A beszéd érzelemi töltetének számítógépes felismerése

Tüske Zoltán[1,3], Simon Márta[2], Mihajlik Péter[1], Gordos Géza[1]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
{tuske, mihajlik, gordos}@tmit.bme.hu

[2]Semmelweis Egyetem
Pszichiátriai és Pszichoterápiás Klinika
{simonmarta}@t-online.hu

[3]AITIA International

Kivonat: Új megközelítést mutatunk be a beszéd érzelmi tartalmának gépi felismerésére. Megmutatjuk, hogy statisztikai módszerekkel, csak a beszéd akusztikus jellemzői alapján, a szöveges tartalom figyelembe vétele nélkül megfelelő érzelemfelismerési eredményeket lehet elérni. Lineáris diszkrimi-náns alapján válogatott beszédjellemzők mennyiségét – azaz a jellemzővektor dimenzióját – adatvezérelt módszerekkel (PCA és LDA) radikálisan csökkent-jük, majd GMM osztályozókat tanítunk be. Sokbeszélős, hat érzelmi állapotra jellemző, magyar adatbázison átlagosan 42,9%-os felismerési pontosságot ér-tünk el. Felismerőnk 60,2%-kal ismerte fel az érzelmeket beszélőfüggő eset-ben. A megközelítés nyelvek közötti hordozhatóságát mutatja, hogy német adatbázison színészek által produkált felvételeken, kötött szöveges tartalom mellett, hét érzelemi osztállyal 71,8%-os beszélőfüggetlen felismerési ered-ményt értünk el, ami nemzetközi élvonalbelinek mondható.

1 Bevezetés

A beszédfeldolgozás területén az érzelemfelismerés mindinkább a figyelem középpontjába kerül. Az automatikus beszédfelismerővel ellátott rendszerekkel kapcsolatban felmerül az az igény, hogy a beszéd szöveges tartalmán kívül egyéb, non-verbális információt is – például a beszélő érzelmi állapotát – képes legyen figyelembe venni és felhasználni, ezáltal téve természetesebbé a felhasználó és a gép közötti kommunikációt.

Az érzelemfelismerési kutatások különböző forrásokból származó jeleken vizsgálódnak, úgymint fiziológiai, mimikai és beszédjelek. Ez a tanulmány a továbbiakban csak a beszédből géppel kinyerhető érzelmi információkkal foglalkozik.

Az ember képes még a telefonon keresztül érkező sávkorlátozott (400-3700 Hz) akusztikus jelből is a vonal túloldalán levő személy érzelmi állapotának meghatározására. Természetesen a vizuális információ, a gesztikuláció, az arcizmok igen kifinomult játékának hiánya gyakran vezet téves emocionális értékeléshez.

Habár a vokális csatorna közvetítette érzelmeket egyre többen vizsgálják, a számtalan kutatási eredmény ellenére nincs egyetértés abban, hogy az érzelmeket mely akusztikus jellemzők alapján lehet azonosítani, illetve egymástól elkülöníteni [8]. Az mindenesetre igazolt, hogy passzív érzelmek (pl. bánat) esetén az alapfrekvencia (F0) átlaga, tartománya és szórása csökken, míg aktív érzelmek esetén (pl. harag, öröm) növekszik.

Mind az emóciók kifejezése, mind azok észlelése jelentős kulturális, nyelvi, nemi és nem utolsó sorban egyéni különbségeket mutatnak, ebből következően minőségi és mennyiségi megjelenésük is jelentős eltéréseket tükröznek [1].

Az érzelem kifejeződése a verbális tartalomban is jelentkezhet. Általában más szavakat használ egy mérges, mint egy nyugodt ember. Schuller és társai [14] által készített érzelemfelismerő a kombinált vokális és verbális információval pontosabb felismerési eredményt ért el. Természetesen léteznek olyan szituációk, ahol érzelemtől függetlenül azonos mondatok hangozhatnak el, ilyenkor csak a vokális üzenet alapján történhet az érzelem felismerése. Az általunk használt felismerési módszer nem használja föl a beszéd szöveges tartalmát. Ezt azzal indokolhatjuk, hogy ugyan valamivel gyengébb hatásfokkal, de képes az ember egy számára teljesen idegen nyelven beszélő ember érzelmi állapotát is megállapítani [12].

