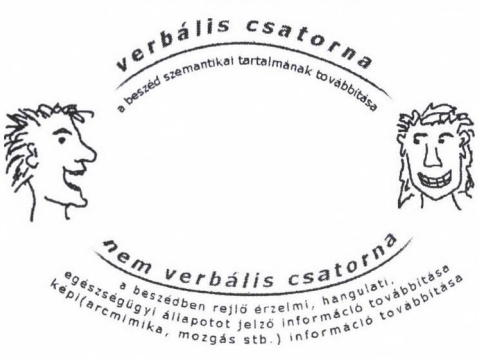
NEM VERBÁLIS HANGJELENSÉGEK SPONTÁN TÁRSALGÁSBAN

Vicsi Klára – Sztahó Dávid – Kiss Gábor

Bevezetés

Az emberi beszédkommunikációban a beszédinformáció feldolgozása két egymástól elkülönült módon történik. Az egyik feldolgozási mód esetében az üzenet nyelvi tartalmát dolgozzuk fel (verbális csatorna); a másik információfeldolgozási mód (a nem verbális csatorna) ahol a beszélő aktuális érzelmi, egészségi állapotát, hangulatát érzékeljük (Burkhardt et al. 2005). Az utóbbi évtizedben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna kutatása iránt az érdeklődés ez idáig kisebb volt, és működését kevésbé értjük.



1. ábra
Az emberi kommunikáció két egymástól elkülönült feldolgozási csatornája

Az emberi beszéddel a beszédtartalmon túl sok mást is ki lehet fejezni. A hangszínezet, az intonáció (hanglejtés), a ritmusváltozások mind széles körben használatosak arra, hogy a beszélő az érzelmi, hangulati vagy egészségi állapotát is a közlendő szöveg mellett, azzal egyidejűleg kifejezzék. Korábban a beszédtartalom vizsgálatakor rendszerint olvasott, vagy szépen megformált beszéd volt a vizsgálat alapja, viszont a beszédtechnológiai alkalma-

zásokban a valóságos spontán beszéd feldolgozása szükséges! Spontán társalgásban számos nem nyelvi elem fordul elő, amelyek hozzájárulnak ahhoz, hogy a beszélgető partnerek jobban megértsék egymást. A beszédkommunikációban a lelki állapot, az érzelem, az egyetértés vagy egyet nem értés közvetítése azt a célt szolgálja, hogy a beszélgető partnert informáljuk, még ha ezeket az információkat szavakkal nem is fejezünk ki a társalgás során. A spontán társalgás jelfeldolgozás szempontjából történő megismeréséhez elengedhetetlenül szükséges ezeknek a nem verbális jelenségeknek a kutatása. A BME TMIT Beszédakusztikai Laboratóriumban éppen ezért, ezeket a beszédben rejlő nem verbális információkat hordozó hangjelenségeket vizsgáljuk. Ezek a nem verbális hangjelenségek a következők:

1. Nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalom, amely prozódiai jellemzőkkel jut kifejezésre a beszédben a nyelvi tartalommal összefonódva. Ilyenek például a szomorúság, izgatottság, idegesség, vidámság, stb. vagy akár az egyetértés és az egyet nem értés prozódiai jellemzőkkel való kifejezése.

2. A nyelvi tartalomtól elhatárolt, attól független hangi események, amelyek további csoportokra bonthatók:

a) Jelentést kifejező hangjelenségek – ezek a hanggesztusok: például a sírás, a nevetés, a különböző érzelmet kifejező felkiáltások, a különböző jelentéstartalmú hümmögések (Markó 2005, 2006).

b) Jelentéssel nem rendelkező hangjelenségek:

– kitöltött szünetek

– egyéb hangjelenségek, mint pl. levegővétel, hangos nyelés, a krákogás, köhögés, egyéb testi hangok stb.

Mindezen hangesemények jelen vannak a spontán beszédben, és szerepük van az információátadásban. Megismerésük elengedhetetlen a természetes gépi beszéd előállítás és a gépi spontán beszéd felismerés megvalósításához.

Ebben a cikkben összefoglaljuk azokat a vizsgálatokat, amelyek a nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalomra vonatkoznak, azokra, amelyek prozódiai jellemzőkkel jutnak kifejezésre a beszédben a nyelvi tartalommal összefonódva. Továbbá csoportosítva tárgyaljuk azokat a nyelvi tartalomtól elhatárolt, attól független hangi jelenségeket, amelyek a spontán beszédben előfordulnak, és bemutatjuk az általunk létrehozott hang gesztustárat. Mindezen vizsgálatokhoz igen nagy mennyiségű spontán hanganyag gyűjtésére és feldolgozására volt szükség.

