Spontán beszédben rejlő nem verbális hangjelenségek – érzelmek, hanggesztusok – vizsgálata

Vicsi Klára, Sztahó Dávid, Kiss Gábor, Czira Anita

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium,
1111 Budapest, Sztoczek u. 2.
{vicsi, sztaho}@tmit.bme.hu

Kivonat: Ebben a cikkben azokat a vizsgálatokat tárgyaljuk, amelyek a spontán beszéd nem verbális hangjelenségeinek a kutatására vonatkoznak. Elsősorban a nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalom jellegzetességeit, feldolgozási nehézségeit tárgyaljuk, 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 hanggesztustárat.

1 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 általános érzelmi, egészségi állapotát, hangulatát érzékeljük [1]. Az utóbbi évtizedben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna jelentősége 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. Ezeket a beszélő különböző beszédformákkal (változatok) tudja érzékeltetni. A hangszínezet, az intonáció, a ritmusváltozások mind széles körben használatosak arra, hogy a beszélő érzelmi, hangulati vagy egészségi állapotát is egyidejűleg kifejezzék.

Korábban a beszédtartalom vizsgálatakor rendszerint olvasott, vagy szépen kiejtett beszéd volt a vizsgálat alapja, viszont a beszédtechnológiai alkalmazásokban a valóságos spontán beszéd feldolgozása szükséges!

Spontán társalgásban számos nem nyelvi elem fordul elő, amelyek hozzájárulnak ahhoz, hogy a beszélgetőpartnerek jobban megértsék egymást. A beszédkommunikációban a lelki állapot, az érzelem, az egyetértés vagy egyet nem értés közvetítése azt a célt szolgálja, hogy a beszélgetőpartnert 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:
   - 2.1. jelentést kifejező hangjelenségek – ezek a hanggesztusok. Ilyenek például a sírás, a nevetés, a különböző érzelmet kifejező felkiáltások.
   - 2.2. jelentéssel nem rendelkező hangjelenségek:
     - 2.2.1. Kitöltött szünetek
     - 2.2.2. 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 beszédelőállítás és a gépi spontánbeszéd-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 hanggesztustárat. Mindezen vizsgálatokhoz igen nagy mennyiségű spontán hanganyag gyűjtésére és feldolgozására volt szükség.

2 Módszer, adatbázisok

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

Magyar Telefonos Ügyfélszolgálati Beszéd Adatbázis (MTÜBA)

Ügyfél és diszpécser beszélgetése került rögzítésre, az adatbázis 1100 ilyen felvételből áll.

Maptask adatbázis

Az adatbázis 1113 „rövid” .wav fájlt tartalmaz, 10 különböző személlyel rögzített spontán beszéd útkeresés témában.

Balázs-show felvételek

135 percnyi műsoridő. 44 női és 99 férfibeszélő hanganyaga került feldolgozásra.

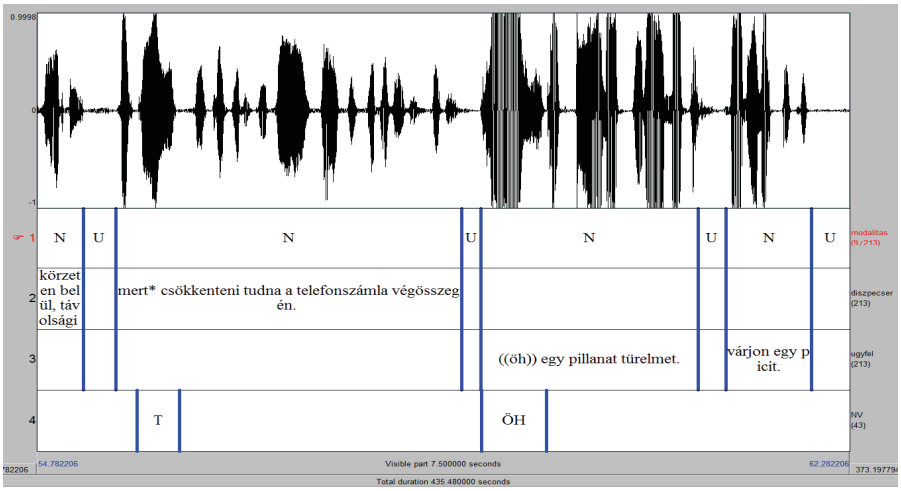
Joshi Bharat-felvételek

Szintén egy beszélgetős műsor, 61 percnyi műsoridővel.

