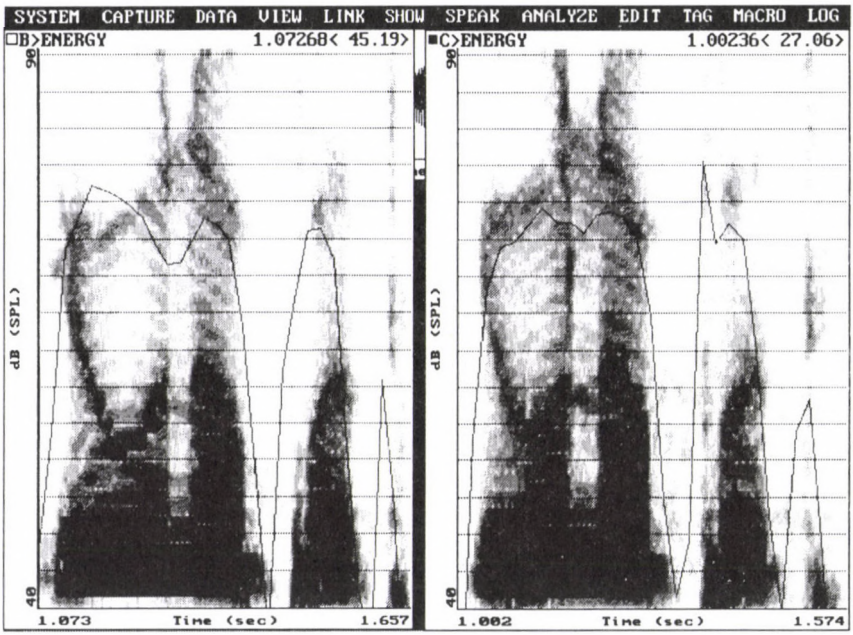
hangmagasság és a testalkat vagy a hangszínezet és az arcforma kapcsolata; ezek tudományos igényű kutatásáról azonban még alig beszélhetünk. A hallgatók asszociációs képessége a beszéd és a beszélő személyének felidézését illetően különböző. Vannak, akiknél gyorsan történik a bevéső dés, gyors a megfelelő neurális spektrogram aktiválása és ennek következtében a beszélő felismerése. Másoknál ezek a folyamatok lényegesen lassabban alakulnak ki, illetőleg mennek végbe.

A beszélő személy felismerésének tényezői

A fentiekben az alapvető feltételt – a beszélő személy ismertségének megfelelő szintjét – már tárgyaltuk. A következő, egzaktan nehezen megfogható, ám a pszicholingvisztikában jól ismert tényezőt veszszük számba, az 'elvárás' faktorát. Saját elvárásaink hatással vannak a beszélő személy sikeres felismerésére. Ha egy jól ismert személynek telefonálunk, rövid ideig egy hozzá hasonló hangú beszélőt is elfogadunk a kívánt beszélőül az elvárás miatt. Ha várjuk valakinek a hívását, azonnal felismerjük, ha megszólal. Ugyanennek a beszélőnek az azonosítása nehezebb, ha nem feltételeztük, hogy éppen az a személy fog telefonálni (Ladefoged 1978).

A beszélőre jellemző neurális spektrogram nyilvánvalóan tartalmazza mindazokat a nyelvi/beszédbeli tényezőket, amelyek alapján azonosítjuk a személyt. Amennyiben ezt nem kérdőjelezzük meg, akkor valójában mi okozza az egyénre jellemző akusztikai tulajdonságok műszeres kimutatásának nehézségét? Elsősorban az, hogy a beszédinformációt továbbító akusztikus rezgések a hangképző rendszer tehetetlensége következtében kvázistacionárius jellegűek. Ez azt jelenti, hogy a rezgések paraméterei általában korlátozott ideig tekinthetők állandónak. Az előbbiekből következik, hogy a beszéd közben létrehozott hangsorok nem ismételhetők meg még egyszer teljesen azonosan. Az 1. ábrán a „*Jó napot*" hangsor spektrogramja és hangsoron belüli intenzitásviszonyai láthatók ugyanazon személy ejtésében 1 nap eltéréssel. A lehető legjobb, torzításmentes megjelenítés érdekében a hangsort 50000 minta/s-os mintavételezési sebességgel digitalizáltuk és Hamming ablakfüggvényű 71 Hz-es szűrővel analizáltuk. Az ábra bal és jobb oldalának vizuális összehasonlítása alapján is megállapítható, hogy az időben később készült hangfelvételről regisztrátumán

(jobb oldali) a formánsok és az intenzitás értékei lényeges eltérést mutatnak a korábbi felvételről készült regisztrátumhoz (bal oldali) képest.