Módszer, adatbázisok

Vizsgálataink során 5 különböző spontán vagy közel spontán beszédadatbázist dolgoztunk fel, amelyeket magunk vettünk fel, vagy médiából gyűjtöttünk. Ezek az alábbiak:

a) Magyar Telefonos Ügyfélszolgálati Beszéd Adatbázis (MTÜBA): ügyfél és diszpécser beszélgetése került rögzítésre, az adatbázis 1100 ilyen felvételből áll (Vicsi–Sztahó 2009);

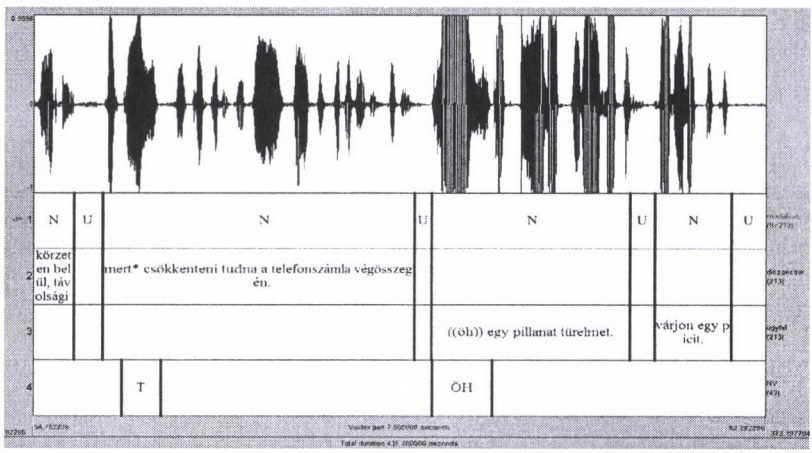
b) Maptask adatbázis: 1113 wav fájl, 10 különböző személlyel rögzített spontán beszéd útkeresés témában (Mády 2005);

c) Balázs-show felvételek: 135 percnyi műsoridő, 44 női és 99 férfi beszélő hanganyaga került feldolgozásra;

d) Joshi Bharatfelvételek: beszélgetős műsor, 61 percnyi műsoridővel;

e) mozi: egy spanyol *Torrente* című 3 részes akció-vígjáték magyar szinkronját használtuk fel gesztusok és egyéb nem verbális hangesemények kigyűjtésre. Ez a hanganyag feldolgozás csak ellenőrzésként került felhasználásra. Azt vizsgáltuk, vajon előfordul-e a filmben olyan nem verbális hangesemény, amit még az előző adatbázisokban nem találtunk.

A hanganyagok feldolgozása prozódiai frázisegységenként (Vicsi–Sztahó 2009) több szinten történt (2. ábra). (A prozódiai frázis értelmezése két szünet közötti beszédszakasz.) Első szinten frázisonként bejelöltük az adott frázisban kifejezésre jutó érzelmet. A következő szinten/eken a nyelvi tartalmat jelöltük beszélőnként külön-külön, ortografikus karakterekkel. Az utolsó szinten a szövegben már csillaggal jelzett helyeknél lévő hangi események időtartamát és típusát jegyeztük be.



2. ábra

Az adatbázisok többszintű feldolgozása.

1. Frázisonkénti érzelembejelölés (N: semleges, U:szünet); 2., 3. Nyelvi tartalom bejelölése (2: diszpécser, 3: ügyfél); 4. Nem verbális hangesemények (T: kitöltött szünet *t* hang után, ÖH: kitöltött szünet *öh*-t ejtve)

A hanganyagok annotálása a beszéd különböző gépi feldolgozási céljainak a figyelembevétele szerint készült. Koncepciójában különbözik a BEA (BEszélt nyelvi Adatbázis, Gósy 2008) és más adatbázisok szupraszegmentális vizsgálatok céljaira készült lejegyzésétől (Markó–Bóna 2006). Például a kitöltött szüneteket az azt követő frázis részeiként jelöltük abban az esetben, ha a kitöltés és a beszédkezdés közötti szünet szakasz kisebb, mint 250 ms.

Ezen adatbázisok vizsgálatával a társalgás során előforduló különböző nem verbális hangjelenségeket gyűjtöttük, amelyeket csoportosítottunk, és akusztikailag elemeztünk.

Nyelvi tartalommal együtt megjelenő érzelmi tartalom

Csak néhány éve kezdődött meg a beszéd különböző, nem verbális tartalmának, főként a hangulat kifejezésének, az érzelemnek a vizsgálata. Már korábban is érdekelte ez a kifejezési forma a kutatókat, de vizsgálataik során számos nehézségbe ütköztek, mivel a probléma igen összetett. A beszédben kifejezésre kerülő érzelmek vizsgálatának számos nehézsége van, melyek közül a leglényegesebbeket az alábbiakban soroljuk fel.

Statisztikai feldolgozáshoz elegendő érzelmet kifejező spontán beszédanyag gyűjtése nehéz. Az irodalomban található ugyan néhány kutatási leírás, amely a beszéd emóciótartalmának vizsgálatával, és az emóció automatikus, gépi felismerésével foglalkozik, de ezek az eredmények mind laboratóriumi körülmények között elhangzó tiszta beszédre vonatkoznak (Dougles-Cowie et al. 2003; Hozian–Kacic 2003; Campbell 2004, 2007). A publikációk legtöbbjében szimulált emóciótartalmú beszédet használnak, leggyakrabban művészek bemondásmintáit. A valós szituációkban elhangzó, spontán beszédre jellemző adatok jelentősen különböznek a színészek által produkált beszédtől (Kostoulas et al. 2007). A beszédtechnológiai alkalmazásokban a valóságos spontán beszéd feldolgozása szükséges. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával (Navas et al. 2006) és információtartalmainak felismerésével (Kohavi 1995) foglalkozik.