Mozi

Végül pedig egy spanyol „Torrente” című 3 részes akcióvígjátékból gesztusok és egyéb nem verbális hangesemények kerültek kigyűjtésre.

A hanganyagok feldolgozása frázisegységenként [2] több szinten történt (lásd 2. ábra). Első szinten frázisonként bejelölésre került az adott frázisban kifejezésre jutó érzelem. A következő szinten/eken a nyelvi tartalom került bejegyzésre beszélőnként külön-külön. Az utolsó szinten a szövegben már csillaggal jelzett helyeknél lévő hangi események időtartama és típusa lett bejelölve.



2.ábra. Az adatbázisok többszintű feldolgozása. 1. Frázisonkénti érzelembejelölés (N: semleges, U:szünet); 2.3. Nyelvi tartalom bejelölése; 4. Nem verbális hangi események (T:kitöltött szünet ’t’ hang után, ÖH kitöltött szünet ’öh’-t ejtve)

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

3 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 feldolgozásra, 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 [3, 4, 5, 6]. 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 [7]. A beszédtechnológiai alkalmazásokban a valóságos spontán beszéd feldolgozása szükséges. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával [8] és információtartalmainak felismerésével [9] foglalkozik.

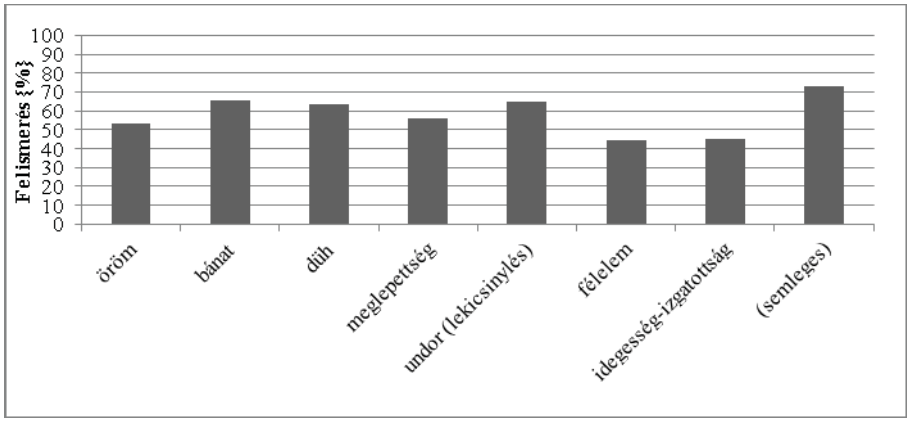
Problémát jelent továbbá az érzelmi kategóriák változatos megjelenése a beszédben. Az emóció jellemzésére a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban hagyományos emóciókategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban [10] e kategóriákat az arcmimika jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták. A beszédtechnológiai szakemberek 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. Kutatási céllal a spontán beszédben leggyakrabban előforduló érzelmi kategóriákat gyűjtöttek ki a PHYSTA 2001 adatbázisból [11]. 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 sz. 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	10	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 [12]. Az említett munkában ugyanazon szemantikai tartalmú mondatok különböző érzelmekkel kerültek bemondásra két csoportban, színészekkel és átlagemberekkel (3 mondat, mondatonként 8 érzelem, 15 személlyel).



3. ábra. Az átlagemberek bemondásainak érzelmek szerinti felismerése percepciós teszttel mérve.

Ezeket a mondatokat meghallgattatták érzelem szerinti megítélésre 20 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

tényezők is szerepet játszhatnak az érzelem kifejezésében). Agyunk az összes érzékszervi csatornán keresztül kapott információ együtteséről dönt [6]. 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 érzelemfelismerési 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 [13].

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 [5].

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 [1, 3, 6, 12]).

3.1 Beszédérzelmek jellemző vektorai a szakirodalomban

A gépi érzelem-felismerés során a meglévő hanganyagból jellemzővektorokat 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, torokbetegségek). 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 érzelemkifejezésben, és melyek nem. A mai napig az ide vonatkozó szakirodalom egyik fő kérdése, hogy az automatikus érzelemfelismeréshez milyen jellemzőket kell kigyűjteni, amelyek alapján majd a felismerés működni fog.

Az irodalomban összefoglalóan az alábbi érzelmekre jellemző fizikai paraméterekkel találkozhatunk [14, 15]:

Alapszintű adatok a jellemzővektorokban

Az úgynevezett alapszintű jellemzők közé tartoznak a keretenkénti alaphangfrekvenciaértékek, a hangintenzitás-értékek, valamint a beszédhangok időtartama.

