Érzelmes beszéd gépi előállítása érzelem specifikus beszédadatbázisok felhasználásával

Fék Márk, Zainkó Csaba, Németh Géza

Beszédtechnológiai Laboratórium,
Távközlési és Médiainformatikai Tanszék,
Budapesti Műszaki és Gazdaságtudományi Egyetem,
1117 Budapest, Magyar tudósok körútja 2.
{fek, zainko, nemeth}@tmit.bme.hu

Kivonat: Tanulmányunkban megvizsgáljuk hogyan lehet érzelem specifikus beszédadatbázisok felhasználásával gépileg érzelmes beszédet előállítani. Kísérletünket magyar nyelvre végeztük, de a módszer nyelvfüggetlen. Felvettünk egy szemantikailag semleges tartalmú mondatot és 26 logatomot amelyek a mondat szintetizálásához szükséges diádokat és CVC triádokat tartalmazták. A hanganyagot egy profi színésznő mondta fel a hat alapérzelemnek megfelelően, illetve semleges érzelmi változatban. A logatomok felhasználásával 7 érzelemfüggő beszédelem adatbázist hoztunk létre. A 7 beszédelem adatbázist összepárosítva a természetes mondatokból kinyert 7 prozódiai kontúrral 49 szintetizált mondatot állítottunk elő. A logatomokban, illetve a természetes és a szintetizált mondatokban hallható érzelmeket 194 tesztalany értékelte ki. A tesztelők a logatomok 99%-ban, illetve az összes természetes mondatban szignifikánsan a véletlen találgatás szintje felett ismerték fel a színésznő által kifejezett érzelmeket. Az érzelmek azonosítási aránya egyes szintetizált mondatok esetén meghaladta a természetes mondatokét.

1 Bevezetés

A napjainkban elterjedten használt elemösszefűzéses beszédszintetizátorok általában adnak valamilyen lehetőséget a prozódiai paraméterek (alapfrekvencia, hangidőtartamok, intenzitásmenet) vezérlésére. A gépi beszéd hangszínezetét azonban alapvetően a beszédadatbázis határozza meg, amelyből az összefűzendő elemeket kiválasztják. Ismert, hogy a beszédben az érzelmeket a prozódia és a hangszinezet együttesen hordozza [1].

Korábbi kísérletek több nyelv esetében vizsgálták [2,3] az érzelem specifikus beszédadatbázisok használatának lehetőségét elemösszefűzéses beszédszintézis esetében. A kísérletek minegyikében a bemondók felolvastak egy szöveget az előzetesen definiált érzelmeknek megfelelő változatokban. A felvett beszédből minden érzelmehez egy-egy érzelem specifikus adatbázist hoztak létre. Elemkiválasztáson alapuló beszédszintézis segítségével szemantikailag semleges mondatokat állítottak elő az egyes érzelmi adatbázisokból.

A [2] és [3]-ban leírt kísérletekben a szintetizált mondatok prozódiai paramétereit természetes mondatokról másolták, illetve [2] második kíséletében a beszédszintézis algoritmus állította elő a prozódiát. Ezután meghallgatásos teszteket végeztek, hogy meghatározzák a szintetikus mondatokkal kifejezni szándékolt érzelmek felismerési arányát.

Montero és szerzőtársai [2] szemantikailag semleges szöveget vettek fel spanyolul egy professzionális színész segítségével, aki négy érzelemmel és semleges stílusban mondta be a szöveget. A szerzők a a természetes mondatokról átmásolt prozódia segítségével szintetizált mondatokon a véletlen szintnél magasabb azonosítási arányt értek el (semleges 76%, öröm 62%, meglepetés 91%, szomorúság 81%, harag 95%) egy 6 választási lehetőséget megengedő meghallgatásos tesztben (amely a "nem azonosítható érzelem" opciót is tartalmazta). Hasonló eredményeket értek el (kivéve a meglepetésre 53%) az érzelmi adatbázisokon betanított, automatikusan generált prozódia használata esetében is.

