# Magyar nyelvű, kötött témájú korpusz-alapú beszédszintézis és a kötetlenség felé vezető út vizsgálata

ZAINKÓ CSABA

*BME Távközlési és Médiainformatikai Tanszék*
*zainko@tmit.bme.hu*

**Lektorált**

***Kulcsszavak: korpusz-alapú beszédszintézis, beszédadatbázisok, prozódiai modul***

*A beszédszintetizátorok között a korpusz-alapú szintetizátorral lehet jelenleg a legjobb minőségű beszédet előállítani. Ennek ára, hogy csak adott témájú szövegek szintetizálását tudja ilyen minőségben garantálni. A cikk azt tárgyalja, hogy ha egy ilyen kötött témájú korpuszos szintetizátort kívánunk kötetlen szövegekre kibővíteni, akkor annak milyen lehetőségei és korlátai vannak. A vizsgálat során a szintetizátor beszédadatbázisát elemeztük és megvizsgáltuk, hogy elegendően változatos-e tetszőleges szöveghez, illetve megfelelő számú elemet tartalmaz-e a jó minőséghez. Végül a szintetizált mondatokat egy meghallgatásos teszt keretében értékeltettük tesztelőkkel.*

## 1. Bevezetés

A korpusz-szintetizátorokat általában meghatározott témájú szövegek szintetizálására fejlesztik (például időjárásjelentés, menetrendi tájékoztató, árlista felolvasó) [1]. A szintetizátor egy válogató algoritmusból és a hozzá tartozó beszédadatbázisból áll. Egy új témakörre való fejlesztés során általában csak a beszédadatbázist kell elkészíteni, mivel a szintetizátor válogató algoritmusa már megfelelően tesztelt, jól válogat. A munka nagy részét ebben az esetben az adatbázis elkészítése jelenti, azaz a megadott témájú szövegek felolvasása és előkészítése a szintézishez (tisztítás, címkézés, zöngés-zöngétlen határok bejelölése stb.). Ezek után a szintetizátor az adott témában tetszőleges mondatokat képes beszéddé alakítani, amely a technológiából adódóan közel emberi minőségű.

A kísérletben megvizsgáljuk, hogy a BME-TMIT-en készített, kötött tematikára készült beszédadatbázisok összeépítésével (egyazon bemondó hangjára) milyen minőségben lehet tetszőleges tartalmú mondatokat szintetizálni. A kutatás irányt ad arra is, hogy a korpuszos technológiánál milyen problémákkal kell számolni, ha kötetlen, általános beszédszintézist kívánunk megcélozni.

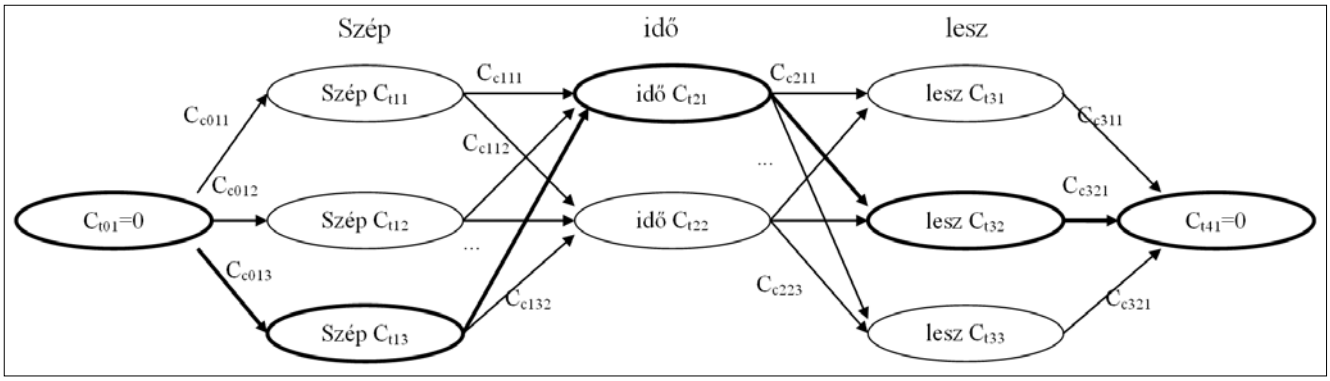
Az első részben bemutatjuk a szintetizátor működését, majd a beszédadatbázist vizsgáljuk meg, hogy milyen a mennyiségi és a minőségi összetétele. Alapvetően ez határozza meg, hogy milyen mondatok szintetizálására alkalmas a rendszer. A beszédadatbázis részletes elemzése után bemutatjuk, hogy milyen kísérleti mondatokat állítottunk elő és azokat a tesztelők hogyan értékelték. Az utolsó részben végül megvizsgáljuk a hangminőség javításának lehetőségeit.