Fontos megemlíteni, hogy azokban a kísérletekben, ahol az alanynak előre megadott öt-hat érzelem alapján kellett számára ismeretlen beszélővel készült felvételeket osztályoznia, az emberi felismerési képesség körülbelül a 60%-ot érte el [8, 10]. Hasonló tesztekben a bemondók a saját érzelmeiket kb. 80%-ban ismerték fel helyesen [10].

Mivel a szakirodalomban nincs egységes álláspont, hogy konkrét emóciókat milyen információk alapján lehet hatékonyan, szabályok alapján megkülönböztetni, ezért mi is statisztikai módon közelítettük meg az érzelemfelismerést.

A géppel történő felismerés pontosságát is, mint akármelyik statisztikus mintaillesztési feladatban, döntően befolyásolják a választott jellemzők, illetve az ezekből összeállított tulajdonságvektor. Tudomásunk szerint az automatikus érzelemfelismerés területén magyar nyelvre vonatkozóan még nem publikáltak hatékonyan használható paramétereket, a külföldi publikációk alapján azonban igyekeztünk minél több akusztikus tulajdonságot összegyűjteni. A szakirodalom által javasolt általános jellemzőket (pl. alapfrekvenciából és intenzitásból származtatott statisztikák – [4, 9, 15]), nem találtuk elég hatékonynak, ezért szükségesnek éreztük, hogy a rengeteg fellelt akusztikus paraméter közül – az általunk kifejlesztett módon – válogassunk, és csak az optimálisnak talált jellemzőkből képzett tulajdonságvektorral dolgozzunk. Ez utóbbi a mintafelismerés szempontjából is kívánatos, hiszen így elkerüljük a túl komplex modellezést, és az ebből eredő problémákat.

A hasznosnak vélt és publikált jellemzők nagy mennyisége, főként arra vezethető vissza, hogy az eredmények nagymértékben függnek a felhasznált adatbázisoktól. Tovább bonyolítja a helyzetet, hogy a statisztikai alapon működő beszélőfüggő és beszélőfüggetlen érzelemfelismerés más-más paramétereket tart hasznosabbnak. Előbbi esetben sokkal jobb eredményt értek el [9, 14], hiszen a tanuló rendszernek nem kell az egyéni különbségekből adódó változatosságot elsajátítania. Másik nagyon fontos tényező az érzelmes felvételek forrása, spontán avagy mesterségesen, színészek által keltett érzelmekről van-e szó. Utóbbi esetben biztosabb a felismerés. A

pusztán a beszéd prozódiájából történő felismerés esetén hasznos, ha érzelmenként azonos szöveges tartalommal rendelkező felvételeket használhatnánk, így a megfelelőnek ítélt akusztikus jellemzők biztosan az érzelmek közötti prozódiai eltéréseket ragadnák meg. Ebben az esetben le kell mondanunk arról az igényről, hogy a tanításra használható adatbázis felvételei spontán érzelemkifejezést tartalmazzanak.

Érdemes kiemelni, hogy a felismerendő érzelmek számának növelésével a publikált eredmények drámaian romlanak. Például, egy telefonos beszédinterfészen keresztül irányított ügyfélszolgálatnak érdeke, hogy az ideges ügyfeleket valódi operátorokhoz kapcsolja. Ebben az esetben két érzelmi állapot elegendő a felismerési feladat szempontjából, a publikált eredmények 90% fölöttiek [7, 14]. Egy diagnosztikai rendszer esetében komplex érzelmeket kell kezelni, ezért tíznél is több állapot lenne szükséges, spontán, 5 osztályos klasszifikáció esetén az 50%-os hatékonyság is már jónak számít [7]. A csak prozódiai jellemzők alapján történő érzelemfelismerésről szóló kutatási eredmények túlnyomó többségében az alapérzelmek szintjén megállnak, nem elég hatékonyak, így az összetettebb érzelmek felismerése még várat magára.

Kutatásunk hat érzelmi állapot - *harag, szomorúság, undor, neutrális, öröm, meglepődés* - pusztán vokális információ alapján történő automatikus felismerését magyar felvételeken tűzte ki célul. Ezt kétféle, beszélőfüggő és beszélő független saját adatbázison végeztük el. Célunk volt érzelmekhez köthető jellemzők keresése, és az ezekkel elérhető hatásfok összehasonlítása a nemzetközi tapasztalatokkal. Bemutatjuk az általunk hasznosnak talált jellemzőket és a válogatásukra használt, nyilvános német adatbázison is tesztelt módszerünket. Az egyes adatbázisokon nyert akusztikus paraméterek közlése után ismertetjük a felismerőrendszerünkkel elért eredményeinket.