1. ábra
A „*Jó napot*" hangsor spektrogramja 0-3 kHz-es tartományban

A hatvanas, hetvenes évek nem túlzottan széleskörű kutatásai a beszédhangok akusztikai szerkezetében jelölték meg a meghatározó paramétereket. Elsősorban a magánhangzók harmadik formánsát gondolták ebből a szempontból jelentősnek, amelyről azóta egyértelműen bebizonyosodott, hogy nem is igazán jellemző és messze nem ele-

gendő az egyén azonosításához. Minden egyes akusztikai paramétert igyekeztek megvizsgálni azzal a nem titkolt céllal, hogy a mindent kizárót vagy legalábbis a legjellemzőbbet megtalálják. Ha azonban csak egy formánst nézünk is (jelen esetben a harmadikat), akkor is három, numerikusan kifejezhető adattal állunk szemben: a formáns frekvenciaértékével, sávszélességével és intenzitásával. Figyelembe véve azt az egyáltalán nem elhanyagolható tényt, hogy e három összetevő állandó változása a beszéd velejárója, akkor nehéz elméletileg is feltételezni azt a számértéket, amely az egyénre jellemző lehet. Ha pedig nem tudunk meghatározni egy vagy néhány konkrét frekvenciaértéket (maximum ±30 Hz eltéréssel), akkor a személyazonosítás számértékek alapján nem valószínűsíthető. Egyelőre még nem vettük figyelembe azt, hogy a formánsok értéke függ a hang hangkörnyezetétől is.

A beszélő felismeréséhez – spektrális elemzéssel – a megoldás a teljes frekvenciatartományban megjelenő beszédhang azonosítása lenne (80 Hz-től 16 000 Hz-ig). Tekintetbe kell venni azonban azt is, hogy a beszédhangok akusztikai elemzésével a beszédhangok határértékeit határozták meg. A beszédészlelési kísérletek eredményei pontosították (minden vizsgált nyelvben) azt, hogy a hallgató személy a kérdéses beszédhangot milyen frekvenciatartományban azonosítja biztosan. Nem egyetlen érték felel meg tehát egy hang valamely formánsának a percepcióban, hanem egy határérték, amely 100-200 Hz-es sávot is felölelhet. A formánsstruktúrában az F1 értéke általában szűkebb, mint az F2-é, míg az F3-é nagyobb az F2-énél. A formánseltolódás értéke a beszélő nemétől is függ. Női ejtésben az F2, F3 jobban változhat, mint férfi ejtésben. Ezek a formánsérték-változások ugyanazon beszélő ejtésében kevésbé térnek el egymástól, viszont azonos hangsorok különböző időpontban történő ejtésekor már 30-100 Hz-es eltérések is fennállhatnak. Létezik olyan kutatási eredmény is (Hollien 1977), amelyik nemcsak az F3 jelentőségét kérdőjelezi meg, hanem az egyéni hangszín akusztikai megfelelőjét a telefonsávon kívülre eső összetevőkben feltételezi.

A hetvenes évek végének kutatási eredményei szerint az alaphangmagasság majdnem elegendő kulcs az egyén hangjainak felismerésére (innentől már csak egy lépés magának, a személynek az azonosítására). Úgy tűnik azonban, hogy a pozitív eredménnyel zárult megkülön-

böztetési kísérletek hátterében inkább a hallgatók jól működő rövid idejű memóriája állt, semmint az F0 mint egyértelmű felismerési tényező (Doehring–Ross 1971). Azt gondolták, hogy a vokális traktus fontosabb a beszélő azonosításában, mint a larynx-forrás (vö. Hecker 1971). Ezek a laryngográfiás kísérletek is sikerrel zárultak; ismert személyek közül egy mondat alapján azonosították a kérdéses személyt. Valamennyi beszélő felismerése csak az F0 alapján azonban csak 60-70%-os eredményt hozott.