Az érzelemi megnyilvánulások vizsgálatánál problémát jelent a különböző érzelmi kategóriák változatos megjelenése. Az emóció jellemzésére kezdetben a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban hagyományos emóciókategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban (MPEG-4 1999) e kategóriákat az arcmimika jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták. A beszédtechnológiai szakemberek kezdetben ezeket a kategóriákat vették át a beszédben rejlő érzelem vizsgálatára is. Ha ezt összevetjük a valós helyzettel, az látszik, hogy a spontán beszédben sokkal változatosabb az érzelmi kategóriák tárháza, és ezek a téma szerint erősen változhatnak is, egyszerre két érzelem is kombinálódhat, és ezt az automatikus érzelemfeldolgozásnál is érdemes fi-

gyelembe venni (Laurence et al. 2005). Kutatási céllal a spontán beszédben leggyakrabban előforduló érzelmi kategóriákat gyűjtötték ki a PHYSTA 2001 adatbázisból (Cowie 2001; Nogueiras et al. 2001). Ez az adatbázis spontán társalgást, televíziós beszélgetőműsorok, és különböző vallási műsorok gyűjteményét tartalmazza (298 egység, 1 egység 10–60 s hosszú). A kiválasztott leggyakoribb érzelem és azok gyakorisága a 1. táblázatban látható.

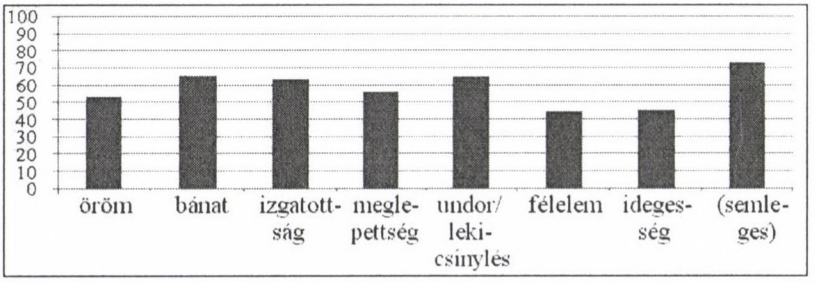
1. táblázat: Érzelmek csoportosítása és gyakoriságuk a PHYSTA 2001 spontán audiovizuális adatbázisban

Címke	Használati gyakoriság	Csoport
semleges	273	nem erősen érzelemvezérelt
dühös	114	erősen negatív
szomorú	94	erősen negatív
örvendező	44	nem orientáltan pozitív
boldog	37	nem orientáltan pozitív
jókedélyű	26	nem orientáltan pozitív
aggódó	19	erősen negatív
csalódott	17	nem erősen érzelemvezérelt
izgatott	17	orientáltan pozitív
félelem	13	erősen negatív
magabiztos	13	nem erősen érzelemvezérelt
érdeklődő	12	nem erősen érzelemvezérelt
gyengéd	1	orientáltan pozitív
elégedett	4	nem erősen érzelemvezérelt
szeretetteljes	3	orientáltan pozitív

További problémát jelent a beszédben kifejezésre kerülő érzelmek vizsgálatánál, hogy a szemantikus tartalom (verbális csatorna) és a beszélő hangulatának, általános érzelmi állapotának a tükröződése (nem verbális csatorna) egyazon beszédfolyamatban valósul meg, és a szemantikus tartalom hozzájárul a beszéd emóciótartalmának a felismeréséhez is. Nyelvi tartalom nélkül az emberi emóciófelismerés sem jobb, mint 60–65% a korábbi percepciós kutatások szerint (Tóth–Sztahó–Vicsi 2007). Az említett munkában ugyanazon szemantikai tartalmú mondatok különböző érzelmekkel kerültek bemondásra két csoportban, színészekkel és átlagemberekkel (3 mondat, mondatonként 8 érzelem, 15 személlyel).

Ezeket a mondatokat meghallgattatták érzelem szerinti megítélésre húsz személlyel. A szubjektív lehallgatás eredményeit a 3. ábra mutatja. A színészek és átlagemberek bemondásával kapott szubjektív lehallgatási eredmények között szignifikáns eltérés nem volt. A helyzetet tovább bonyolítja, hogy az érzelmeinket a kommunikáció során, több érzékszervi csatornán keresztül juttatjuk el a másik félhez, e csatornák közül a legjelentősebb, a beszédhang maga, és az arcmimika (de még a testbeszéd, bőrpír és egyéb té-