2 Adatbázisok

A kísérletekhez kétféle adatbázis készült, mindkettő 44.1 kHz-es mintavételezéssel és 16 bites kvantálással. Az első korpusz (HU_SI) 34 beszélőtől tartalmaz érzelmenként 2-3 példamondatot. Összesen 243 spontán bemondásból áll.

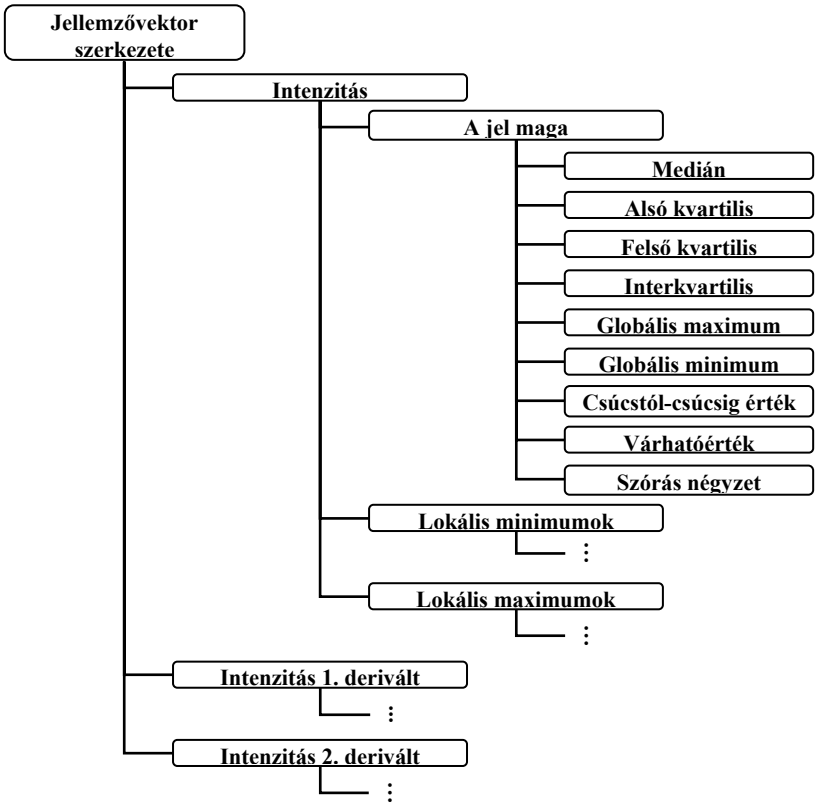
A második adatbázis csupán két beszélőtől felvett érzelmes mondatokból áll. Tartalmaz olyan nem-spontán bemondásokat, amelyek mindkét beszélőtől minden érzelemmel elhangoznak; érzelmenként különböző, de minkét beszélő esetében azonos tartalmú mondatokat; valamint spontán, érzelmenként és beszélőnként is különböző, egyedi mintákat, összesen 198 felvételt (HU_SD).

A szöveges tartalom nélküli megközelítés előnye, hogy lehetőség volt – apróbb módosítások után – német nyelvű, a Berlini Műszaki Egyetemen készült, nyilvános, érzelmes beszédadatbázison [3] is tanítani és tesztelni. A korpusz színészek által mesterségesen keltett, hétféle érzelemmel készült: *semleges, harag, félelem, öröm, bánat, undor, unalom*. Összesen 537 felvételből áll, és 10 beszélővel készült. A szöveges tartalom beszélőnként, érzelmenként azonos volt. (DE_SI)

3 Akusztikus jellemzők

A fellelhető szakirodalomban nem találni egyértelmű javaslatot a sikeres érzelemfelismeréshez szükséges jellemzőkre vonatkozóan. A legtöbb kutatási eredmény [2, 6, 14] a hosszúidejű jelszakaszokból (mondat, több szó; kb. néhány másodperc) nyert paraméterekből indul ki, így rendel minden egyes bemondáshoz egy jellemzővektort.
Általánosan alkalmazottak az alapfrekvencia (F0) és az energia (E) időjeleiből származtatott statisztikák (szórás, átlag, minimum, maximum stb.). A beszédjel energiáját általában további alsó és felső energiára osztják [16], a határ 4-600 Hz körüli. Fontos a beszéd sebessége és annak ingadozása is. A beszédfelismerési tapasztalatokból ismert, hogy a beszéd rövid idejű szakaszait (kb. 32 ezredmásodperc) igen tömören jellemzik a kepsztrális együtthatók (MFCC = Mel Frequency Cepstral Coefficient). Általános módszer, hogy ezekből az együtthatókból származtatott hosszúidejű statisztikákat is bevonják az érzelemfelismerésbe [7].