Bulut és szerzőtársai [3] örömöt, szomorúságot, haragot és semlegességet közvetítő rövid szövegrészeket vettek fel egy fél-profi színésznő segítségével. Az adatbázisokból előállított mondatokat a véletlen találgatás szintje felett azonosították (harag 86%, szomorúság 89%, öröm 44%, semleges 82%) egy 4 választási lehetőséget tartalmazó meghallgatásos tesztben.

Schröder és Grice [4] egy teljes német diádos elemkészletet vettek fel három különböző vokális erőfeszítés mellett (lágy, modális és hangos). Meglátásuk szerint a különböző hangszalag feszességeknek megfelelő vokális erőfeszítések fontos szerepet játszanak az érzelmek kifejezésében. A diádokat értelem nélküli hordozó szavakba (logatomokba) ágyazták és egy német anyanyelvű bemondó olvasta fel őket, akit megkértek, hogy állandóan tartott hangmagasságon beszéljen. Egy (elemkiválsztás nélküli) elemösszefűzéses szintetizátort használtak a tesztmondatok előállításához. A meghallgatásos teszt folyamán kimutatták, hogy a szintetizált mondatokban a szándékoltnak megfelelően észlelhető a vokális erőfeszítés.

A fent felsorolt munkák ellenére még mindig kérdéses, hogyan lehet olyan érzelmes beszédszintetizátort készíteni, amely a megfelelő hangszínezettel állítja elő az egyes érzelmeket. Tanulmányunkban az érzelem specikfikus diád és triád elemkészlet használatát vizsgáljuk érzelmes beszéd előállítására.

Kísérletünk hasonló a korábbiakhoz a következő lényegi eltérésekkel. Az általunk használt szintézis módszer elemkiválasztás nélküli elemösszefűzésen alapul és [4]-hez hasonlóan szintén értelem nélküli logatomokat vettünk fel, de a bemondott logatomok közvetlenül hordozták az egyes érzelmeket, hasonlóan a [2] és [3]-ban felolvasott szövegekhez. A mi beszédelem-készletünk mind diádokat, mind CVC triádokat tartalmazott növelve az összefűzés hangzásának folytonosságát a csak diádokból álló elemkészlethez képest. Meghallgatásos teszttel ellenőriztük az érzelmek felismerhetőségét az egyes logatomokban. A forrásanyag ilyen jellegű formális ellenőrzése a korábbi munkákban nem történt meg. Ellentétben a [4]-ben leírtakkal, a színésznő szabadon variálhatta a hangmagasságát, amely jelentős változatosságot mutatott a logatomokon belül és azok között, illetve a mondtokon belül is. Emiatt a szintézis folyamán alkalmanként jelentős alapfrekvencia módosításra volt szükség, amely nemkivánt torzulást vitt a szintetizált jelbe. Az elemkiválasztáson alapuló szintézis módszerek is érzékenyek az érzelmes beszédre jellemző jelentős alapfrekvencia

változásokra, ami fokozottan jelentkezhet a [2] és [3]-ban használt kisméretű elemkészletek esetében. Az alapfrekvencia változtatás okozta torzulások feltehetően befolyásolták ezen kísérletek eredményeit. Az idézett szerzőkhöz képest több érzelmet vizsgáltunk, ami több választási lehetőséghez vezetett a meghallgatásos tesztek folyamán rontván a várható érzelemazonosítási arányt. Kísérletünket magyar nyelvre végeztük, de a korábbi kísérletekhez hasonlóan a módszer más nyelvek esetében is alkalmazható.

2 A hanganyag felvétele

A kísérletben a következő szemantikailag semleges mondatot használtuk: „A menüben minden szükséges információ elhangzik”. Meghatároztuk mondat fonetikus átírásának megfelelő, diádokból és CVC triádokból álló szekvenciát. Összeségében 14 CVC triádot, 3 VV és 9 CC diádot használtunk. A diádokat és triádokat jelentés nélküli szavakba (logatomokba) ágyaztuk, amelyek a CC diádok esetében kétszótagosak, a VV diádok és a CVC triádok esetében három szótag hosszúak voltak.