nyezők is szerepet játszhatnak az érzelem kifejezésében). Agyunk az összes érzékszervi csatornán keresztül kapott információ együtteséről dönt (Hozian–Kacic 2003). Például egyes érzelmeket hallva az ember maga sem tud különbséget tenni a két érzelem között, de látva az arckifejezést, már könnyebben dönt. Az is megfigyelhető, hogy az ember érzelem felismerési képessége csupán az arckifejezést látva meglepően jó. Az, hogy a hangi információ ad több információt vagy pedig a kép az érzelem felismeréséhez, az attól függ, hogy a hangi információban a nyelvi tartalom is benne van, vagy nincs. Amennyiben a hangi információ nyelvi tartalmat is ad, akkor csak hangi információ alapján lényegesen jobb a felismerés, mint csak az arckifejezés alapján. Ha viszont a hangi információ nyelvi tartalmat nem ad, pl. idegen nyelv esetén, akkor az arckifejezés alapján lesz jobb felismerés (Esposito 2009). A hang- és képinformációt kombinálva javul a legjobban a felismerés minősége, eddig az automatikus felismerésben a kutatóknak megközelítőleg 80% körüli felismerést sikerült elérniük a kombinált információ felhasználásával (Douglas-Cowie et al. 2003).



3. ábra
Az átlagemberek bemondásainak érzelmek szerinti felismerése (percepciós teszt eredményei) (Tóth–Sztahó–Vicsi 2007)

Továbbiakban célunk csak a hang alapján történő érzelem kifejezés jellemző paramétereinek a vizsgálata. A fenti felsorolt nehézségek talán magyarázatul szolgálnak arra, hogy az eddig elért kutatások, kizárólag hang alapján, 60% körüli gépi felismerést értek el legjobb esetben is (Cowie et al. 2001; Hozian–Kacic 2003; Campbell 2004; Burkhardt et al. 2005).

Beszédérzelmek jellemző vektorai a szakirodalomban

A gépi érzelem-felismerés során a meglévő hanganyagból jellemző vektorokat nyerünk ki, és ezeket használjuk fel az automatikus felismerő tanításához, majd ezekkel hajtjuk végre a felismerést. Ehhez persze tudni kell, hogy

mik azok a jellemzők, amelyek jól leírják az emberi beszéd érzelmi tartalmát. Tehát először a beszédérzelem jellemzőit kell definiálni, kategorizálni.

A beszéd semleges érzelem kifejezésekor is rendkívül változatos, két különböző személy ugyanazt a mondatot másképp ejti ki, továbbá ugyanazt a mondatot, ugyanaz a személy sem ejti kétszer ugyanúgy. A kiejtett hangok fizikai paraméterei függhetnek a beszélő egészségi, fizikai állapotától is (megfázás, stressz, fáradtság, különböző hangképző szervi megbetegedések). Mindezekhez hozzájárul még az a tény, hogy a beszélő a szándékától, érzelmi állapotától függően is változtathat egy mondat hangzásán, ezzel is kifejezve érzelmi állapotát. A beszédhang fizikai jellemzői tehát ugyanannál a szemantikai tartalomnál is sokfélék lehetnek.

Ez megnehezíti az érzelem gépi felismerését, hiszen meg kell tudnunk mondani, hogy mely változások játszanak fontos szerepet az érzelem kifejezésben, és melyek nem. A mai napig az ide vonatkozó szakirodalom egyik fő kérdése, hogy az automatikus érzelem felismeréshez milyen jellemzőket kell kigyűjteni, amelyek alapján majd a felismerés működni fog.

Az irodalomban összefoglalóan az alábbi érzelmekre jellemző fizikai paraméterekkel találkozhatunk (Seppänen et al. 2003; Álvarez et al. 2007).

Alapszintű adatok a jellemző vektorokban

Az úgynevezett alapszintű jellemzők közé tartoznak a keretenkénti alaphangfrekvencia-, a hangintenzitás-értékek, valamint a beszédhangok időtartamai.

Az alaphang erősen beszélőfüggő, személyenként és időben változó érték. Mégis az irodalomban érzelmet tükröző alapszintű jellemzőnek tekintik.

A beszédhangok intenzitása és annak deriváltja is fontos paraméter, kifejezi a nyomatékokat, a hangsúlyokat. A témával foglalkozó cikkek mind besorolják a vizsgálandó paraméterek közé.

A harmadik alacsonyszintű jellemző a szótagok, beszédhangok akusztikai időtartama. Ezek meghatározzák a beszéd tempóját, ritmusváltásait.

Származtatott adatok a jellemzővektorokban

A származtatott jellemzőket az alap szintűekből képezzük, azok valamilyen változását, statisztikáját tekintve, melyet jellemzően egy mondatnyi hosszúságú beszédre számítanak ki. A cikkek szerint ezek a származtatott jellemzők meghatározzák az egyén beszédének prozódiai jegyeit. Információt hordoznak az intonációról, a tempóról és a hangerőről. Ilyen származtatott jellemzők az alaphang és az intenzitás maximuma, minimuma, átlagértéke, deriváltja, értéktartománya egy hosszabb közlésre, például egy mondatra. Újabban már a színképi jellemzőket például a mel skálás frekvencia tartomány együtthatóit (MFCC együtthatók) is besorolják az érzelmek jellemző paraméterei közé (Cowie et al. 2001).

