Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval

Csapó Tamás Gábor[1], Németh Géza[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{csapot, nemeth}@tmit.bme.hu

Kivonat: A prozódiai változatossággal kiegészített szövegfelolvasó rendszer olyan alkalmazásokban lehet hasznos, ahol hasonló jellegű, ismétlődő mondatok szintetizálására van szükség. A cikkben bemutatunk egy új módszert, amellyel egy adott szöveghez különböző prozódiával rendelkező mondatváltozatokat lehet szintetizálni. A prozódia komponensei közül a dallammal és hangsúllyal foglalkozunk az alapfrekvencia (F0) változtatásán keresztül. Ehhez egy statisztikai F0-modellt használunk fel rejtett Markov-modell alapú beszédszintetizátorban. A betanításhoz használt eredeti beszédkorpuszt a SOFM (Self Organizing Feature Map) módszerrel felbontjuk több részkorpuszra. A különböző beszédkorpuszokból betanult modellekkel eltérő dallamú mondatváltozatokat szintetizálunk azonos szöveghez. A mondatváltozatok közötti különbségeket megvizsgálva a szubjektív kísérletek azt mutatják, hogy az alapfrekvencia eltérése sok esetben elég jelentős ahhoz, hogy ez az emberi fül számára is észlelhető legyen.

1 Bevezetés

A szövegfelolvasó rendszerek érthetősége elérte a megfelelő szintet, viszont más tulajdonságokban még hiányosságok fedezhetőek fel. Ezek közé tartozik az emberi beszéd változatossága, amelyet ritkán modelleznek beszédszintetizátor rendszerekben. Az emberi beszédben a prozódia (dallam, hangsúly, ritmus) rendkívül változékony jellemző. Egy-egy mondatot még akarattal sem tudunk többször ugyanúgy elmondani, a mindennapi beszédben pedig nagy különbségek tapasztalhatóak mindegyik fenti jellemzőben. A legtöbb szövegfelolvasó rendszer ezzel szemben determinisztikusan állítja elő a prozódiát, azaz egy-egy bemeneti szöveghez ismételt szintéziskor mindig ugyanaz a prozódia tartozik. Ez sokszor ismétlődő, monoton minták túlzott előfordulásához vezet, ami zavaró lehet a szintetizált beszédben. A prozódiai minták ismétlődése azért fordulhat elő a szövegfelolvasó rendszerekben, mert a beszédszintetizátor mindig a legjobb prozódiát próbálja egy-egy mondathoz rendelni. Így az emberi beszéd változatossága lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, ami a változékonysághoz szokott, könnyen felismerhető, és hosszabb szintetizált beszédrészlet hallgatása során zavaró lehet.

1.1 Prozódiai változatosság

Az a cél, hogy a szövegfelolvasó egy-egy bemeneti mondatához ne mindig ugyanolyan prozódiájú mondatot szintetizáljunk, úgy valósítható meg, ha a bemeneti szöveghez többféle dallammenetet és ritmusszerkezetet tudunk generálni, és ezek közül a rendszer szintéziskor egyet kiválaszt. Ekkor ugyanis csökken a monotonitás, hiszen nem-determinisztikussá válik a mondatokhoz történő dallammenet- és ritmushozzárendelés. Ezen elv segítségével a hasonló szerkezetű egymás után előforduló mondatokhoz is eltérő prozódiát tudunk kialakítani. A cikk további részében a prozódia dallam és hangsúly részével foglalkozunk, az alapfrekvencia (F0) megfelelő beállításán keresztül.

Korábbi kutatásaink során a fenti célt korpuszalapú prozódiai modellel kíséreltük meg elérni. Egy nagyméretű beszédkorpuszból kigyűjtöttük a jellemző mondatdallam-mintázatokat, majd ezeket rendeltük a szintetizálandó szöveghez, hasonlósági mértékként a mondatrészek szótagszámát felhasználva. Ezeket a vizsgálatokat egy diádos beszédszintetizátorral végeztük el [2, 8]. Jelen cikkben a korábbiakhoz hasonló kísérleteket végzünk, statisztikai alapú prozódiai modellt felhasználva.

A nemzetközi szakirodalomban Díaz és Banga foglalkozott a prozódiai változatosság témájával egy korpuszos, elemkiválasztásos beszédszintetizátoron végzett kísérletek keretében [3, 4]. A módszer megőrzi az eredeti beszélő intonációjának változatosságát, mivel az összefűzendő elemek kiválasztásakor több lehetséges sorozatot megtart, melyek mindegyike hasonló minőségű szintetizált beszédet eredményez.