Az akusztikai elemzések döntően a spektrográfián alapszanak; a következő paramétereket vizsgálják (különböző nyelvekben): formáns-sávszélesség, központi formánsfrekvenciák, maximumpontok, a rés- és zárhangok zörejfrekvenciái, átmenetek és még valami, amit úgy neveznek, hogy 'sajátos spektrográfiás alakzat', de közelebbről nem meghatározható paraméter (Künzel 1995). Tekintetbe veendők még a beszédtempó, illetőleg az artikulációs sebesség, a hezitációs jelenségek és a dallammenet. A kutatók azonban egyetértenek abban, hogy a spektrogramok elemzése nem nyújt egyértelmű kulcsot a beszélő személy felismeréséhez. Az alapvető kiindulás mégis a beszéd akusztikuma. A Los Angelesben kifejlesztett beszélőazonosító rendszer (Nakasone–Melvin 1988) például 14 paramétert használ (az időtől a spektrumig). Ezzel a rendszerrel állítólag 98%-os pontosságot lehet elérni (a kísérletek 50 férfi beszélőtől származó beszédmintát tartalmazó adatbázison folytak).

A Hollien és munkatársai által kifejlesztett fonetikai alapú rendszer (SAUSI) olyan paramétereket használ az azonosításhoz, mint az F0, a csendes szünetek száma és hossza, a beszédtempó vagy a magánhangzók időtartama (Hollien 1990).

Több kísérletsorozatban vizsgálták a jelentés szerepét a beszélő felismerésében. Nem a nyelvi, stilisztikai sajátosságok tekintetében, a kérdés csupán az volt, hogy a szöveg érthetősége összefügg-e a beszélő személyének felismerésével. Az eredmények azt mutatták, hogy nem, az a beszélő azonosítása független a beszéd szemantikai sajátosságaitól (Janota 1967; Lariviere 1972; Schlichting–Sullivan 1998).

A leírtakból látható, hogy meglehetősen eltérőek a vélemények abban a tekintetben, hogy melyik a beszédnek az az összetevője, amelyik egyértelmű azonosítást tesz lehetővé. Az alaphangmagasság értéke, a

formánsfrekvenciák, a beszédhang mikrointonációs szerkezete, a beszédhangok egymáshoz viszonyított intenzitása, a beszéd időszerkezete mind-mind olyan paraméterek, amelyeket újra és újra meg kell vizsgálni az egyéni hangszínezet szempontjából. Azt vagy azokat a paramétereket kell megtalálnunk, amelyek mind a szegmentális, mind a szupraszegmentális szerkezetet tekintve, a legkisebb értékkel változnak, azaz közel állandó jelleggel reprezentálják a beszélő személy beszédét.