A származtatott jellemzők, amelyet az irodalomban mondategységekre számítottak ki, folyamatos spontán beszédben nem vezettek eredményre, mivel a hosszabb összetett mondat szerkezete függvényében a mondat más-más

részében jelenik meg az érzelem kifejezése. Éppen ezért, a legújabb kutatások szerint (Vicsi–Sztahó 2009) az érzelem kifejezésének alapegységeként a frázist tekintjük. Amennyiben frázisonként vizsgáljuk az érzelmek kifejezését, akkor nagyobb részben már ki tudjuk küszöbölni a mondat szerkezetétől való függést, ugyanakkor a frázis, már elég hosszú beszédegység ahhoz, hogy érzelmet tükrözhessen. A kérdés tehát az, hogy milyen fizikai paraméterek és azok milyen kombinációi tükrözik az egyes érzelmeket a frázisokban.

Beszédérzelmek jellemző vektorai frázisokban

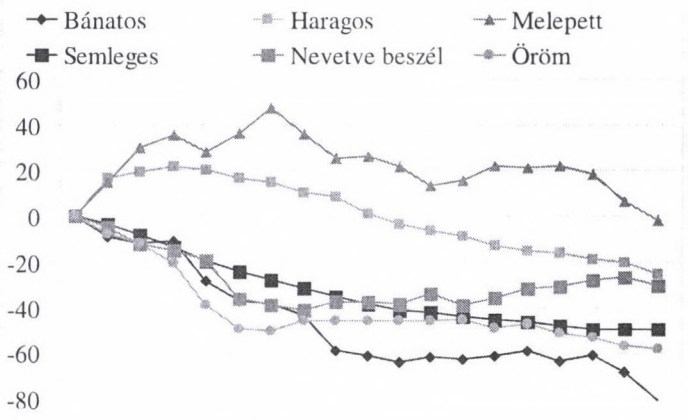
Jellemző vektorok vizsgálatát az összegyűjtött 5 különböző adatbázis felhasználásával végeztük el. Ezeknek az adatbázisoknak a feldolgozása során már kiderült, hogy az alapérzelmek (boldogság, szomorúság, düh, meglepetés, undor) jelölése sem egyértelmű feladat, és rendszerint a címkézőt a döntésben a szövegkörnyezet nagymértékben befolyásolja. Amennyiben azokat a prozódiai jellemzővektorokat akarjuk meghatározni, amelyek az érzelmi, hangulati tartalmat hordozzák a beszédben a nyelvi tartalom nélkül, akkor olyan mintákat kell elemeznünk, amelyek biztosan hordoznak ilyen információt. Az elemzéshez szükséges ilyen minták kiválasztása a szövegtartalomból kiragadott frázisok szubjektív lehallgatásával történt. (20 egyetemi hallgató, férfiak, nők vegyesen). Azokat a frázisokat tartottuk meg a további vizsgálatokhoz, amelyek esetében a hallgatók legalább 70%-a egy adott érzelemre ítélt. Így spontán 43 beszélő 1000 frázisát választottuk ki és osztottuk be hat különböző érzelemi kategóriába: a semleges, bánatos, haragos-ideges, meglepett, nevetve beszélő, örömöt kifejező. Az alap szintű jellemzőket vizsgálva a kiválasztott hanganyagon, az volt a tapasztalat, hogy az alapfrekvencia, és az intenzitás időbeli változása egy frázison belül jellemző a különböző érzelmekre.

A vizsgálati anyagban a különböző hosszúságú frázisokat lineárisan vetemítettük, hogy mindegyik minta „n" hosszúságú legyen, majd a mért adatokat normalizáltuk a frázisban mért első átlagadat értékére úgy, hogy a mintavételezési pontoknál mért adatokból az első minta értékét levontuk. Végül az érzelem szerinti csoportok frázisonkénti értékeit átlagoltuk, vagyis minden érzelemre elkészült az adott érzelemre jellemző átlagos hangminta-dinamika: mind alapfrekvenciában, mind összintenzitásban.

Az átlagos alapfrekvencia-dinamikája n = 19 értékek esetén a 4. ábrán láthatók, ahol az alapfrekvencia szórás értékei 5–10 Hz közötti értékeknek adódtak. Az átlagos alapfrekvencia-dinamika érzelem szerint szépen elkülönül az alábbiak szerint.

Bánatos beszédben az alapfrekvencia folyamatos és nagymértékű csökkenését figyelhetjük meg, majd körülbelül a frázis felénél, 60 Hz-es csökkenés után egy stagnálást, a végén újabb csökkenést. Haragos érzelem esetén az elején nő az alapfrekvencia, majd folyamatosan csökken. A meglepettség hatására az elején nagymértékű alapfrekvencia-növekedés látható, majd némi csökkenés. Ennél az érzelemkategóriánál figyelhető meg leginkább az alap-