A semleges mondatot és a 26 logatomot 7 változatban vettük fel a hat alapérzelmet (öröm, harag, meglepetés, undor, szomorúság, félelem) és semlegességet kifejezve. A hanganyagot egy professzionális (30 éves) magyar színésznő mondta be, akinek már volt tapasztalata beszédszintézishez szükséges elemkészlet felvételével, illetve korábbi kísérletekhez már mondott fel érzelmes mondatokat. Minden egyes érzelmhez iteratívan több (4-7) változatban kerültek bemondásra a logatomok, amíg a cikk szerzői megfelelőnek nem ítélték őket. A felvétel után informális meghallgatásos tesztet végeztünk, melynek során a cikk két szerzője kiválasztotta az érzelmileg legmeggyőzőbb logatomokat és mondatokat. A logatomok kb. 20%-át nem volt elég kifejező, ezért azokat egy második alkalommal újra bemonddattuk. Az első felvétel folyamán készült legmeggyőzőbb mintákat lejátszottuk a színésznőnek az egyes érzelmek másodszori felvétele előtt. A logatomok több (4-6) változatban is bemondásra kerültek, melyek közül a cikk két szerzője kiválaszotta a legjobban sikerülteket.

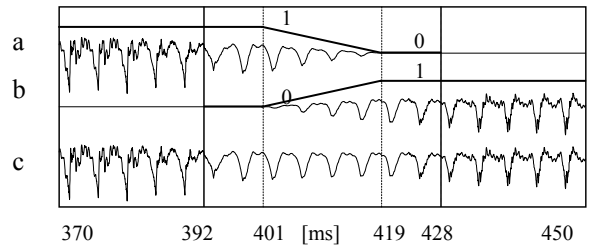
A felvételeket süket szobában készítettük AKG C-414 B-ULS kondenzátor mikrofonnal, a kardioid karakterisztika és a 75 Hz-es aluláteresztő szűrés bekapcsolásával. A mikrofon elé pop szűrőt helyeztünk. A beszédjelet 44.1 kHz-en, mintánként 16 biten digitalizáltuk.

3 Szintetizált minták előállítása

A logatomokból készített hét érzelem specifikus beszédelem adatbázist és a hét változatban bemondott mondatokat az összes lehetséges módon összepárosítottuk. Mind a 49 pár esetén a természetes mondat prozódiáját az érzelem specifikus adatbázisból szintetizált mondatra másoltuk.

A szintézist és a prozódia másolását a következő módon végeztük. A banghatárokat manuálisan bejelöltük a felvett logatomokban és mondatokban. Az alap-

frekvencia menetet a Praat-ban implementált autokorrelación alapuló algoritmus [5] segítségével detektáltuk. A cél lapfrekvencia görbék periódusonként egy frekvencia értéket definiáltak. Az intenzitást 8 ms-onként számítottuk 32 ms széles ablak segítségével. A Praat-ban implementált PSOLA alapú prozódia módosító algoritmust [5] használtuk a hangidőtartamok, az alapfrekvencia és az intenzitások diádok és triádokra másolásához.



1. ábra: Szomszádos triádok összefűzése időtartománybeli átmosással (50%-os átfedéssel).

Fontos megjegyezni, hogy a diádok két egymást követő hangot teljesen lefednek, míg a triádok három egymás követő hang teljes időtartamán át tartanak. Mint az 1. ábrán látható, a prozódia módosítás után a diádokat és a triádokat időtartománybeli átmosással fűztük össze, olymódon, hogy a szomszédos szegmensek első, illetve utolsó hangjai között 50% átfedés legyen. Az ábrán a függőleges folytonos vonalak a hanghatárokat, a függőleges szaggatott vonalak pedig az átfedési tartományt jelölik. A felső két görbe (a és b) mutatja az összefűzött triádokat az átlapoló ablak feltüntetésével. Az alsó görbe (c) mutatja az összefűzés után kapott jelet.