A fentiek alapján tehát adott bemondásra mértük a következő alábbi időjeleket: intenzitás, alsó energia, felső energia, alapfrekvencia, MFCC vektor hossza, 10 darab MFCC. Beszédfelismerőt alkalmazva lehetőség adódott az elhangzott szavak rejtett Markov-modellből történő kijelölésére, így a beszélő által egységnyi idő alatt kiejtett hangok és szavak mennyiségének (artikulációs sebesség és „szóráta") mérésére is. Számoltuk az első és a második deriváltakat (sebesség, gyorsulás) is. Ezekből a jelekből további „hosszú idejű" jeleket származtattunk. Ezzel az időjelek szélsőértékeinek változásit igyekeztünk figyelembe venni: lokális maximumok, lokális minimumok. Majd minden hosszúidejű jelen számoltuk a következő statisztikákat: medián, alsó kvartilis, (a legkisebb és a medián között középen elhelyezkedõ adat számértéke a rendezett mintában), felső kvartilis (hasonlóan a medián és a legnagyobb érték között van középen), interkvartilis (felső és alsó kvartilis különbsége), maximum, minimum, maximum és minimum különbsége (csúcstól-csúcsig érték), tapasztalati várható érték, tapasztalati szórás. Összesen 1377 (=17*3*3*9) darab jellemzőt vizsgáltunk (*1. ábra*).



1. Ábra: Az előállított jellemzővektor szerkezetének illusztrálása a beszédintenzitás jeléből származtatott statisztikákkal

4 A jellemzők válogatása

A beszédjelből nyert paraméterek vizsgálata egyenként történt. A Fisher-féle lineáris diszkrimináns analízisből ismert osztályok közötti és osztályon belüli variancia számítás alapján képzett hányadosok mutatják az egyes jellemzők szeparáló képességét. Esetünkben ennek alkalmazása úgy történt, hogy vettünk egy érzelmes osztályt (pl. harag), míg a többi érzelemhez tartozó adatokat összevontuk egy közös osztályba (pl. nem-harag). Ezután minden egyes jellemzőre kiszámoltuk az erre a két osztályra vonatkozó szeparáló képességet. Ezt minden érzelmi osztályra elvégeztük, majd a legdiszkriminálóbb jellemzőket gyűjtöttük össze az egyes szeparációvizsgálatból. Összesen 40 darab különböző jellemzőt.

Azért, hogy valóban a legjobb jellemzőket találjuk meg, a fent leírt módszert a keresztkiértékelésből ismert leave-one-out módszerrel használtuk. Beszélőfüggetlen esetben minden egyes tesztből kihagytunk egy beszélőt, beszélőfüggő esetben érzel-

menként az adatok 1/10-ét. Így például a HU_SI adatbázison 34 darab negyven elemű vektort kaptunk Végül csak az összes tesztben szereplő jellemzőket tartottuk meg. Az *1. táblázat*ban az egyes adatbázisokon ilyen módon nyert jellemzők számát láthatjuk.

1. Táblázat: Az egyes adatbázisokból kinyert leghasznosabb jellemzők száma

ADATBÁZIS	JELLEMZŐK SZÁMA
HU_SI	24
HU_SD	16
DE_SI	18

A *2.* és *3. táblázat*ban az egyes magyar adatbázisokon nyert néhány hasznos jellemző látható. Kaptunk olyan paramétereket, melyek a többszörös teszt alapján egyértelműen egy érzelem többitől való megkülönbözetésére szolgál, valamint olyan általános akusztikus tulajdonságokat, amik minden teszben jó szereparálási képességet mutat, de szorosan egyik osztályhoz sem köthető. Ami meglepő, hogy beszélőfüggetlen esetben csak az MFCC együtthatókból származtatott statisztikákat találunk, beszélőfüggő esetben a paraméterek 1/5-e az intenzitásból származik.