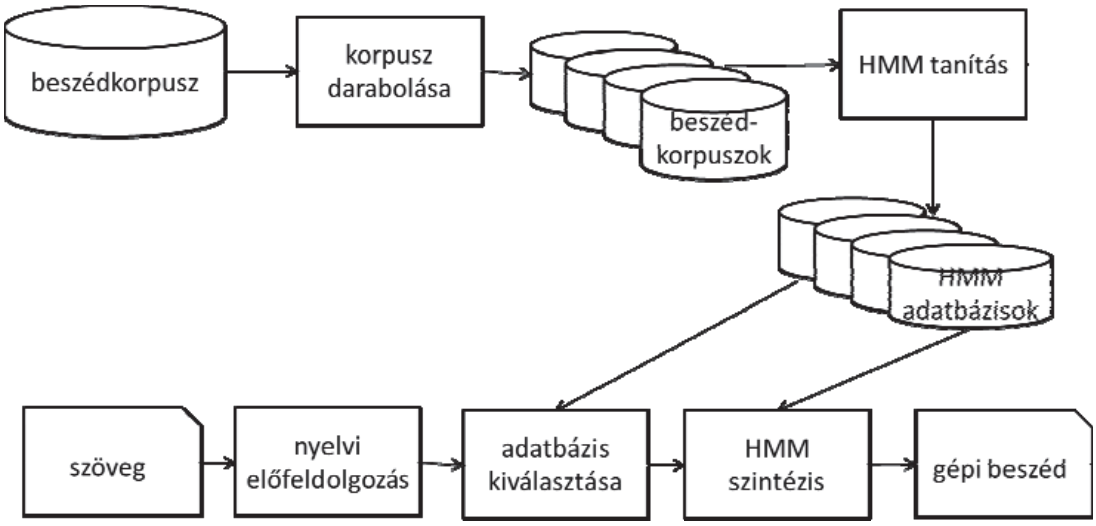
1.2 Rejtett Markov-modell alapú beszédszintézis

A szövegfelolvasó technológiák közül az elmúlt években a rejtett Markov-modell (Hidden Markov Model, HMM) alapú beszédszintetizátorral foglalkozott sokat a szakirodalom, melynek előnye a korábbi megoldásokhoz képest az alacsonyabb erőforrásigény és a statisztikai alapú parametrikus működés. A statisztikai beszédszintézisben a rendszer a tanulási fázis során kinyeri a tanító beszédadatbázisból a beszélő hangjára jellemző tulajdonságokat, és ezek alapján határozza meg később a szintézis során a beszéd generálásához szükséges paramétereket, majd egy beszédkódoló eljárás ez alapján létrehozza a szintetizált beszédet. Ezen paraméterek közé tartoznak például a beszéd alapfrekvenciája, hang- és szünetidőtartamai, illetve spektrális együtthatói.

A kutatás során a HTS [13] nyílt forráskódú HMM-alapú beszédszintetizátor magyar nyelvre adaptált változatát alkalmaztuk [12]. A kísérletekhez egy professzionális női bemondóval készült fonetikailag gazdag beszédadatbázist használtunk fel, amely 2 órányi 16 kHz-en mintavételezett, 16 bites kvantálású beszédet tartalmaz összesen 1940 kijelentő mondatban.

2 Módszerek

Amennyiben a HMM-alapú beszédszintézisben az eredeti tanító adatbázist több részre bontjuk, és ezekre külön-külön elvégezzük a statisztikai alapú tanítást, akkor ez alapján különböző paraméterértékeket tanul be a rendszer. A különböző részadatbázisok paramétereit egy beszédszintézisre épülő alkalmazásban párhuzamosan felhasználva (azaz felváltva használva az eltérő paraméterhalmazokat) elérhető, hogy egy adott mondathoz ne mindig ugyanaz a prozódia tartozzon. Ha a részadatbázisok mondatai elég különbözőek voltak, akkor a generált ismétlődő mondat tulajdonságai is eltérőek lesznek ismételt szintézis során, illetve azt várjuk, hogy hasonló szerkezetű mondatok is lényegesen eltérő prozódiával fognak rendelkezni. A HTS rendszerrel végzett betanítási és szintetizálási, valamint adatbázis feldarabolási lépéseket az 1. ábra mutatja be.



1. ábra: A beszédkorpusz feldarabolása, majd HMM tanítási fázis (felső rész). A bemeneti szöveghez HMM adatbázis kiválasztása, majd szintézis fázis (alsó rész).