4 Meghallgatásos teszt

Egy webes felületen végzett meghallgatásos teszt során meghatároztuk, hogy a tesztelők milyen érzelmeket azonosítanak a logatomokban, illetve a természetes és a szintetizált mondatokban. 208 magyar anyanyelvű alany vett részt a vizsgálatban. A tesztalanyok többsége motivált informatika szakos hallgató volt.

14 tesztelő eredményét nem vettük figyelembe, mert vagy nem fejezték be a tesztet, vagy véletlenszerű válaszokat adtak. A kizárt tesztelők közül néhányan hanglejátszási problémát jelzetek. A maradék 194 résztvevő közül 159 férfi és 35 nő volt, átlagos koruk 23 év. 83 tesztelő fej- vagy fülhallgatót használt, míg 111 tesztelő hangszóró segítségével hallgatta meg a felvételeket.

A teszt hat részből állt és átlagosan 18 percet vett igénybe. A résztvevők az első részben a logatomokat, a negyedikben pedig a természetes és szintentizált mondatokat értékelték ki. A második és a negyedik rész egy a jelen tanulmánytól független vizsgálathoz tartozott. A harmadik és a hatodik részt azért iktattuk be, hogy ki tudjuk szűrni a véletlenszerűen válaszoló tesztelőket. Ebben a két részben egy 5-fokozatú skálán kellett minősteni egy mondat természetes, illetve három különböző

szintetizátorral előállított változatát. Négy tesztelőt kizártunk, mivel az általuk adott minősítések nem tükrözték a felvételek közötti különbséget.

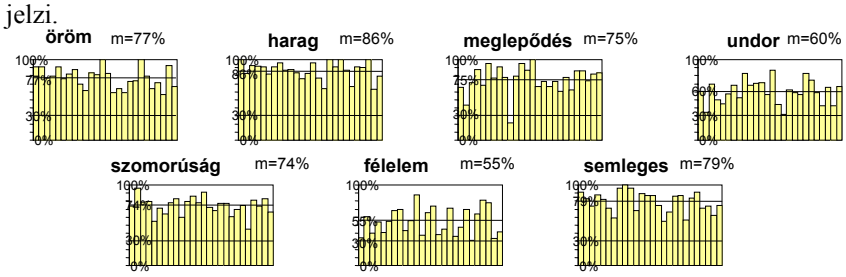
Hogy csökkentsük a tesztelők terhelését, egy résztvevőnek nem kellett az összes mondatot, illetve logatomot meghallgatnia. A 49 szintetizált és 7 természetes mondatot összekevertük és véletlen választással két 28-28 mondatot tartalmazó csoportra bontottuk. 98 tesztelő csak az első, míg 94 csak a második csoportot hallgatta meg. Hasonlóképpen a logatomokat is hét egyforma méretű csoportra osztottuk. Egy ilyen csoportot legalább 23-an hallgattak meg.

A tesztelőknek 7 lehetőség (hat alapérzelem, illetve semlegesség) közül kellett kiválasztani a logatomok, illetve a természetes és a szintetizált mondatok által kifejezett érzelmet. A teszttel kapcsolatos ismertető szöveget ugyanaz a színésznő olvasta fel, mint aki a hangját adta a kísérlethez, így a résztvevők megismerhették érzelmileg semleges beszédstílusát. Azért, hogy ne lehessen ezt a részt átugorni, a továbblépéshez meg kellett adni egy az ismertető végén elhangzó, véletlenszerűen generált kódot. Az egyes elemek között a tesztelők maguk szabályozták a továbblépést. Egy mintát akárhányszor meg lehetett hallgatni, viszont egy korábban már kiértékelt mintára nem volt megengedett a visszatérés. A minták lejátszási sorrendje tesztelőnként véletlenszerűen változott. A tesztben szereplő hangminták honlapunkon elérhetőek: http://speechlab.tmit.bme.hu

5 Eredmények

Annak eldöntésére, hogy az azonosítási arányok szignifikánsan a 14%-os véletlen szint felett vannak-e binomiális próbát ($p<0.05$) használtunk nemegyenlő (1:7) arányokkal. A logatomos teszt esetében (23 tesztelő) a 30% feletti azonosítási eredmény volt szignifikánsan a véletlen szint felett. A mondatok kiértékelésénél (94 tesztelő) a 21% feletti azonosítási arány volt szignifikánsan a véletlen szint felett.