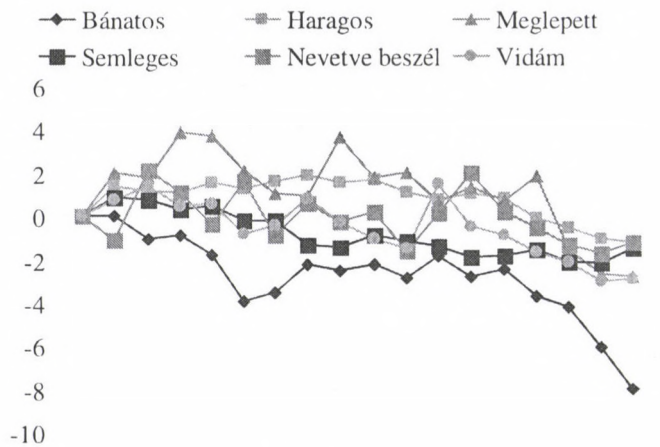
frekvencia növekedése. A semleges mintákban az alapfrekvencia folyamatos szabályos csökkenése figyelhető meg, bár annak mértéke nem igazán jelentős. Amikor a beszélő nevetve beszél, az alapfrekvencia csökkenése, majd körülbelül a frázis felétől, alacsony növekedése jellemzi. Öröm esetében az elején az alapfrekvencia lényeges csökkenése figyelhető meg a frázis felétől körülbelül 50 Hz, majd utána stagnál, igen hasonlóan a nevetve beszél kategóriához. Tehát a kísérlet alapján kijelenthető, hogy egy frázison belül az alapfrekvencia dinamikája jól jellemzi az érzelmeket. A kísérlet tanulsága szerint alapvetően az egyes érzelemkategóriák átlagos intenzitásdinamikái nem különülnek el olyan szépen, mint az alapfrekvencia változásának esetében, amint ez az 5. ábra alapján látható. Itt az értékeket nem az első mintavételezési helytől ábrázoljuk, hanem a másodiktól, emiatt az utolsó mintavételezési hely sorszáma a 18-as.



4. ábra
A különböző érzelmek „átlagos alapfrekvencia dinamikája"
(A vízszintes tengelyen a mintavételezési pontok láthatók.)

A szórás értékek körülbelül 3dB értékűek voltak. Ez itt relatíve magas érték. Amit érdemes megfigyelni, az az, hogy a „bánatos" érzelemnél jól látható és a többi érzelemtől elkülönült az intenzitás csökkenése, stagnálása, majd újabb csökkenése, illetve a „haragos" érzelemnél az intenzitás növekedése körülbelül a frázis feléig. A „semleges" érzelemnél az elején kicsi növekedés figyelhető meg, majd az érték folyamatos csökkenése. A „nevetve beszél" és a „vidám" érzelmeknél, az intenzitás folyamatos változása figyelhető meg. Az intenzitás értékek kevésbé tükrözik a különböző érzelmeket, bár azért jellemző dinamika jegyek az intenzitásnál is fellelhetők. Érzelmekre jellemző

lényeges színképi változás az idő függvényében a frázison belül nem tapasztalható, ugyanakkor egy frázisra átlagolt színképi paraméterek már érzelemre jellemző eltéréseket mutatnak.



5. ábra
A különböző érzelmek „átlagos intenzitás dinamikája".
(A vízszintes tengelyen a mintavételezési pontok láthatók.)

Összefoglalva, a 43 beszélő 6 különböző spontán beszédben felvett érzelmi kategóriáinak statisztikai vizsgálata alapján elmondható, hogy az alapfrekvencia és az intenzitás frázison belüli időbeli változása, valamint egy frázis egészére átlagolt színképi paraméterek együttesen jellemzik a különböző érzelmeket. Az, hogy meg tudjuk mondani, melyik paraméter mikor és milyen súllyal járul hozzá a komplex érzelmi jellemzés kialakításához, még további kutatást igényel.

A nyelvi tartalomtól független hangi események

A nyelvi tartalomtól elhatárolt spontán beszédben előforduló egyes hangeseményeket, kitöltött szüneteket (Horváth 2009), hümmögéseket (Markó 2005, 2006) korábban már a BEA adatbázis (BEA: BEszélt nyelvi Adatbázis, Gósy 2008) és más korpuszok felhasználásával részletesen vizsgáltak a magyar fonetikai szakirodalomban.

A BME TMIT Beszédakusztikai Laboratóriumában egy-egy kiragadott hangesemény részletes vizsgálata helyett azt kerestük, hogy a magyar spontán beszéd felvételekor milyen hangesemények fordulnak elő: a jelentést kifejező hangjelenségek, vagyis a hanggesztusok, valamint a jelentéssel nem

rendelkező hangjelenségek, kitöltött szünetek, testhangok. A testhangok a beszédhez ugyan nem tartoznak, de a beszéddel együtt jelennek meg. Tehát az öt felsorolt adatbázisban egy külön szinten jelöltük azokat a hangeseményeket, amelyek a nyelvi tartalomtól elhatároltan, attól függetlenül jelentek meg. Bejelölésre kerültek még olyan hangok is, amelyek nem vokális eredetűek, mint például a csók, vagy taps, mivel akusztikailag ezek a hangesemények is rajta vannak a felvételeken. Az öt adatbázisban előforduló hangjelenségeket a 2. táblázat mutatja.