2.1 Prozódiai távolságmértékek

Két mondat prozódiájának objektív összehasonlítására számos módszer található a szakirodalomban. Amennyiben csak a mondatok alapfrekvencia-menetét akarjuk összehasonlítani, többek között az átlagos négyzetes közép távolság (Root Mean Square Error, RMSE) [6], a Hermes-korreláció [5], vagy ez utóbbinak DTW-vel (Dynamic Time Warping) kiegészített változata [10] használható.

Az RMSE a következő módon számítható két mondat dallama között [6]:

$$RMSE_{f1,f2} = \sqrt{(\frac{1}{n}\sum_{i=1}^{n}(f_1(i) - f_2(i))^2)}$$

ahol f_1 és f_2 jelöli a két összehasonlítandó mondat F0 értékeit, n pedig a mérőpontok száma.

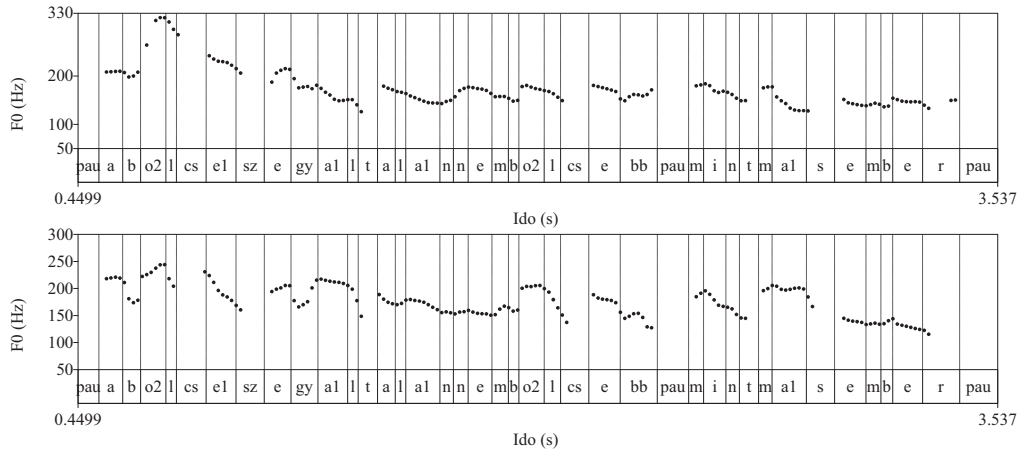
A Hermes-korreláció számítása [10] alapján:

$$Hermes_{f1,f2} = \frac{\sum_i w(i)(f_1(i)-m_1)(f_2(i)-m_2)}{\sqrt{\sum_i w(i)(f_1(i)-m_1)^2 \sum_i w(i)(f_2(i)-m_2)^2}}$$

ahol f_1 és f_2 jelöli a két összehasonlítandó mondat F0 értékeit, m_1 és m_2 ezeknek az átlagos F0-ja, ezen kívül a *w(i)* egy súlyozó faktor az adott jelszakasz intenzitásának függvényében. Az alapfrekvenciát sok esetben nem közvetlenül Hz-ben, hanem logaritmizálva alkalmazzák ezen képletekben [10].

A DTW alapú Hermes-korreláció akkor lehet hasznos, ha olyan mondatok alapfrekvenciájának összehasonlítására van szükség, amelyeknek időszerkezete jelentősen eltérő.

A 2. ábra egy példát mutat két mondat F0-menete közötti RMSE távolság és Hermes-korrelácó értékére. A továbbiakban a Hermes-korrelációt használtuk fel prozódiai távolságmértéknek, mert a szakirodalom alapján ez alkalmasabb az alapfrekvencia-különbségek kimutatására, mint az RMSE távolság [5].



2. ábra: Egy mondat két különböző F0-menettel rendelkező változatának összehasonlítása (amennyiben a mondatok időszerkezete megegyezik). A szótagonkénti átlagos F0 értékek alapján számolva az RMSE távolság 0,1619; a Hermes-korreláció pedig 0,6337.