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 alacsony szintű jellemző a szótagok, beszédhangok időtartama. Ezek meghatározzák a beszéd tempóját, ritmusváltásait.

Származtatott adatok a jellemzővektorokban

A származtatott jellemzőket az alapszintűekből 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 frekvenciatartomány együtthatóit (MFCC-együtthatók) is besorolják az érzelmek jellemző paraméterei közé [12].

A származtatott jellemzők, amelyet az irodalomban mondategységekre számítottak ki, folyamatos spontán beszédben nem igazán 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 [2] 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.

3.2 Beszédérzelmek jellemző vektorai frázisokban

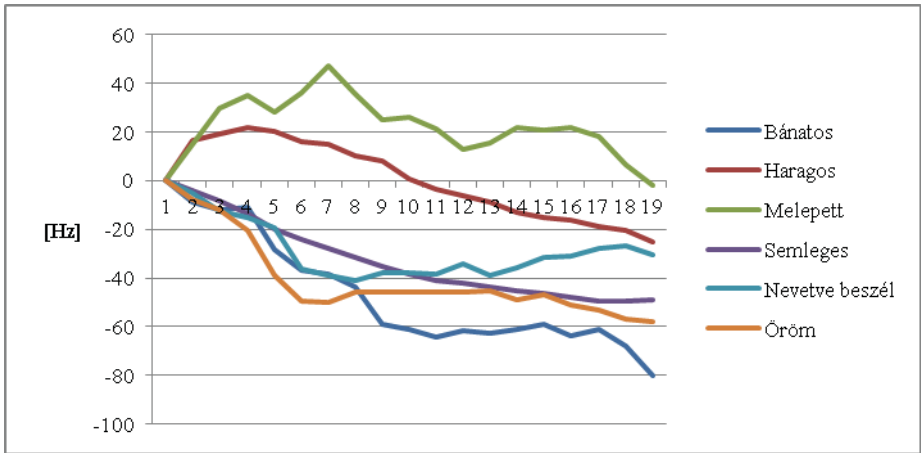
Jellemző vektorok vizsgálatát az összegyűjtött 5 spontán adatbázis felhasználásával végeztük el. Ezeknek az adatbázisoknak a feldolgozása során már kiderült, hogy csupán a tiszta érzelmek jelölése sem egyértelmű feladat, és rendszerint az annotátort a döntésben a szövegkörnyezet nagymértékben befolyásolja. 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 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 osz-

tottuk be 6 különböző érzelmi kategóriába, amelyek a semleges, bánatos, haragos-ideges, meglepett, nevetve beszélő, örömöt kifejező.

Az alapszintű 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 frázisok lineárisan vetemítésre kerültek úgy, hogy mindegyik minta „n"hosszúságú lett, majd a mért adatokat normáltuk a frázisban mért első átlagadat értékére úgy, hogy a mintavételezési pontoknál mért adatokból az első minta értéke levonásra került. Végül az érzelem szerinti csoportok frázisonkénti értékei átlagolásra kerültek, vagyis minden érzelemre elkészült az adott "**érzelemre jellemző átlagos hangminta-dinamika**" mind alapfrekvenciában, mind összintenzitásban.

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



4. ábra. A különböző érzelmek átlagos alapfrekvencia-dinamikája. Vízszintes tengelyen a mintavételezési pontok láthatók.

Bánatos:

Alapfrekvencia folyamatos és nagymértékű csökkenését figyelhetjük meg. Majd körülbelül a frázis felénél, 60Hz-es csökkenés után egy stagnálást, majd a végén újabb csökkenést.

Haragos:

Az elején nő az alapfrekvencia, majd folyamatosan csökken.

Meglepett:

Az elején nagymértékű alapfrekvencia-növekedés látható, majd valamelyes csökkenés. Ennél az érzelemkategóriánál figyelhető meg leginkább az alapfrekvencia növekedése.

Semleges:

Az alapfrekvencia folyamatos szabályos csökkenése figyelhető meg, bár annak mértéke nem igazán jelentős.