2. Táblázat: Néhány a beszélőfüggetlen (HU_SI) adatbázison nyert érzelemhez köthető és általánosan jól teljesítő jelparaméter

Harag	3. MFCC szórása MFCC vektor hossz szórása
Undor	10. MFCC 2. deriváltjának csúcstól-csúcsig értéke 10. MFCC 2. deriváltjának szórása
Öröm	10. MFCC maximumainak mediánja 10. MFCC alsó kvartilise
Neutrális	1. MFCC felső kvartilise 1. MFCC maximumainak mediánja
Szomorúság	MFCC vektor hosszának maximumainak szórása 10. MFCC szórása
Meglepődés	1. MFCC maximumainak csúcstól-csúcsig értéke 1. MFCC maximumainak minimuma
Általános	10. MFCC együttható maximumainak szórása 9. MFCC 2. deriváltjának felsőkvartilise 1. MFCC együttható maximumainak mediánja 1. MFCC együttható felső kvartilise

3. Táblázat: Néhány a beszélőfüggő (HU_SD) adatbázison nyert érzelemhez köthető és általánosan jól teljesítő jelparaméter

Harag	intenzitás medián
Undor	MFCC vektor hosszának alsó kvartilise intenzitás maximumainak mediánja
Öröm	intenzitás maximumainak alsó kvartilise
Neutrális	felső energia mediánja
Szomorúság	*nem találtunk egyértelmű jellemzőt*
Meglepődés	MFCC vektor hosszának alsó kvartilise
Általános	MFCC vektor hosszának csúcstól-csúcsig értéke MFCC vektor hosszának minimumainak alsókvartilise Felső energia 1. deriváltjának alsó kvartilise

5 Tanítás és felismerés

Adott felvételből képzett vektort a Bayes-döntéssel soroltunk egyik vagy másik érzelmes osztályba, azaz a legnagyobb valószínűségű érzelemre döntöttünk, az egyes érzelmek valószínűségét azonosnak tételeztük fel.

$$\hat{C} = \arg\max_i \{P(C_i|z)\} = \arg\max_i \{P(z|C_i)P(C_i)\}$$

Ahol z jelenti a döntés előtt álló, beérkezett vektort, C_i pedig az egyes érzelmi osztályokat. A döntéshez szükséges feltételes eloszlásfüggvényeket Gauss függvények keverékével (Gaussian Mixture Modell = GMM) becsültük. A válogatott mennyiségű jellemzőkön az alábbi transzformációk elvégzése után kapott vektorokkal tanítottuk az egyes érzelmek modelljeit. A tanítás és felismerés során alkalmazott lépéseket a *2. ábra* foglalja össze.

Standardizálás: A tanításhoz használt adatok alapján egységnyi szórásúvá és nulla várhatóértékűvé tettük az egyes dimenziókat, standardizáltuk az adatokat.

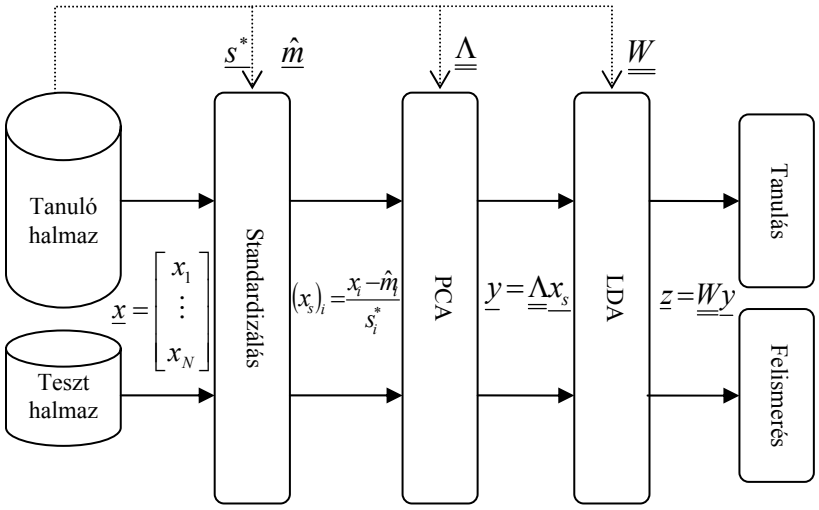
Főkomponens analízis (PCA): A kiválasztott jellemzők között előfordulhatnak olyanok, melyek között szoros összefüggés, korreláció lehet. A standardizált adatok korrelációs mátrixa az alábbi módon becsülhető.