2.2 Tanító adatbázis felbontása

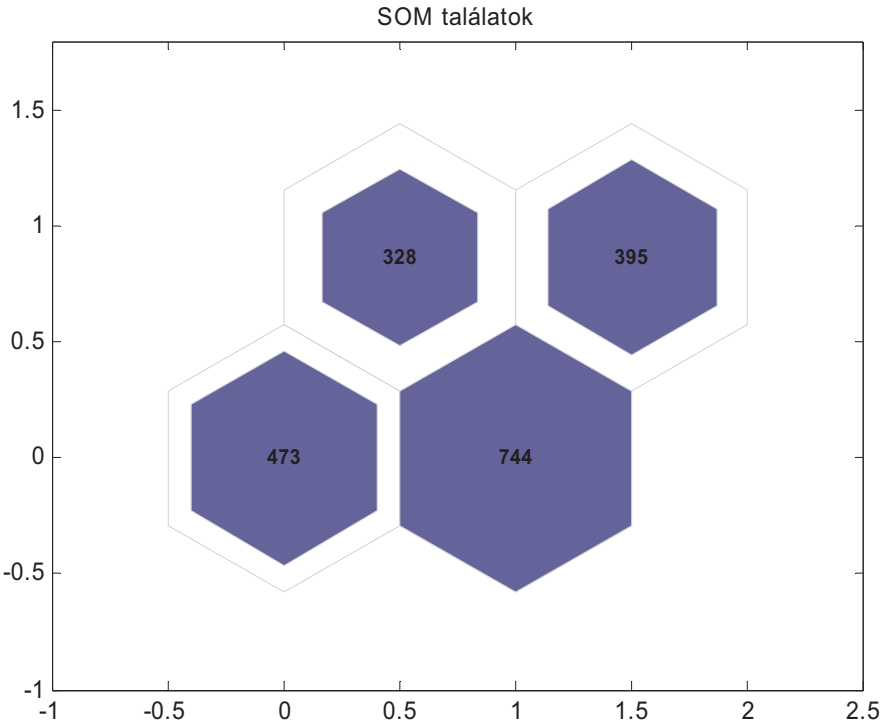
A kutatás során megvizsgáljuk, hogy egy adott beszélőtől származó különböző részt-anítóadatbázisokkal mennyire különböző prozódiájú mondatok állíthatóak elő a dallam, illetve alapfrekvencia tekintetében.

Az eredeti 1940 mondatból álló beszédkorpuszt több eltérő módon választottuk külön csoportokba. Első kísérletként véletlenszerűen szétválogattuk a mondatokat 2, 4, 8, illetve 16 csoportra, majd mindegyik rész-tanítóadatbázis segítségével elvégeztünk egy tanítást a HTS beszédszintetizátorral, majd leszintetizáltunk 40 mondatot. A szintetizálás során csak a betanult dallam modellt változtattuk (a gerjesztési, hangidőtartam és egyéb paramétereket változatlanul hagyva).

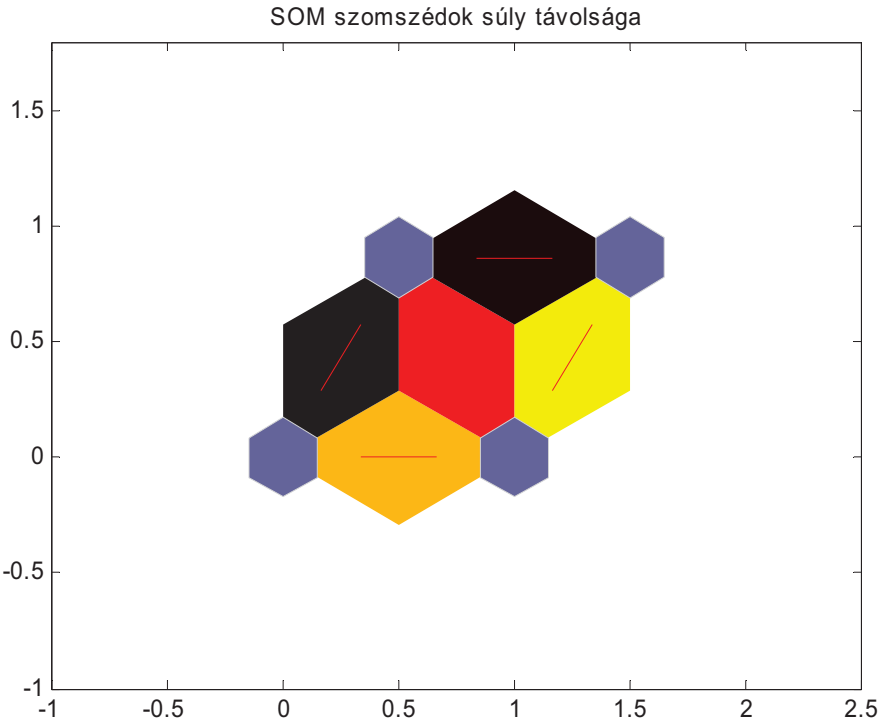
Ezután a 2.1 szakaszban ismertetett Hermes-korreláció objektív távolságmértéket felhasználva ellenőriztük, hogy egy adott szöveghez tartozó szintetizált változatok mennyire különböznek egymástól a mondat F0-menetének szempontjából. Ehhez a szótagonkénti átlagos F0 érték alapján számoltuk a Hermes-korrelációt. A véletlen szétválasztás esetén a mondatváltozatok közötti Hermes-korreláció magas volt (a legtöbb esetben 0,95 fölötti érték), azaz olyan mondatokat sikerült így szintetizálni, melyeknek F0-menetében nem fordult elő ezen mérték szerint jelentős különbség.