Nevetve beszél:

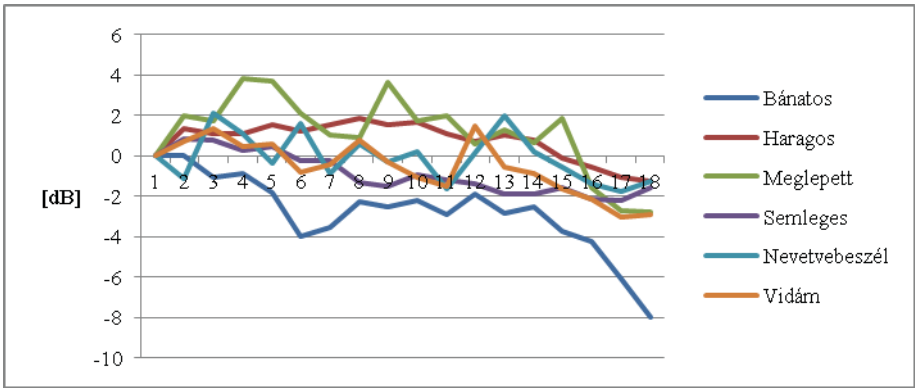
Az alapfrekvencia csökkenése, majd körülbelül a frázis felétől alacsony növekedése jellemzi.

Öröm:

Elején az alapfrekvencia lényeges csökkenése figyelhető meg a frázis felétől körülbelül 50Hz, 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 intenzitásának dinamikái nem különülnek el olyan szépen, mint az alapfrekvencia változásának esetében, amint ez az 5. ábra alapján látható. Itt az értékek nem az első mintavételezési helytől kerültek ábrázolásra, hanem a másodiktól, emiatt az utolsó mintavételezési hely sorszáma a 18-as.



5. ábra. A különböző érzelmek átlagos intenzitásdinamikája. Vízszintes tengelyen a mintavételezési pontok láthatók.

A szórásértékek 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 intenzitásértékek kevésbé tükrözik a különböző érzelmeket, bár azért jellemző dinamikajegyek 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.

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. Ezen jellemző vektorok alapján végzett automatikus érzelem-felismerési kísérletekről jelen kötetben egy másik cikk fog beszámolni.

4 A nyelvi tartalomtól független hangi események

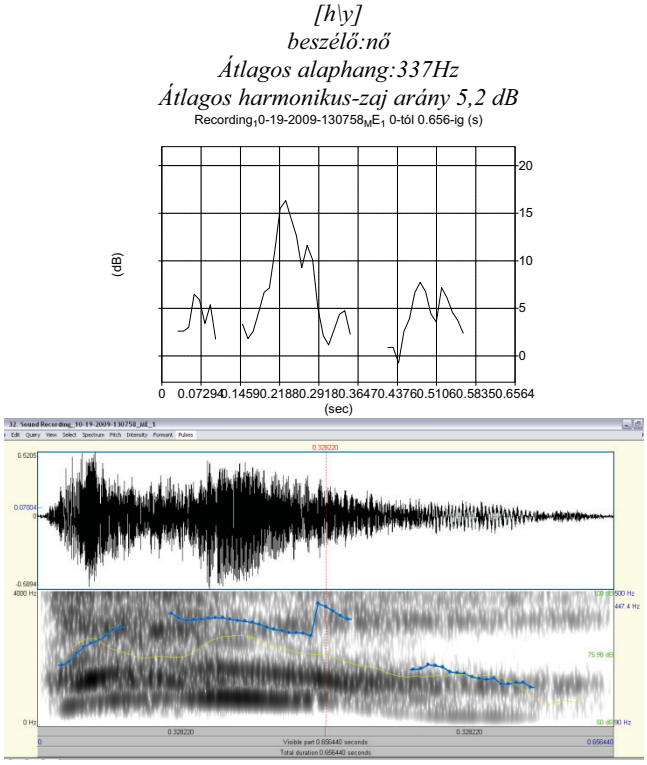
Az 5 felsorolt adatbázisban jelölésre kerültek azok a hangesemények is, amelyek a nyelvi tartalomtól elhatároltan, attól függetlenül jelentek meg. Ezek a jelentést kifejező hangjelenségek, vagyis a hanggesztusok, valamint a jelentéssel nem rendelkező hangjelenségek, kitöltött szünetek, testhangok. Bejelölésre kerültek még olyan a beszélgető partnerektől származó hangok, amelyek nem vokális eredetűek, mint például a csók vagy taps. Az 5 adatbázisban előforduló hangjelenségeket a 2. táblázatban soroljuk fel.

2. táblázat: A nyelvi tartalomtól elhatárolt, attól független hangi események.