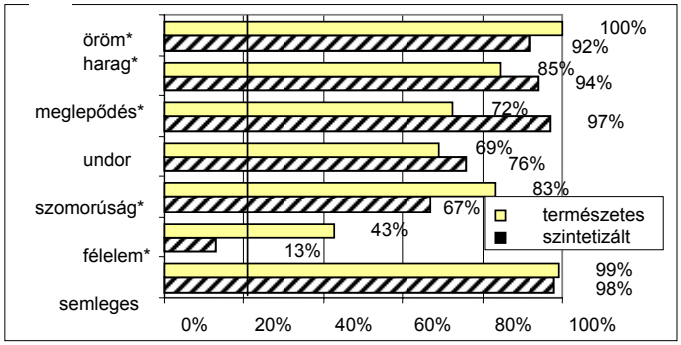
A 2. ábra minden egyes logatomra mutatja a kifejezni szándékolt érzelem felismerési arányát. Az alsó vonal mutatja a szignifikánsan a véletlen szint feletti szintet. A felső vonal az adott érzelemhez tartozó adatbázis átlagos felismerési arányát jelzi.



2. ábra: Érzelmek azonosítási aránya a 26 logatomra.

A kifejezni szándékolt érzelmet a 182 logatomból 180 esetében szignifikánsan a véletlen szint felett azonosították. A harag volt az átlagban legjobban felismert érzelem (86%), míg az undor (60%) és a félelem (55%) sokkal kevésbé volt felismer-

hető. Ugyanakkor több undort és félelmet kifejező logatom 80% felett teljesített, ami mutatja, hogy a tesztelők ezeket az érzelmeket képesek rövid mintákon is felismerni. A többi logatomra kapott alacsonyabb eredmény oka valószínűleg az, hogy a színésznőnek nehéz volt ezen érzelmek konzisztensen az összes mintán kifejezni.



3. ábra: Érzelmek azonosítási aránya a természetes mondatokra, illetve az egyező érzelmi adatbázissal és prozódiával előállított szintetizált mondatokra. A szignifikáns különbséget * jelöli.

A 3. ábra mutatja szándékolt érzelem felismerési arányát természetes mondatokra, illetve az egyező érzelmi adatbázissal és prozódiával rendelkező szintetizált mondatokra. Két-mintás t-próba ($p<0.05$) segítségével ellenőriztük, hogy a természetes és a szintetizált mondatok felismerési arányai között szignifikáns-e az eltérés. A függőleges vonal által jelölt értékénél magasabb felismerési arányok szignifikánsan a véletlen szintje felettiek.

A természetes mondatok közül a legjobb eredményt az öröm (100%) érte el. A haragot (85%) semlegesnek ítélte a tesztelők 9%-a. A meglepetést (72%) örömnek ítélte a tesztelők 28%-a. Ennek az lehetett az oka, hogy a színésznő pozitív meglepetést kifejezendő mosolyogva fejezte be a mondatot. A színésznőnek nehéz volt az undor (69%) kifejezése, amit 12% örömnek, illetve 10% haragnak gondolt. A szomorúságot (83%) semlegesnek ítélte a tesztelők 11%-a. A legkevésbé felismert félelmet (43%) a tesztelők 30%-a semlegesnek, míg 22%-a szomorúnak tartotta. A színésznő a mondat második felében abbahagyta a hangmagasság remegtetését, ami magyarázatot adhat a semleges válaszok magas arányára. Ráadásul a színésznő lágy hangon mondta be a mondatot, ami megmagyarázhatja miért tartották azt sokan szomorúnak. A mondat még a félelmet kifejező logatomok átlagánál is rosszabb eredményt ért el, ami szintén mutatja a kevésbé szerencsés realizációt. A semleges mondatot (99%) könnyen azonosították a tesztelők.