A véletlen választás mellett a továbbiakban azt vizsgáltuk, hogyan lehet gépi tanuló algoritmussal célzottan szétválasztani az eredeti beszédkorpuszt több klaszterre. Ehhez a választásunk a felügyelet nélküli tanításon alapuló Self-Organizing Feature Map (SOFM) eljárásra esett. A Kohonen által bemutatott megoldást [7] használtuk fel egy Matlab-alapú implementációban [1]. A SOFM-ot korábban sikeresen alkalmazták hangoskönyvek beszédanyagának expresszivitás szerinti szétválasztására [11]. A SOFM alkalmasnak látszik az alapfrekvencia szerinti szétválasztás feladatára, mivel felügyelet nélküli gépi tanulási módszer. A betanítás során azt kell beállítanunk, hogy hány részre bontsa szét a korpuszt az algoritmus. A SOFM bemeneteként felhasznált tulajdonságoknak az F0 bizonyos statisztikáit választottuk (minimum, maximum, átlag, szórás 1-1 mondaton belül), azaz mondatonként ezek a paraméterek álltak rendelkezésre a felügyelet nélküli tanításhoz.

A SOFM további előnye, hogy a többdimenziós adat kétdimenziós térképen ábrázolható. A 3. ábrán a klaszterezés eredményeként kapott 4 csoport látható, melynek során az 1940 mondat egy nagyobb és három kisebb részkorpuszra lett felbontva. A 4. ábra a szomszédos klaszterek közötti távolságok térképét mutatja. A hexagonok a bemeneti változókon (vagyis az F0 paraméterei) elvégzett felügyelet nélküli tanításból származó klaszterek. Azok a kapcsolatok, amelyek nagyobb távolságot mutatnak a klaszterek között, sötétebb színnel vannak jelölve. Az ábráról az látható, hogy a bal felső csoport távolsága nagy a többi csoporttól, míg a többi távolság ehhez képest alacsonyabb. Ez alapján azt várjuk, hogy azok a szintetizált mondatok, amelyek a bal felső mondatokkal mint tanító adatbázissal készülnek, dallam szempontjából nagyobb távolságra lesznek a többi tanító adatbázissal készült szintetizált mondatoktól, mint azok egymástól.



3. ábra: A SOFM alapú klaszterezés eredményeként felbontás után kapott négy tanítóadatbázis mondatainak elemszáma.