2. táblázat: A nyelvi tartalomtól elhatárolt, attól független hangesemények. (A zárójelben lévő számok az előfordulások számát jelölik.)

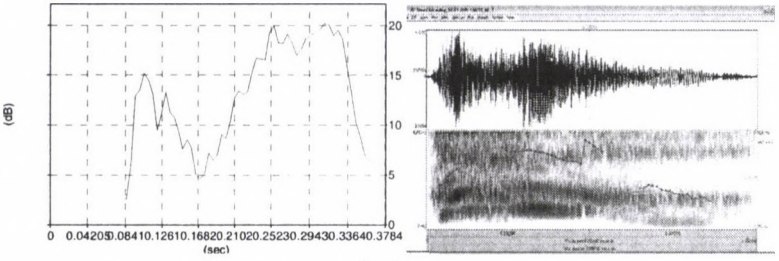
hanggesztusok	nevetés (62), fojtott nevetés (29), sírás (10), fütyülés (6), kitartott üvöltés (24), kitartott *a* (26), *aha* (80), kitartott *á* (1), *áo* (2), *csss* (1), *fu* (1), *ja* (28), *jé* (1), *hohó* (1), *hoppá* (2), *izé* (30), *na!?* (30), *húha* (14), *hát* (52), kérdő hümmögés (14), igenlő hümmögés (144), elgondolkodó hümmögés (20), *na? na!* (30), *psz!* (1), sziszegés (3), *vao* (1);
kitöltött szünetek	különálló *ö*-zés (410), egybeolvadva a megelőző mássalhangzóval – *(hogy)ő* (122), *(hát)ő* (51), *(csak)ö* (8), *(mert)ö* (106), *(ez)ö/(az)ö* (7), *(mikor)ö* (10), egybeolvadva egy mássalhangzóval – *öh* (21), *öm* (44);
nem vokális eredetű hangok	csók hangja (4), pofon (2), tapsolás (2);
testhangok	ásítás (1), böfögés (2), csámcsogás (2), köhögés és/vagy krákogás (102, csuklás (4), lélegzés (7), lihegés (2), nyelvcsettintés (4), nyögés (3), sóhaj (30), szipogás (19), tüsszentés (3)

A kijelölt hangeseményeket kivágtuk és csoportokba gyűjtöttük. Megadtuk a csoportonként jellemző akusztikai jellemzőket. Így hoztunk létre egy ún. Hanggesztustárat, amelybe a hanggesztusokon kívül a 2. táblázat összes hangeseményét feltüntettük. A tárban az 5 adatbázis vizsgálatával kapott hangesemények gyűjteménye található, az akusztikai jellegzetességeikkel együtt, továbbá egy-egy jellemző minta hangképe (spektrogram, alaphang, intenzitás, dinamika, harmonikus-zörej arány dB-ben), amint az a 6. ábrán látható. A tár alapja elkészült és azóta is folyamatosan bővül.

Összefoglalás

A beszéd nem verbális hangjelenségei közül e tanulmányban elsősorban a nyelvi tartalommal együtt megjelenő érzelmi, hangulati kifejeződés jellegzetességeit, feldolgozási nehézségeit tárgyaltuk, valamint bemutattuk a nyelvi

tartalomtól elhatárolt hangesemények gyűjtéséből létrehozott ún. Hanggesztustárat.



6. ábra

A *hű* meglepődésgesztus adatai a Hanggesztustárban.
(Balra: harmonikus-zörej arány dB-ben az idő függvényében.
jobbra: oszcillogram és spektrogram.)
(Női beszélő, átlagos alaphang: 337Hz, átlagos harmonikus-zaj arány: 5,2 dB)

A spontán, kvázispontán beszédadatbázisok feldolgozása során derült ki, hogy még az alapérzelmek (boldogság, szomorúság, düh, meglepetés, undor) jelölése sem egyértelmű feladat az adatbázist feldolgozó szakember számára. Döntését a szövegkörnyezet nagymértékben befolyásolja. Mivel vizsgálataink egy adott érzelem prozódiai jellemzővektorait akartuk meghatározni a nyelvi tartalom nélkül, ezért az elemzéshez a folyamatos szövegből kiragadtuk a különböző érzelmeket tartalmazó frázisokat. Az összegyűjtött mintákból szubjektív lehallgatási kísérlettel választottuk ki a 6 különböző érzelmet (semleges, bánatos, haragos-ideges, meglepett, nevetve beszélő, örömöt) hordozó frázismintákat. Az így nyert 43 beszélő 1000 frázismintáján statisztikai vizsgálatokat végeztünk. A statisztikai vizsgálat alapján elmondható, hogy az alapfrekvencia és az intenzitás frázison belüli időbeli változása, valamint egy frázis egészére átlagolt színképi paraméterek együttesen jellemzik a különböző érzelmeket. Az, hogy meg tudjuk mondani, hogy ezen paraméterek közül melyik milyen súllyal járul hozzá a komplex érzelmi jellemzés kialakításához, még további kutatást igényel.