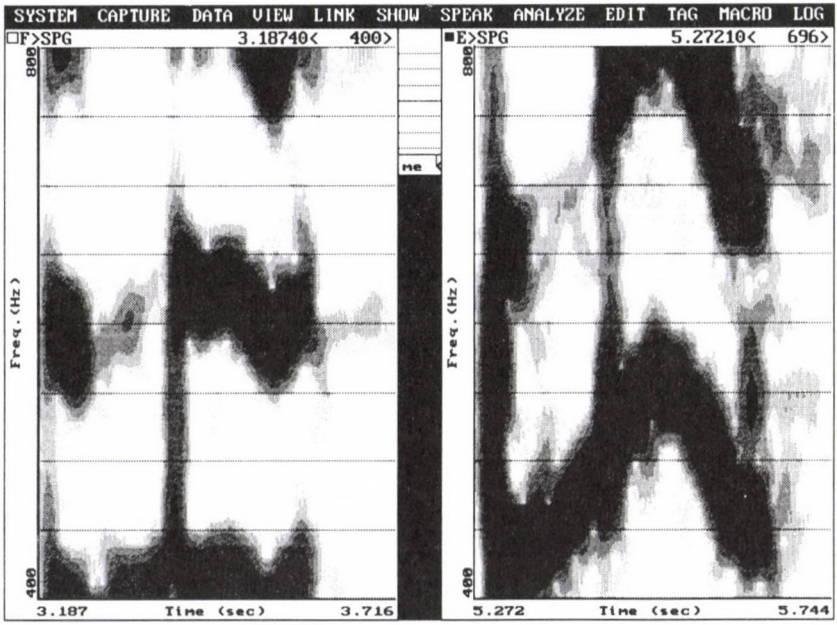
Láttuk, hogy nagy jelentősége van a beszédminták minőségének. A fizikai értelemben jó minőséggel rögzített minták összehasonlítását a beszéd teljes spektrumában el lehet végezni. Megemlítendő az az eset, amikor az összehasonlítandó hangfelvételek rossz jel/zaj viszonyúak, és a kérdéses felvétel nem egységes telefonhálózaton belül készült. A minőségen kívül fontos szerepe van a minták időtartamának, az egységnyi időtartam alatt elhangzó információnak, valamint a szöveg spontaneitásának.

Módszertani megközelítések

Elsődleges célunk az volt, hogy a már korábban említett, ún. „sajátos spektrográfiás alakzat" további tanulmányozásával, illetve kísérleti igazolásával olyan vizsgálati módszert dolgozzunk ki, amelynek segítségével 1. a korlátozott (rövid) időtartamú, rossz jel/zaj viszonnyal rendelkező beszéd vizuálisan és numerikusan is összehasonlító és 2. a kapott eredmények alapján az azonosítás nagy biztonsággal elvégezhető.

Az MTA Nyelvtudományi Intézetének Fonetikai Laboratóriumában olyan eljárást dolgoztunk ki, amely lehetővé teszi rossz technikai körülmények között rögzített hangfelvételeknél a személyazonosítást (azonos hangsorok esetében). A jelen tanulmányban ismertetendő elemző eljárásunk lényege az, hogy a hang spektrumából csak a felharmonikusok vizsgálatával foglalkozunk, függetlenül attól, hogy az adott tartományban van-e formáns vagy nincs. Mivel a felharmonikusok alacsonyabb frekvenciatartományban intenzívebben vannak jelen, ezért alkalmasabbak műszeres elemzésre. A 400 Hz és 800 Hz közötti tartományban lévő felhangok dinamikus változását és elhelyezkedésének numerikus értékét vesszük figyelembe. Megfelelően választott

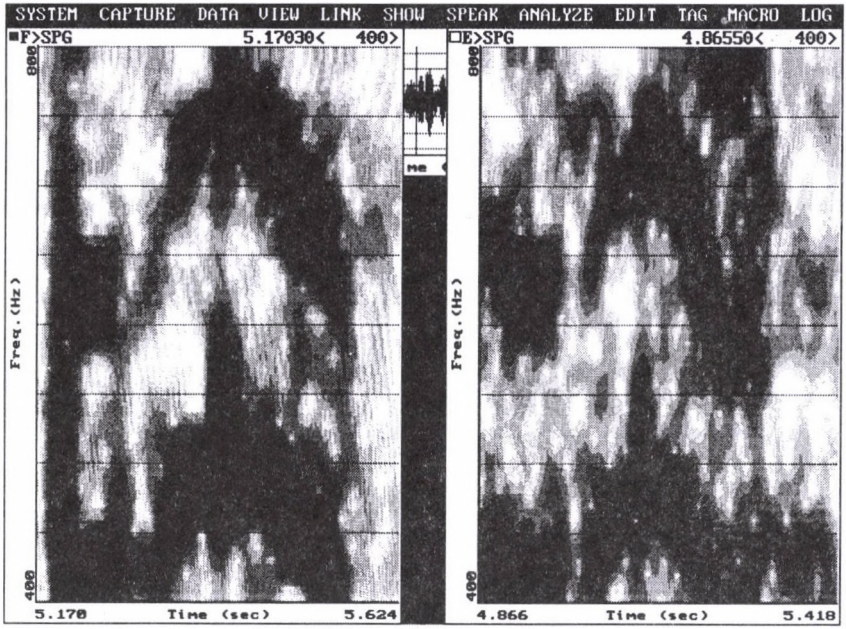
hangszelet (frame), illetőleg sávszélesség esetében a szegmentált részről készített hangszínkép alapján az összehasonlítás elvégezhető. A „*hogy ma este*" hangsor „*maes*" részletéről készült regisztrátumok illusztrálják a fentieket a 2. és 3. ábrán.