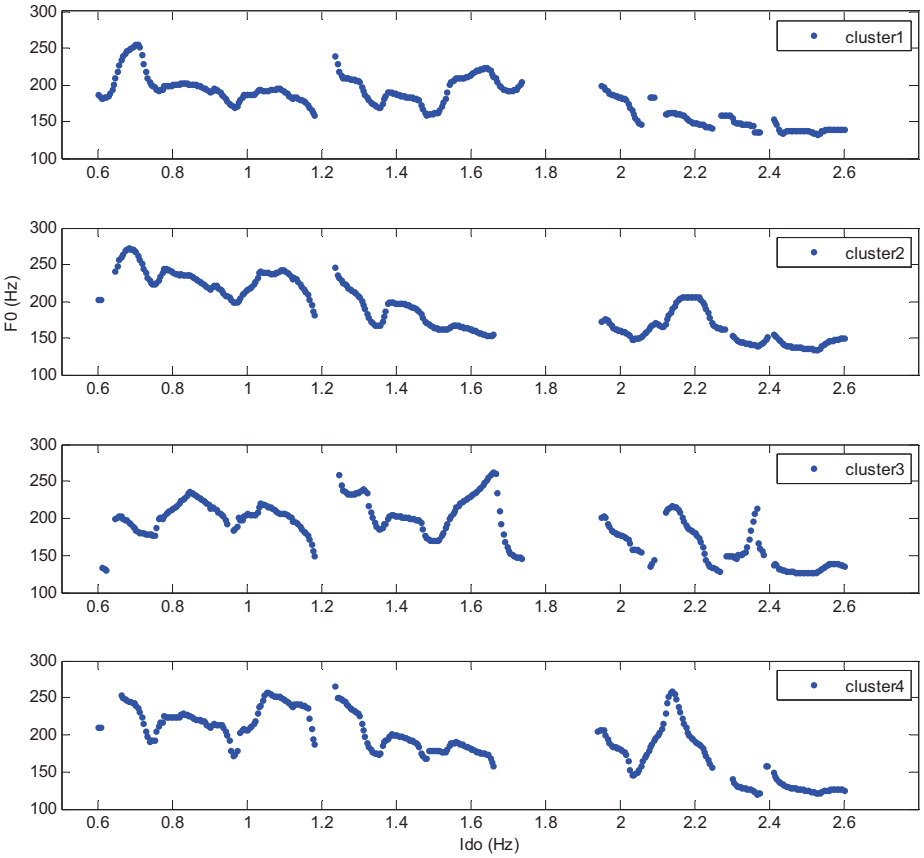
4. ábra: A SOFM alapú klaszterezés eredményeként felbontás után kapott négy tanító adatbázis egymástól mért távolsága. A világosabb szín kisebb, a sötétebb szín nagyobb távolságot jelöl.

3 Eredmények

A SOFM alapú klaszterezés eredményességét objektív és szubjektív vizsgálatokkal is ellenőriztük. 2000 kiválasztott mondatot leszintetizáltunk a 4 tanító adatbázisból származó F0-modellel külön-külön (a gerjesztési és időtartam paramétereket a teljes tanító adatbázisból származó modellből felhasználva).

3.1 Objektív különbségek

A mondatváltozatok közötti dallambeli különbség vizsgálatára a 2.1 szakaszban ismertetett Hermes-korrelációt használtuk fel. A szintetizált mondatok 4 változatát páronként összehasonlítottuk, majd kiszámoltuk az egyes mondatváltozatok közötti Hermes-korrelációt, melyre egy példát az 5. ábra és az 1. táblázat #1625 része mutat.



5. ábra: A #1625 mondat („Zsigmond nem tagadja, hogy ő zsidó.”) négy szintetizált változata, különböző tanító adatbázisokból kiindulva. Az alapfrekvencia-menet (és így a mondatdallam, illetve a hangsúlyok helye és erőssége) eltérő a különböző változatokban.

Ezután a 2000 mondatból kiválasztottunk 10 mondatot, melyeknél a változatok közötti F0 szerinti Hermes-korreláció a legalacsonyabb volt (így várhatóan ezek között észlelhető a legnagyobb különbség a mondatdallamban).

3.2 Szubjektív különbségek

A 10 legnagyobb objektív különbséggel rendelkező mondat 4-4 változatát választottuk ki a szubjektív teszt hanganyagához páros összehasonlítás keretében, így összesen 60 mondatpár állt rendelkezésre. A meghallgatásos teszt célja az volt, hogy ellenőrizzük, a Hermes-korreláció milyen mértékben mutatja meg a mondatdallambeli különbséget egy percepciós vizsgálathoz képest. Hasonló vizsgálatot végeztek korábban például német mondatokon [9].

A meghallgatásos tesztet internetes tesztfelületen végeztük. A mondatokat páronként kellett meghallgatniuk a tesztelőknek, és arra a kérdésre válaszolniuk, hogy „Hallasz-e különbséget a két mondat dallama között? Igen – Nem". Ezután ha „Igen"-nel válaszoltak, egy második kérdést is meg kellett válaszolniuk: „Ha hallottál különbséget, akkor milyen mértékű? Kicsi – Közepes – Nagy".

A mondatpárok meghallgatását 9 tesztelő végezte el. A tesztelők mindannyian ép hallású, magyar anyanyelvű emberek voltak, a 23-60 év közötti korosztályból (átlagosan 33 év). Egy részük a témához értő beszédtechnológiai szakértő vagy fonetikus volt, míg a többiek egyetemi hallgatók köréből kerültek ki. A teszt átlagos meghallgatási ideje 12 perc volt.