A spontán beszédben igen nagy gyakorisággal jelennek meg a nyelvi tartalomtól elhatárolt, attól független hangesemények (kitöltött szünetek, hanggesztusok). Ezeknek a vizsgálata a spontán beszéd tudományos leírása szempontjából igen fontos. De fontos a vizsgálatuk az automatikus szövegtartalom felismerése szempontjából is, hiszen ha jelen vannak, akkor ezekre az akusztikailag különböző hangi eseményekre is be kell tanítanunk akusztikus modelleket. Csak így kaphatunk jó eredményt. Ezért hoztuk létre a Hanggesztustárat. Távlati cél, annyi hanggesztus példa összegyűjtése egy-egy fajtából, hogy alkalmas legyen az adott hanggesztus akusztikai modelljének a felépíté-

sére, ami majd a spontán beszéd szövegtartalmának automatikus felismerését fogja segíteni.

Irodalom

Álvarez, Aitor – Cearreta, Idoia – López, Juan Miguel – Arruti, Andoni – Lazkano, Eelena – Sierra, Basilo – Garay, Nestor 2007. A comparison using different speech parameters in the automatic emotion recognition using feature subset selection based on evolutionary algorithms. *TSD LNAI* 4629. 423–430.

Gósy Mária 2008. Magyar spontánbeszéd-adatbázis – BEA. *Beszédkutatás 2008.* 194–207.

Burkhardt, Felix – Paeschke, Astrid – Rolfes, Miriam – Sendlmeier, Walter – Weiss, Benjamin 2005. A Database of German Emotional Speech. In: *Proceedings of Interspeech 2005.* 1517–1520.

Campbell, Nick 2004. Getting to the heart of the matter; speech as the expression of affect, rather than just text or language. *Language Resources and Evaluation Conference* 39/1. 109–118.

Campbell, Nick 2007. Individual traits of speaking style and speech rhythm in a spoken discourse. *Lecture Notes in Computer Science* 5042. 107–120.

Cowie, Rody – Douglas-Cowie, Ellen – Tsapatsoulis, Nicolas – Votsis, Gorge – Kollias, Stefanos – Fellenz, Winfried – Taylor, John 2001. Emotion recognition in human-computer interaction. *IEEE Signal Process* 18/1. 32–80.

Douglas-Cowie, Ellen – Campbell, Nick – Cowie, Rody – Roach, Peter 2003. Emotional speech: Towards a new generation of databases. *Speech Communication* 40. 33–60.

Esposito, Anna 2009. The perceptual and cognitive role of visual and auditory channels in conveying emotional information. *Cognitive Computation* 1/3. 268–278.

Horváth Viktória 2009. *Funkció és kivitelezés a megakadásjelenségekben.* PhD-értekezés. ELTE, Budapest.

Hozian, Vladimir – Kacic, Zdravko 2003. Context-independent multilingual emotion recognition from speech signals. *International Journal of Speech Technology* 6. 311–320.

Kohavi, Ron 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2/12. 1137–1143.

Kostoulas, Theodros – Ganchev, Todor – Fakotakis, Nikos 2007. Study on speaker-independent emotion recognition from speech on real-world data. *Lecture Notes in Computer Science* 5042. 235–242.

Devillers, Laurence – Vidrascu, Laurence – Lamel, Lori 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18. 407–422.

Mády Katalin 2005. MAPTASK. http://www.phonetik.uni-muenchen.de/~mady/corpora/maptask/

Markó Alexandra 2005. „Szavak nélkül". Nonverbális vokális közlések fonetikai elemzése. *Magyar Nyelvőr* 129. 88–104.

Markó Alexandra – Bóna Judit 2006. A spontán beszéd lejegyzésének néhány módszertani kérdése. *Beszédkutatás 2006.* 124–133.

Markó Alexandra 2006. Nonverbális vokális jelek a társalgásban. *Beszédkutatás 2006.* 57–68.

MPEG-4 1999. ISO/IEC 14496 standard. http://www.iec.ch

Navas, Eva – Hernáez, Imma – Luengo, Iker 2006. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Transactions on Audio, Speech, and Language Processing* 14/4. 1117–1127.

Nogueiras, Albino – Moreno, Asunción – Bonafonte, Antonio – Marino, José B. 2001. Speech emotion recognition using Hidden Markov Models. In *Eurospeech 2001.* 2679–2682.

Seppänen, Tpio – Väyrynen, Eero – Tovanen, Juhani 2003. Prosody-based classification of emotions in spoken Finnish. In *Eurospeech 2003.* 717–720.

Tóth Szabolcs Levente – Sztahó Dávid – Vicsi Klára 2007. Speech emotion perception by human and machine. In *Proceeding of COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007: Revised papers in verbal and nonverbal features of human-human and human-machine interaction.* 213–224.

Vicsi Klára – Sztahó Dávid 2009. Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): *VI. Magyar Számítógépes Nyelvészeti Konferencia.* JATEPress, Szeged, 217–225.

A kutatás a Jedlik OM-00102/2007 számú „TELEAUTO" projekt
és a TÁMOP-4.2.2-08/1/KMR-2008-0007 projekt keretein belül készült.