2. ábra
A *hogy maes* hangsor ejtése RM jelű (bal oldal) és SZM jelű személy (jobb oldal) esetében

A példában a személyazonosítás telefonvonalról, rossz minőségben rögzített hangfelvételből szegmentált közlésrészlet alapján történt. A kiválasztott rész időtartama 560 ms. A lehetséges beszélő személyektől az előre meghatározott szöveget stúdióminőségű felvételen rögzítettük, és a négy hangsorról készített spektrumot vizuálisan

elemeztük. Az RM jelű spektrogramról elmondhatjuk, hogy a felhangstruktúra dinamikus változása semmiképpen nem azonos az etalonfelvétellel. Az SZM jelű személy hangjáról készült spektrogram már mutat némi hasonlóságot, de numerikus értékei nem azonosak. A PH jelű személy azonban minden tekintetben azonosnak mondható a telefonról rögzített személy hangjával. (Megjegyezzük, hogy a vizsgálatban valójában 9 személy vett részt, de a többi részvevő hangszínképe lényeges eltérést mutatott az eredetihez képest, ezért ezekre itt nem tértünk ki).



3. ábra
A *hogy maes* hangsor ejtése PH jelű személytől (bal oldal) és a telefonról rögzített hangfelvétel

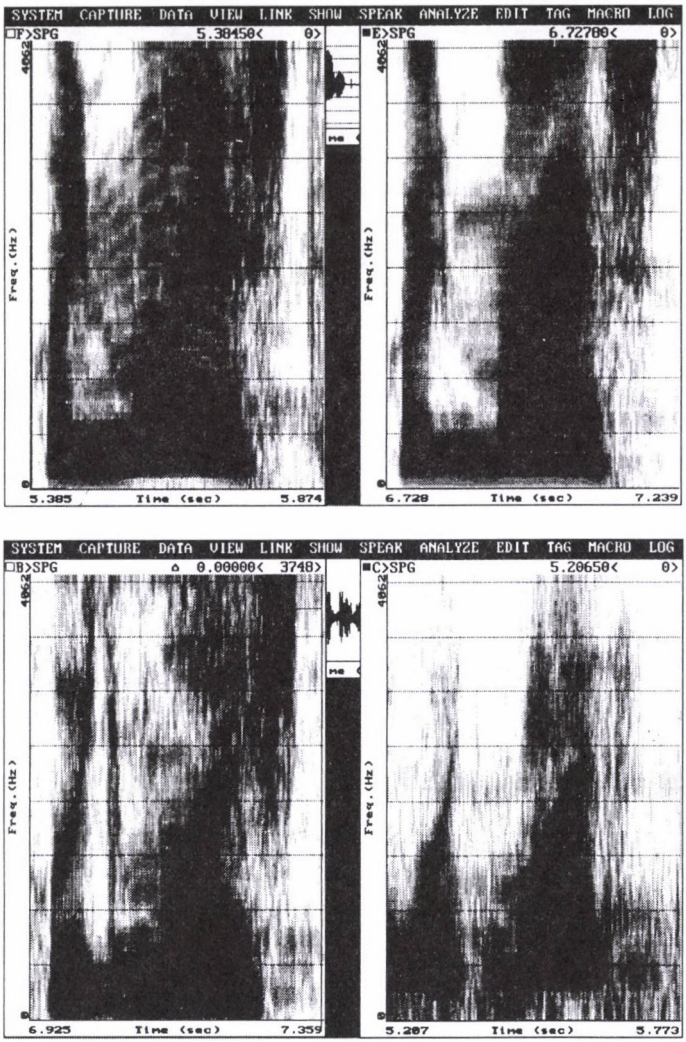
E módszer hátránya, hogy csak azonos hangsorok esetében alkalmazható. A szemléletesség kedvéért közöljük a hangsorok teljes spektrumát 5 kHz terjedelemig (4a, 4b. ábra). Az ábrákból jól látható, hogy az azonosítás ilyen megjelenítési formában nem végezhető el.