Az 1. táblázatban hasonlítjuk össze a mondatváltozatok között mért Hermes-korrelációt, és a tesztelők „Igen" válaszainak arányát. A szubjektív teszt 2. kérdését, (azaz a dallambeli különbség mértékét) itt nem vettük figyelembe, de az észrevehető volt a válaszok között, hogy a tesztelők leggyakrabban „kicsi" és „közepes" különbséget jelöltek csak be. A táblázatban a Hermes-korrelációnál az alacsonyabb érték jelent nagyobb F0 eltérést, míg az „Igen" aránynál a nagyobb szám jelenti azt, hogy többen észleltek különbséget a mondatváltozatok dallamában. Az eredmények alapján az objektív és a szubjektív mérték között nem található erős összefüggés ($R^2 = 0{,}115$).

A 60 mondatpárból összesen 35 esetben válaszolta a tesztelők legalább 65%-a, hogy hall különbséget a változatok között. A maradék 25 mondatpárt megvizsgálva az derült ki, hogy ezekben az esetekben a mondatváltozatok közötti szótagonkénti átlagos F0 különbsége legfeljebb 10-20 Hz volt. Azoknál a mondatpároknál, ahol hallottak különbséget a tesztelők, a legnagyobb F0 különbség akár a 70 Hz-et is elérte, és több helyen előfordult, hogy a mondat hangsúlya (az ereszkedő jellegű alapfrekvencia-menetből lényegesen kiugró rész) is másik szóra került. A #0074-es mondat (*„A bölcsész egyáltalán nem bölcsebb, mint más ember."*) esetén például a négy változatban különböző pozíciókra helyeződött a mondathangsúly: *„bölcsész"*; *„egyáltalán"*; *„bölcsebb"*; *„más"*. Ezek közül nem minden változat megfelelő, a *„más"* szóra helyezett hangsúly például helytelen hangsúlyozást jelent.

1. táblázat: A 10 kiválasztott mondat 4-4 változata közötti Hermes-korreláció és a szubjektív teszt alapján számolt különbség.

Mondat	v1	v2	Hermes-korreláció	Szubjektív „Igen”
#0044	1	2	0,7833	88,89%
#0044	1	3	0,7416	66,67%
#0044	1	4	0,8271	55,56%
#0044	2	3	0,9408	55,56%
#0044	2	4	0,9071	33,33%
#0044	3	4	0,9385	33,33%
#0046	1	2	0,7697	44,44%
#0046	1	3	0,7410	44,44%
#0046	1	4	0,7185	77,78%
#0046	2	3	0,9356	22,22%
#0046	2	4	0,9158	66,67%
#0046	3	4	0,9644	88,89%
#0069	1	2	0,7663	77,78%
#0069	1	3	0,8016	66,67%
#0069	1	4	0,8260	77,78%
#0069	2	3	0,9273	22,22%
#0069	2	4	0,8608	55,56%
#0069	3	4	0,9381	77,78%
#0074	1	2	0,6337	88,89%
#0074	1	3	0,8452	77,78%
#0074	1	4	0,8101	77,78%
#0074	2	3	0,7819	44,44%
#0074	2	4	0,7759	66,67%
#0074	3	4	0,8971	77,78%
#0091	1	2	0,9034	66,67%
#0091	1	3	0,6437	66,67%
#0091	1	4	0,9006	66,67%
#0091	2	3	0,8481	44,44%
#0091	2	4	0,9777	0,00%
#0091	3	4	0,8189	55,56%

Mondat	v1	v2	Hermes-korreláció	Szubjektív „Igen”
#0186	1	2	0,8515	44,44%
#0186	1	3	0,7416	77,78%
#0186	1	4	0,7650	66,67%
#0186	2	3	0,8877	66,67%
#0186	2	4	0,9575	33,33%
#0186	3	4	0,9108	66,67%
#0849	1	2	0,6929	77,78%
#0849	1	3	0,7921	44,44%
#0849	1	4	0,8694	55,56%
#0849	2	3	0,9327	55,56%
#0849	2	4	0,8991	22,22%
#0849	3	4	0,9406	66,67%
#1342	1	2	0,9205	55,56%
#1342	1	3	0,7346	77,78%
#1342	1	4	0,9032	55,56%
#1342	2	3	0,8172	55,56%
#1342	2	4	0,9127	77,78%
#1342	3	4	0,7591	66,67%
#1425	1	2	0,8240	66,67%
#1425	1	3	0,8310	66,67%
#1425	1	4	0,7815	77,78%
#1425	2	3	0,9546	11,11%
#1425	2	4	0,8546	88,89%
#1425	3	4	0,9040	66,67%
#1625	1	2	0,7812	44,44%
#1625	1	3	0,8299	44,44%
#1625	1	4	0,8523	77,78%
#1625	2	3	0,6547	77,78%
#1625	2	4	0,9233	66,67%
#1625	3	4	0,8081	66,67%