$$\underline{\underline{R}} = \frac{1}{n-1}\sum_{\underline{x}} \underline{x}_s \underline{x}_s^T$$

Érdemes a paraméterek számát oly módon csökkenteni, hogy a túlzottan korreláló paraméterek helyett csak azok valamilyen lineáris kombinációját tartjuk meg. Az ilyen kapcsolatok feltárására, ezáltal dimenziócsökkentésre használhatjuk a korreláci-

ós mátrix legnagyobb sajátértékeihez tartozó sajátvektorok (főkomponensek) alapján képzett transzformációs mátrixot ($\underline{\Delta}$). Erre többek között azért van szükség, mert a következő lépés numerikus problémákat vet fel, ha az adatok túlságosan korrelálnak [13].

LDA: A főkomponens analízis után az adatvektorokat kisebb dimenziójú térbe vetítettük a Fisher-féle diszkrimináns analízisnek (LDA) megfelelően [5] kapott mátrix segítségével ($\underline{W}$). Az így nyert ötdimenziós vektorokkal végeztük a tanítást – ahol 1 illetve 2 Gauss függvény keverékével próbáltuk a sűrűségfüggvényeket közelíteni – és a felismerést.



2. Ábra: A válogatott jellemzővektoron képzett transzformációk tanítás és felismerés előtt

A gépi felismerő rendszerek teljesítőképessége keresztkiértékeléssel jellemezhető. Esetünkben ez azt jelenti, hogy például a 34 beszélővel készített adatbázison, 34 tanítási és felismerési tesztet futtatunk, az egyik beszélőt mindig kihagyva a tanításból, a felismerési teszteket pedig a kihagyott beszélő adatain mértük. A 34 teszt eredményét átlagolva kaptuk meg a rendszerünk felismerési eredményét.

6 Eredmények

A sokbeszélős magyar adatbázison (HU_SI) elért beszélőfüggetlen eredmények a *4. táblázat*ban láthatók, az átlagos felismerési pontosság 42,9%. Figyelembe véve, hogy nem színészek által produkált érzelmeket hordozó, tartalmilag kötetlen felvételekről van szó, az eredmény a nemzetközi publikációkkal összemérhető, és az emberi közelítőleg 60%-os hatásfokhoz képest is bíztató.

4. Táblázat: Magyar, beszélőfüggetlen (HU_SI) érzelemfelismerés eredménye

Érzelem	Felismerési arány [%]
Harag	42,7
Undor	43,5
Öröm	33,3
Neutrális	62,0
Szomorúság	36,7
Meglepődés	39,0
Átlag	**42,9**

A beszélőfüggő esetben – ahol beszélőnként külön-külön tanítottunk és teszteltünk, majd a független eredmények átlagát vettük – felismerőnk az alábbi eredményeket mutatta. (*5. táblázat*).

5. Táblázat: Kétbeszélős magyar adatbázison (HU_SD) elért átlagos felismerési hatásfokok

Érzelem	Felismerési arány [%]
Harag	50,0
Undor	80,0
Öröm	80,0
Neutrális	60,0
Szomorúság	53,3
Meglepődés	38,0
Átlag	**60,2**

Ebben az esetben a felismerő sokkal jobban teljesített, érzelmi kategóriánként átlagolva 60% körül. A kevesebb tanítóminta dacára a beszélőfüggő felismerési eredmények lényegesen jobbak lettek.

Azért, hogy rendszerünket másokéval is összehasonlíthassuk, a kísérleteket lefuttattuk a német adatbázison is (*6. táblázat*).

6. Táblázat: Tízbeszélős, német adatbázison (DE_SI) elért felismerési eredmények

Érzelem	Felismerési arány [%]
Harag	65,6
Unalom	76,5
Undor	80,3
Félelem	73,0
Öröm	51,3
Semleges	73,7
Bánat	82,0
Átlag	**71,8**

Meglepően magas felismerési eredményt sikerült elérni, mely a nemzetközi irodalomban használt komplexebb tanuló rendszerek (például SVM) eredményeivel is összevethető [14]. Véleményünk szerint ez a magas felismerési eredmény annak köszönhető, hogy az adatbázisban kötött a szöveges tartalom, és ez korlátozza az érzelemkifejezés lehetőségeit. Nem szabad elfelejteni azt sem, hogy itt színészek által produkált felvételekről van szó, melyek nem adhatják vissza az egyes érzelmek teljes skáláját. Általában is elmondható, hogy a színészekkel készült felvételek „hevesebb" érzelmeket tartalmaznak.