A felhangok változásának elemzésén alapuló eljárás magában hordozza a hibalehetőséget is. Abban az esetben, ha a rögzített szöveg – rövidsége miatt – nem tartalmaz hasonló beszédhangokat, hangkapcsolatokat, ez az azonosítást erősen megnehezíti. Ezért következő lépésként olyan módszerrel kísérleteztünk, ahol nem a beszélő személy által ejtett hangsorokat tekintjük kiindulási alapnak, hanem annak kisebb egységét, magát a hangot. A hosszan ejtett magánhangzó, illetve a zöngés mássalhangzó ugyanis bizonyos fokig magában foglalja a személy egyéni hangszínezetét is.

Feltételezésünk szerint a felhangoknak is hordozniuk kell a beszélő személyre jellemző egyéni sajátosságokat. A formánsstruktúrát leválasztva tehát, olyan felharmonikusokat keresünk, amelyek a legkevésbé esnek egybe (azaz megfelelő távolságban vannak) a formánshelylyel. A formánshelytől távol levő felhang ugyanis a kívánt felbontással elemezve, magában hordozza a hangszalagrezgés egy teljes periódusában bekövetkezett változást. Ez az eltérés pedig spektrálisan megjeleníthető. Az így megjelenített, megfelelő számú spektrum összehasonlításával, kialakítható egy olyan analizáló stratégia, amely rövid idő alatt nyújt értékelhető adatot, és jól reprezentálja a személy hangjának bizonyos sajátosságait.

Mindezek igazolására kísérletet végeztünk, ahol a beszédminta formáns- és felhangstruktúráját vizsgáltuk. A kiválasztott „tiszta" felhangot a jó melléknyaláb-csillapítás érdekében Hamming ablakfüggvényű keskeny sávú digitális szűréssel megtisztítva (a többi felhangtól), vetjük alá az összehasonlító spektrális vizsgálatnak. A cél annak megítélése, hogy az így kiválasztott felhang spektruma milyen képet ad a hang teljes időtartamában, illetve, hogy elegendő-e 8-10 periódus a vizsgálat céljára.

A kísérlet első fázisában 5 beszélővel végeztük el a vizsgálatokat úgy, hogy a kísérleti személyektől rendelkeztünk 23 évvel ezelőtt készített hangfelvételekkel.



4b. ábra

Az SZM és az RM jelű személyek ejtéséről készült teljes spektrum (fent); a PH jelű személy hangjáról (lent, bal oldal) és a telefonról készített hangfelvétel teljes spektruma (lent, jobb oldal)

Az öt személy hanganyagát CSL 4300B típusú digitális jelfeldolgozóval vizsgáltuk. A vizsgálat menete a következő volt:

1. A kiválasztott szöveget digitálisan rögzítettük. A rögzítés mintavételezési sebessége 50000 minta/s.

2. A bevételezett hanganyagból a kiválasztott hangot (ebben az esetben ez az [l] mássalhangzó volt) szegmentáltuk.

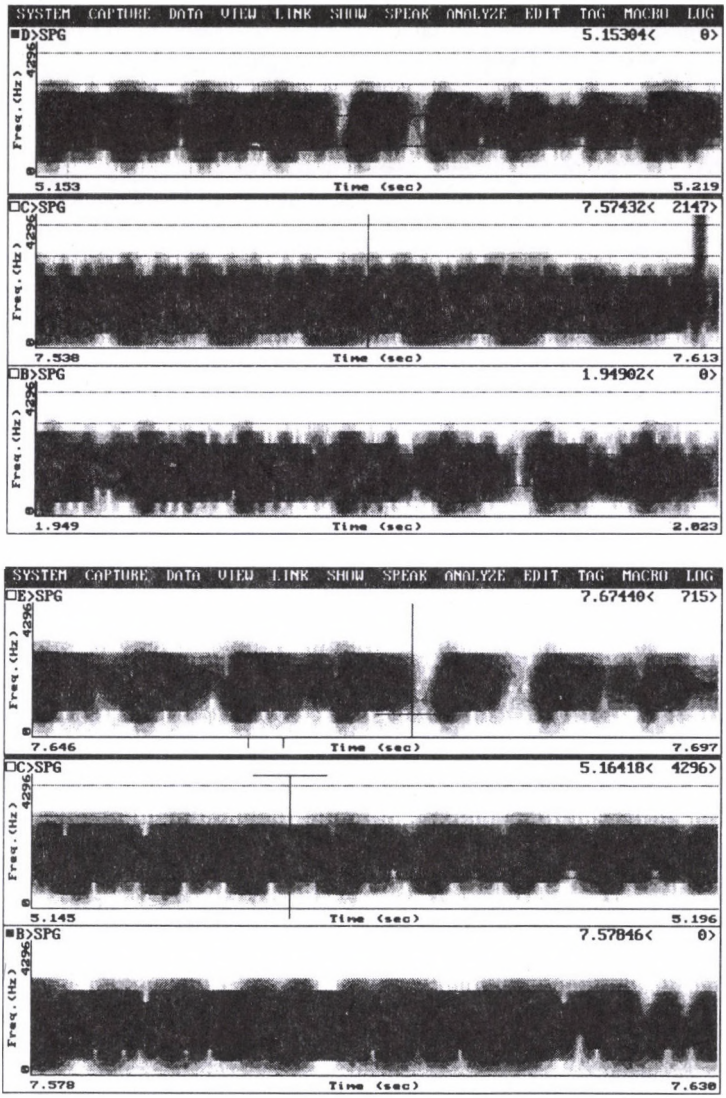
3. A megfelelő hosszúságú mintát (kb. 70 ms) keskeny sávú Hamming ablakfüggvénnyel 300 Hz sávszélességben szűrtük, majd az intenzitását többszörösen megnöveltük.

4. A keletkezett mintáról keskeny sávú spektrogramot készítettünk.

A spektrogramok összehasonlítása a következő eredményhez vezetett. A különböző személyektől származó zöngés mássalhangzók periódusonként más-más elhelyezkedésű intenzitásmaximumot mutatnak. A 6. ábrán látható, hogy a felső részben lévő spektrogramon a periódusok gömbszerű alakot vesznek fel, lefelé mutató nyúlvánnyal, a középső részben elnyújtott formát láthatunk, felfelé mutató nyúlványokkal. Az alsó részben látható forma hasonlít ugyan kissé a felsőhöz, de a gócok maximumpontjai inkább felfelé mutatnak. Az ábra 4. sorában az első sorban lévő személy hangja ismétlődik meg, de a két ejtés között 1 hét különbség van. Az 5. sorban, a kissé ellaposodó periódustól eltekintve, nagyjából azonos jelek láthatók. A 6. sor viszont periódusonként eltérő képet mutat lefelé mutató nyúlványokkal.

Következtetések

A beszélő személy azonosítása a beszéde alapján már bizonyos múltra tekinthet vissza a magyar szakirodalomban (Gordos–Takács 1983; Gósy 1996; Nikléczy 1996;), de rendszeres akusztikai-fonetikai és percepciós vizsgálata alig két éve indult meg. A munkálatok részlegesen ugyan támaszkodhatnak a nemzetközi szakirodalomban leírt eredményekre, azonban a nyelvspecifikusság ténye mindig új feladat elé állítja a kutatót. Az alábbiakban összegezzük azokat a megállapításokat, amelyek részben elméleti meggondolás, részben gyakorlati tapasztalat, illetőleg saját kísérleti eredményeink alapján már egyértelműen megfogalmazhatók. Ezek a megfogalmazható kijelentések nemegyszer sok-sok órás elemző munkán, számtalan adat sokféle feldolgozásán alapulnak (magukban foglalva a kutatás zsákutcáit is).