A kísérletet végighallgatóknak a teszt végén megjegyzések hozzáfűzésére is volt lehetőségük. Az egyik tesztelő a mondatdallambeli különbséget jóval nagyobbnak érezte azokban az esetekben, amikor a hangsúly is másik szóra került (esetleg olyan szóra, amit valójában nem is kellett volna hangsúlyozni), mint amikor a hangsúly pozíciója azonos volt a két változatban, de az alapfrekvenciában mégis jelentős különbség volt.

4 Összefoglalás

A kutatás során bemutattunk egy egyszerű módszert, amivel egy adott szöveghez különböző dallammal rendelkező mondatokat lehet szintetizálni. Ehhez egy statisztikai F0-modellt használtunk fel HMM-alapú beszédszintetizátorban. Az eredeti beszédkorpuszt az SOFM módszerrel bontottuk fel négy részre. A különböző beszédkorpuszokból betanult modellekkel eltérő dallamú mondatváltozatokat szintetizáltunk (azonos szöveghez). Ezután megvizsgáltuk a mondatváltozatok közötti különbségeket. A szubjektív kísérletek azt mutatják, hogy az alapfrekvencia eltérése a vizsgált mondatpárok felében annyira jelentős volt, hogy ez az emberi fül számára is észlelhető (azonban ez nem áll szoros összefüggésben az objektív távolságmértékkel). Ahhoz, hogy percepciós szempontból eltérő prozódiájú mondatokat tudjunk létrehozni, az szükséges, hogy az eredeti beszédkorpusz felbontása minél jobban eltérő részekre történjen, melyre a SOFM módszer alkalmasnak látszik.

A változatosabb prozódiával kiegészített beszédszintézis azokban a rendszerekben jelenthet javulást a felhasználók számára, ahol hosszabb szövegek felolvasása történik, illetve gyakran előfordulnak ismétlődő, hasonló szerkezetű mondatok. Ezek közé tartozik a könyv és az e-levél felolvasás.

A kutatást részben a TÁMOP-4.2.1/B-09/1/KMR-2010-0002 projekt támogatta.

Bibliográfia

1. Bealen, M.H., Hagan, M.T., Demuth, H.B.: Neural Network Toolbox, Revised for Version 7.0, Release 2010b, http://www.mathworks.com/help/toolbox/nnet/ (2010)
2. Csapó, T.G., Zainkó, Cs., Németh, G.: A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System. Infocommunications Journal, Vol. LXV, No.1 (2010) 32–37
3. Campillo Díaz, F., Rodríguez Banga, E.: A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. Speech Communication Vol. 48 (2006) 941–956
4. Campillo Díaz, F., van Santen, J., Rodríguez Banga, E.: Integrating phrasing and intonation modelling using syntactic and morphosyntactic information. Speech Communication, Vol. 51, No.5 (2009) 452-465
5. Hermes, D.J.: Measuring the perceptual similarity of pitch contours. Journal of Speech Language Hearing Research Vol. 41 (1998) 73–82
6. Klabbers, E., van Santen, J., Wouters, J.: Prosodic factors for predicting local pitch shape. In Proceedings 2002 IEEE Workshop on Speech Synthesis. Santa Monica, CA (2002)

7. Kohonen, T., Kaski, S., Lappalainen, H.: Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. Neural Computation Vol. 9, No. 6 (1997) 1321–1344
8. Németh, G., Fék, M., Csapó, T.G.: Increasing Prosodic Variability of Text-To-Speech Synthesizers. In: Proc. of Interspeech (2007) 474–477
9. Reichel, U.D., Kleber, F., Winkelmann, R.: Modelling similarity perception of intonation. In: Proc. of Interspeech (2009) 1711–1714
10. Rilliard, A., Allauzen, A., Boula de Mareüil, P.: Using Dynamic Time Warping to compute prosodic similarity measures. In: Proc. of Interspeech (2011) 2021–2024
11. Székely, E., Cabral, J. P., Cahill, P., Carson-Berndsen, J.: Clustering expressive speech styles in audiobooks using glottal source parameters. In: Proc. of Interspeech, (2011) 2409–2412
12. Tóth B.P., Németh G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 246–256
13. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system version 2.0. In: Proc. of ISCA SSW6 (2007)