A szintetizált örömöt (92%) könnyen felismerték. A kapott arány jelentősen meghaladja a [2]-ben és [3]-ban közölt korábbi eredményeket. [2] szerzői feltételezték, hogy a mosolyt tartalmazó, illetve nem tartalmazó beszédelemek keveredése csökkenti a szintetizált mondat örömteli hangszínezetét. Az öröm viszonylag alacsony felismerési arányát [3] szerzői részben a kevésbé sikeres színészi teljesítménnyel magyarázták. A természetes örömöt kifejező mondatra, illetve logatomokra kapott magas felismerési arányok mutatják, hogy a fenti problémák egyike sem jelentkezett

az általunk végzett kísérletben. A szintetizált haragot (94%) szintén könnyen felismerték. Érdekes módon a szintetizált harag felismerési aránya szignifikánsan meghaladta a természetes harag felismerését. Megfigyeltük, hogy a prozódia módosító algoritmus érdesebb hangzásúvá tette a beszédet, ami magyarázatot adhat a jobb eredményre. Hasonló megfigyelést [2] is közölt. A szintetizált meglepetést (97%) könnyen felismerték, és a természetes mondattal ellentétben nem tévesztették össze az örömmel. Ez megmagyarázható, ha figyelembe vesszük, hogy a szintetizált mondat hangszínezete (a természetes mondattal ellentétben) nem ment át mosolygásba a mondat végén. A szintetizált undort (76%) a tesztelők 13%-a a haraggal keverte össze. Továbbá a szintetizált felismerési arány meghaladta (de nem szignifikánsan) a természetes mondatét. Mindkét megfigyelés magyarázza a szintetizélt mondat érdes hangszínezete. A szintetizált szomorúságot (67%) a tesztelők 17%-a semlegesnek, míg 7%-uk félelemnek ítélte. Megfigyeltük, hogy a szintetizált szomorú mondat a többi érzelemhez képest torzabban hangzott, ami megmagyarázza a kapott viszonylag alacsony eredményt. A szintetizált mondaton leginkább eltorzult szakaszokat megvizsgálva észrevettük, hogy az azoknak megfelelő szomorú logatomok szabálytalan zönge periódusokat (glottalizációt) tartalmaztak. A szomorú beszédelem adatbázis használatával előállított összes mondatban hasonló torzítást találtunk. Továbbá azt is észrevettük, hogy a természetes szomorú mondat is tartalmaz glottalizációt. A színésznő valószínűleg részben glottalizációval fejezte ki a szomorúságot, de a szabálytalan zönge periódusok megzavarták az alapfrekvencia detektáló és módosító algoritmusok működését, ami torzulást okozott. A szintetizált félelmet (13%) a véletlen szint alatt ismerték csak fel. A mondatot a tesztelők 64%-a haragnak, 16%-a semlegesnek, 6%-a pedig undornak azonosította. Az alacsony felismerési eredmény részben a természetes félelmre kapott alacsony eredménnyel magyarázható. A szintetizált semlegesség (98%) majdnem olyan jó eredményt ért el, mint természetes párja.

Az 1. táblázat mutatja az összes (49) adatbázis és prozódia párosításra kapott felismerési eredméyeket. Az egy-egy adott érzelmet legjobban kifejező párosításhoz tartozó cellákat árnyékolással jelöltük meg. Eltérő adatbázis és a prozódia esetén a függőleges nyíl jelzi, ha a mondatot az adatbázis érzelmi tartalmának megfelelően azonosították. A vízszintes nyíl jelzi, ha a mondatot a célprozódiával megegyezően azonosították. A felismert érzelmet külön kiírtuk, ha az sem az adatnázisnak, sem a prozódiának nem felelt meg. A félkövérrel jelzett számok jelzik az adott párosításhoz tartozó legmagasabb felimerési arányt. A táblázatban csak a véletlen találgatás szintjét szignifikánsan meghaladó értékeket tüntettük fel.