Hanggesztusok	**Kitöltött szünetek**
L – nevetés(15)	A: – [A:] (25)
S –sírás(0)	d'2 – [(ho)d'2], [(pEdi)g2], [(ho)d'2], [(E)d'2],(72)
[jO], [jOj] (15)	h2 – [hA:t 2:] (47)
[nO]! (16)	ER – er… (18)
[(h\)u:h\O] (4)	k2**:** – [(ki:vA:no)k2:],[(tSO)k2:], [(mond'u)k2:] (7)
[h\A:t] (31)	2: – [2:] (66)
[h\y]?, [h\yF]? (8)	2:x: – [2:x:] (15)
yep (67)	2F: – [2F:], [2h\F:], [2yFh\:] (30)
[F:] (9)	2: – extrém hosszú[2:] (67)
h\F – hum! (csukott szájjal) (76)	r2: – [(Omiko)r2:] (7)
[ps]! (1)	t2: – [(mEr)t2:], [(tEh\A:)t2:] (72)
egyéb – (hahaha, há, fú, hóhó, phöhö, éé, hoppá, ú, húha, ajaja, háhá, nya, aő, au) (36)	

Nem vokális eredetű hang	**Testhangok**
KISS – csók hangja (3)	B – böfögés (2)
SLAP – tapsolás (2)	CO – köhögés (32)
	MO – csámcsogás (2)
	HIC –csuklás (4)
	BR – lélegzés (7)
	S – szipogás (11)
	SN – trüsszentés (1)

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 a kigyűjtött 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.



6. ábra. A [h\y] meglepődésgesztus adatai a hanggesztustárban. Balra: harmonikus-zörej arány dB-ben az idő függvényében, jobbra: amplitúdó-idő függvény, alatta spektrogram.

Távlati cél, annyi hanggesztus példa összegyűjtése egy-egy fajtából, hogy alkalmas legyen az adott hanggesztus akusztikai modelljének a felépítésére, ami majd az automatikus spontánbeszéd-felismerést fogja segíteni.

Köszönetnyilvánítás

Ez 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.

Bibliográfia

1. Burkhardt, F., Paeschke A. et al.: A Database of German Emotional Speech. In: Proc. of Interspeech2005 (2005) 1517–1520
2. Vicsi, K., Sztahó, D.: Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In: Tanács A., Szauter D., Vincze V. (szerk.):VI. Magyar Számítógépes Nyelvészeti Konferencia. JATEPress, Szeged (2009) 217–225
3. Campbell, N.: Getting to the Heart of the Matter. Keynote Speech. In Proc. Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal (2004)
4. Campbell, N.: Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. In COST Action 2102 International Conference on Verbal and Nonverbal Features. Patras, Greece (2007) 107–120
5. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional Speech: Towards a New Generation of Databases. Speech Communication Vol. 40 (2003) 33–60
6. Hozian, V., Kacic, Z.: Context-Independent Multilingual Emotion Recognition from Speech Signals. International Journal of Speech Technology Vol. 6 (2003) 311–320
7. Kostoulas, T., Ganchev, T., Fakotakis, N.. Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data. In: COST Action 2102 International Conference on Verbal and Nonverbal Features. Patras, Greece (2007) 235–242
8. Navas, E., Hernáez, I., Luengo, I.: 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 Vol. 14, No. 4 (2006)
9. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence Vol. 2 No. 12 (1995) 1137–1143
10. MPEG-4 (1999): ISO/IEC 14496 standard. http://www.iec.ch

Nogueiras, A., Moreno, A., Bonafonte, A., Marino, J. B.: Speech Emotion Recognition Using Hidden Markov Models. In: Eurospeech (2001)

11. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G.,Kollias, S., Fellenz, W.,Taylor, J. Emotion Recognition in Human–Computer Interaction. IEEE Signal Process. Mag. Vol. 18 No.1 (2001) 32–80
12. Tóth, Sz. L., Sztahó, D., Vicsi, K.: Speech Emotion Perception by Human and Machine. In: Proceedings 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 2008. ISBN: 978-3-540-70871-1. Springer LNCS (2008) 213–224
13. Esposito, A.: The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. Cogn. Comput DOI 10.1007/s12559-009-9017-8. Springer Science+Business Media, LLC (2009)
14. Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., Garay, N.: A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms. In: TSD LNAI 4629 (2007) 423–430
15. Seppänen, T., Väyrynen, E., Tovanen J.: Prosody-based classification of emotions in spoken Finnish. In: Eurospeech (2003)