6b. ábra
Az [l] hang átlagosan nyolc periódusáról készült spektrografikus kép különböző személyek ejtésében

1. A beszéd akusztikuma jellemző a beszélőre, oly mértékben, hogy az akusztikai-fonetikai paraméterek alapján a beszélő azonosíthatóvá válik.
2. Az elméleti megállapítást a humán beszélőfelismerő képességünk is alátámasztja.
3. A beszéd alapján történő közel-objektív személyazonosítás számtalan tényező függvénye. Ezek részben külső faktorok (pl. a beszédrögzítési körülmények) és belsőnek tekinthetők (pl. a beszélő kooperációs készsége).
4. Többféle eljárás is célravezető lehet; az adott módszert mindig a beszélőfelismerés célja, a beszédminta és egyéb körülmények határozzák meg.
5. Pillanatnyilag nincs tudományos válasz arra vonatkozóan, hogy hány vagy mely paraméterek azok, amelyek az egyén felismerését kétséget kizáróan biztosítják. Nem zárható ki az, hogy az agyban tárolt neurális spektrogram aktiválása egészen különböző azoktól az akusztikai eljárásoktól, amelyek révén a beszéd egyéni jegyeit igyekszünk meghatározni.
6. Különféle aspektusú akusztikai-fonetikai és percepciós kísérletek (és rengeteg adatfeldolgozás, -tárolás és összegzés) szükségesek ahhoz, hogy a probléma megoldásához közelebb jussunk.

A továbbiakban a leírt eredmények alapján tervezzük a vizsgálataink és kísérleteink.

Irodalom

Doddington, G.R.–Helms, R.E.–Hydrick, B.M.: Speaker verification III. Texas Instruments Inc. Report for RDAC, Rome, New York 1976.

Doehring, D.G.–Ross, R.W.: Voice recognition by matching to sample. J. of Psycholinguistic Res. 1. 1972, 233-142.

Gordos Géza–Takács György: Digitális beszédfeldolgozás. Műszaki Könyvkiadó. Budapest 1983.

Gósy Mária: A beszéd akusztikai szerkezetének állandóságáról. In: Nyelv, nyelvész, társadalom. Emlékkönyv Szépe György 65. Születésnapjára barátaitól, kollégáitól, tanítványaitól. II. Szerk.: Terts István. Keraban Könyvkiadó. JPTE. Pécs 1996, 66-75.

Hecker, M.: Speaker recognition: an interpretative survey of the literature. A.S.H.A. Monogr. 16. Washington, D.C. 1971.

Hollien, H.: Speaker identification by long-term spectra under normal and distorted speech conditions. JASA 62. 1977, 975-980.

Hollien, H.: The Acoustics of Crime. Plenum Press. New York, London 1990.

Janota, P.: Personal characteristics of speech. Trans. Of the Czechoslovak Academy of Sciences – Social Sciences Series 77/1. 1967.

Künzel, H.J.: Field procedures in forensic speaker recognition. In: Windsor Lewis, J.: Studies in General and English Phonetics. Essays in Honour of Professor J. D. O'Connor. Routledge. London 1995, 68-85.

Ladefoged, P.: Expectation affects identification by listening. Language and Speech 21/4. 1978, 373-375.

Lariviere, C.: Acoustic and perceptual correlates to aural speaker identification. In: Rigault, A. (ed.): Proc. 7th ICPhS. The Hague 1972, 558-564.

Nakasone, H.–Melvin, C.: Computer assisted voice identification system. Proceedings IEEE-ASSP. 1988, 587-590.

Nikléczy Péter: Beszélő személy azonosítása szűk frekvenciás szavak alapján. In: Beszédkutatás '96. Szerk.: Gósy Mária. MTA Nyelvtudományi Intézete. Budapest 1996, 20-31.

Schlichting, F.–Sullivan, K.P.H.: Can voice imitation be detected in voice line-ups in a lnaguage unknown by the listeners? Phonum 6. 1998, 105-118.

Stevens, K.N.: Acoustic Phonetics. MIT Press. Cambridge, Mass. 1998.

A kutatás a T 025965 sz. OTKA-munkálat keretében folyt