Az öröm és a félelem kivételével minden érzelem esetében az egyező adatbázis és prozódia párosítása adta a legjobb eredményt. A félelmet csak 13% azonosította egyező prozódia és adatbázis esetén. Sokkal jobb eredményt hozott, ha természetes félelem mondat prozódiáját a szomorú (62%), vagy a meglepődés (60%) adatbázisból előállított mondatra másoltuk. Az így kapott felismerési arányok még a természetes félelemre (43%) kapott eredményt is meghaladták. Megfigyeltük, hogy a félelem adatbázisban szereplő logatomok „préselt” hangzással lettek bemondva, ellentétben a szomorúságot illetve félelmet kifejező logatomokkal. Ez megmagyarázhatja, miért lett „lágyabb” az utóbbi két adatbázisból szintetizált mondatok hangzása. A 2. ábra alapján a „préselt” logatomok többé-kevésbé a szándékolt érzelmet fejezték ki

(átlag=55%), de a prozódia módosítás által bevitt érdes hangzás a harag felé (64%) tolta el a mondat felismerését.

1. táblázat: Érzelmek azonosíátsi aránya szintetikus mondatokra az összes prozódia és adatbázis párosítás esetén.

↑adatbázis ←prozódia	öröm	harag	meglepetés	undor	szomorúság	félelem	semleges
öröm	**92%**	← 2%, **↑74%**	**←35%,** ↑14% félelem 33%	← 2%, ↑ 14% **félelem 37%**	← 3%, **↑ 50%** félelem 29%	←11%, **↑35%** szomorú 23%	←1%, **↑ 50%** szomorú 23%
harag	←1%, **↑ 93%**	**94%**	←12%, ↑12% **félelem 41%**	**←63%**, ↑ 22%	← 3%, **↑39%** semleges 27% félelem 26%	**← 60%,** ↑ 2%	←35%, **↑ 39%**
meglepetés	**←68%,** ↑ 31%	**← 56%,** ↑41%	**97%**	**← 82%**, ↑ 5%	**←75%**, ↑ 5%	**← 85%**, ↑ 1%	**← 87%,** ↑ 4%
undor	←13%, **↑ 82%**	← 48%, **↑49%**	←23% , ↑10% **öröm 29%**	**76%**	← 14%, **↑45%**	**← 71%,** ↑ 2%	**← 68%**, ↑ 9%
szomorúság	← 1%, **↑ 78%**	← 0 %, ↑ 33% **semleges 34%** undor 30%	← 11%, ↑ 12% **semleges 37%**	←16%, **↑ 43%**	**67%**	← 20%, ↑12 % **undor 36%** semleges 28%	← 32%, **↑ 52%**
félelem	← 3%, **↑ 74%**	←1%, **↑87%**	**←60%**, 11%	←6%, ↑22 % **harag 58%**	**←62%**, ↑18%	13% **harag 64%**	← 9%, **↑77%**
semleges	←22%, **↑ 75%**	←33%, **↑54%**	**←79%**, ↑ 9%	**←80%**, ↑13%	**←79%**, ↑13%	**←85%**, ↑ 0%	**98%**

Az öröm adatbázis harag prozódiával kombinálva (93%) némileg magasabb felimerési arányt ért el, mint öröm prozódiával (92%). Ez azt jelentheti, hogy egy közös prozódiai modell használható ezen érzelmek szintetizálásához.

Ha az 1. táblázatban megvizsgáljuk a semleges prozódiához tartozó párosításokat, észrevehetjük, hogy a szintetizált harag (54%) és öröm (75%) kifejezésében inkább az adatbázis játszik szerepet. A semleges adatbázishoz tartozó párosítások alapján látható, hogy a meglepetést (87%) és az undort (68%) inkább a prozódiával lehet kifejezni. Hasonló eredményre jutottak [2] és [3] szerzői is, az undort (ami nem szerepelt a korábbi kísérletekben) és a szomorúságot kivéve, mely inkább a prozódiától függpt mindkét korábbi kísérletben. Az utobbi eltérés egy lehetséges magyarázata az, hogy a mi kísérletünkben a színésznő glottalizáció segítségével fejezte ki a szomorúságot. A szintézishez használt módszer nem tudta kezelni a szabálytalan zönge periódusokat, ami akadályozta a szomorú prozódia szintetikus semleges mondatra (13%) másolását. Ezt a feltételezést megerősíti a szomorú adatbázis párosítása öröm