7 Összefoglalás

Megmutattuk, hogy statisztikai módszerekkel, pusztán a beszéd akusztikus jellemzői alapján, a szöveges tartalom figyelembe vétele nélkül megfelelő érzelemfelismerési eredményeket lehet elérni. Ez különösen beszélőfüggő esetben lehet igen hatékony. Annak érdekében, hogy ilyenkor ne kelljen egy teljesen új felismerőt betanítani, amihez sok adat kell, érdemes lenne a beszédfelismerésnél is gyakran használt beszélőadaptációt alkalmazni – ebben az irányban tervezzük a további vizsgálatokat.

Az eredmények alapján arra is következtethetünk, hogy az „amatőrök" és a színészek által keltett beszéd érzelmi töltete különböző jellegű, melyek közül az utóbbinak a felismerése jóval eredményesebb lehet.

8 Köszönetnyilvánítás

A kutatást az NKFP-2/034/2004-es projekt keretében az NKTH támogatta.

Bibliográfia

1. Bernáth, László – Révész, György 1994. A pszichológia alapjai. Tertia, Budapest, 1994.
2. Blouin, Christophe - Maffiolo, Valerie 2005. A study on the automatic detection and characterization of emotion in a voice service context. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal 469-472.
3. Burkhardt, Felix - Paeschke, Astrid – Rolfes, Miriam – Sendlmeier, Walter - Weiss, Benjamin 2005. A Database of German Emotional Speech, In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 1517-1520.
4. Cichosz, Jaroslaw – Slot, Krzysztof 2005. Low-Dimensional Feature Space Derivation for Emotion Recognition. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 477-480.
5. Duda, Richard O. - Hart, Peter E. - Stork, David G. Pattern Classification (Second Edition) 2000. John Wiley & Sons Inc, ISBN: 0-471-05669-3, New York
6. Fernandez, Raul – Picard, Rosalind W. 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 473-476.
7. Kwon, Oh-Wook – Chan, Kwokleung – Hao, Jiucang – Lee, Te-Won 2003. Emotion Recognition by Speech Signals. In Proceedings of EUROSPEECH-2003. Geneva, Switzerland, 125-128.
8. Laukka, Petri 2004. Vocal Expression of Emotion, PhD Thesis. Uppsala University, Uppsala.
9. Luengo, Iker - Navas, Eva - Hernáez, Inmaculada – Sánchez, Jon 2005. Automatic Emotion Recognition using Prosodic Parameters. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 493-496.
10. Petrushin, Valery A. 2000. Emotion recognition in speech signal: experimental study, development, and application. In Proceedings of ICSLP-2000. vol.2. Beijing, China, 222-225.
11. Scherer, Klaus R. 2000. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In Proceedings of ICSLP-2000, vol.2. Beijing, China, 379-382.
12. Scherer, Klaus R – Banse, Rainer - Wallbott, Harald G. 2001. Emotion Inferences from Vocal Expression Correlate Across Language and Cultures. Journal of Cross-Cultural Psychology. vol. 32. No. 1. 76-92.
13. Schlüter, Ralf - Zolnay, András - Ney, Hermann 2006. Feture Combination using Linear Discriminant Analysis and its Pitfalls. In Proceedings of INTERSPEECH-2006. Pittsburgh, Pennsylvania, 345-348.
14. Schuller, Björn – Müller, Ronald - Land, Manfred – Rigoll, Gerhard 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features Within Ensenmbles. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 805-808.
15. Ververidis, Dimitrios – Kotropoulos, Constantine – Pitas, Ioannis 2004. Automatic emotional speech classification. In Proceedings of ICASSP'04. vol. 1. Philadelphia, Pennsylvania, 593-596.
16. Ververidis, Dimitros - Kotropoulos, Constantine 2004. Automatic Speech Classification to five emotional states based on gender information. In Proceedings of EUSIPCO-2004. Vienna, Austria, 41-344.