(szomorúság 50%), harag (szomorúság 39%) és undor (szomorúság 45%) prozódiával. Ezen mondatok prozódiáját jobban sikerült a szomorú adatbázisból szintetizált mondatra ültetni, és mindegyik esetben a szomorúság volt a legnagyobb arányban felismert érzelem.

6 Összefoglalás

A kísérlet igazolta, hogy a szintetizálással előállított harag és öröm meggyőzően kifejezhető érzelem specifikus adatbázisok segítségével. Az örömteli beszédelem adatbázis és a haragos prozódia kombinációja meggyőzően fejezte ki az örömöt. Ennek alapján elégséges lehet egy közös prozódiai modul használata ezen érzelmek esetében. A meglepetés és az undor pusztán prozódia segítségével is kifejezhető, de a megfelelő hangszínezet tovább növeli felismerhetőségüket.

Az undort és félelmet kifejező természetes mondatok kevéssbé voltak meggyőzőek, mint a többi mondat. Ez részben megmagyarázza ezek szinetizált változataira kapott rosszabb felismerési arányokat is. A prozódia módosító algoritmus érdessé tette a félelem adatbázisban szereplő erősen „préselt" logatomok hangzását. Az adott technológiai korlátok mellett előnyös lehet kevésbé „préselten" képzett logatomok használata.

A természetes szomorú felvételek szabálytalan zöngeperiódusokat tartalmaztak, amelyeket nem tudott megfelelően kezelni a szintézis algoritmus. A szintetikus mintákban megjelenő járulékos torzítás megmagyarázza a gépi szomorúságra kapott rosszabb felismerési arányt. A szintetizált szomorú kifejező ereje javulhat mesterséges glottalizáció hozzáadásával, ugyanakkor a prozódiamódosítás technológiai korlátai miatt a beszédelem adatbázisokban kerülni kell a glottalizációt.

Az elemösszefűzéses beszédszintézis segítségével lehetséges meggyőzően érzelmeket kifejezni, viszont a különböző érzelmekhez tartozó beszédelem adatbázisok felvétele igen megterhelő. Az egyes érzelmekre jellemző hangszínezet feltehetően inkább marad konzisztens egy rövid logatomon, mint egy teljes mondaton belül. Egy hangszínezet módosító algoritmust egy viszonylag kisméretű, de konzisztens logatom adatbázison betanítva akár hosszabb, elemkiválasztáshoz használt beszédadatbázisokat is érzelmileg kifejezővé lehetne tenni.

7 Köszönetnyilvánítás

A munkát a Nemzeti Kutatási és Technológiai Hivatal támogatta (NKFP 2/034/2004).

Bibliográfia

1. Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K.R., "Evidence for the independent function of intonation contour type, voice quality, and f0 range in signalling speaker affect", Journal of the Acoustic Society of America, 78 (2), pp. 435–444, 1985.
2. Montero, J.M., Arriola, G.J., Colas, J., Enriquez, E., and Pardo, J.M., "Analysis and Modeling of Emotional Speech in Spanish", Proc. of ICPhS, pp. 957-960, 1999.
3. Bulut, M., Narayanan, S. S., and Syrdal, A. K.: "Expressive Speech Synthesis Using a Concatenative Synthesizer", In. ICSLP-2002, pp. 1265-1268, 2002.
4. Schröder, M., Grice, M., "Expressing Vocal Effort in Concatenative Synthesis", Proc. of ICPhS, Barcelona, Spain, pp. 2589-2592, 2003.
5. Boersma, P., "Praat, a system for doing phonetics by computer", Glot International, vol. 5:9/10, pp. 341-345, 2001.