## 2. Kötött témakörre fejlesztett korpusz-alapú beszédszintetizátor működése

A korpusz-alapú, elemkiválasztásos szintetizátor – továbbiakban korpuszos szintetizátor – egy olyan beszédgenerátor, amely nagy mennyiségű előre rögzített beszédből (beszédkorpuszból) válogatja ki a megfelelő elemeket és állítja elő ezek felhasználásával a szintetizált beszédet. A működés menetét az *1. ábrán* látjuk.

A példamondat a következő: *„Szép idő lesz”*. A szintetizátor a beszédkorpuszból válogat, meghatározza, hogy melyek azok a beszédrészletek (főleg szavak), amelyek felhasználhatók a mondat előállításához. Ezeket az ábrán a szavak alatt található ellipszisek jelölik,

*1. ábra Példa az elemek kiválasztására és költségeire*



amelyek most szó-méretű elemek, de lehetnek kisebbek is. A talált jelölteket egy mérőszámmal (célegyezési – target – költséggel, az ábrán $C_{txy}$) látja el, amely meghatározza, hogy mennyire alkalmas az adott elem a keresett pozícióra. A költség egyfajta büntetés, minél nagyobb, annál kevésbé alkalmas az adott helyre. Az egymás melletti pozícióra kiválasztott jelöltek között is kiszámol egy költséget a rendszer (összefűzési – concatenation – költséget, az ábrán $C_{cxyz}$), amely megadja, hogy mennyire illeszkedik a két elem egymáshoz. Itt is annál nagyobb a költség, minél rosszabbul illeszkedik a két elem. A végső elemsor kiválasztásához az összköltség minimalizásának segítségével jutunk el, amely a felhasznált elemek célegyezési és összefűzési költségeinek összegéből áll. A mondat a legkisebb összköltségű elemsorból fog előállni. Ezt a válogatást a Viterbi algoritmus határozza meg [2]. A számításhoz egy kezdő- és végelemet is felhasználunk, amely egy szünet- vagy csendjellegű elem, ezeket az ábrán $C_{t01}$ és $C_{t41}$-el jelöltük.

Az ábrán a példamondat előállításához kiválasztott elemeket megvastagítva láthatjuk. Ennek az elemsornak a költsége: $C_{t01}$+ $C_{c013}$+ $C_{t13}$+ $C_{c131}$+ $C_{t21}$+ $C_{c212}$+ $C_{t32}$+ $C_{c321}$+ $C_{t41}$.

Abban az esetben ha nem található meg a keresett szó, akkor a szóhoz tartozó beszédhangokat keresi a rendszer. Ha a példában szereplő *„idő”* szó nem szerepelne az adatbázisban, akkor az *„i” „d” „ő”* hangokat keresi a rendszer a megfelelő környezetben és a szavakhoz hasonlóan számolja a cél- és az összefűzési költségeket.

Mint korábban említettük, az előállított beszéd minősége nagy részben függ attól, hogy a szintetizálni kívánt mondat mennyire illeszkedik a beszédkorpusz témájához. Ha hasonló szavakból álló mondatot szeretnénk előállítani, mint amilyenek a korpuszban szerepelnek, akkor hosszabb beszédelemekből (szavak, szófüzérek) tudja a szintetizátor előállítani a mondatot, a jelöltek is többen lesznek egy-egy pozícióra, így nagyobb eséllyel tud jobban illeszkedőt találni. Az elemek összeillesztésének száma is kevesebb lesz, így az esetleges illesztetlenségi hibák is kisebb számban és mértékben jelennek meg az előállított beszédben.

Összefoglalva azt mondhatjuk, hogy ebben az esetben kevésbé sértjük meg azt a *tételt*, ami azt mondja, hogy *az optimális beszédjel egyedi és egyszeri produktum*. (Ezt például a dadogó megsérti, mivel szaggatottá teszi a jelet, ezért beszéde távol lesz a köznapi normától). A tétel vonatkoztatása az adatbázisra azt jelenti, hogy minél hosszabb beszédegységeket sikerül kiválasztani, annál optimálisabb lesz a hangzás. A legoptimálisabb az a helyzet, amikor a teljes keresett mondat benne van az adatbázisban. Ilyenkor az előbbi tétel teljes mértékben teljesül [3]. Ha eltérő tematikájú mondatot szintetizálunk, akkor kisebb elemeket kell használni, azok az adatbázis különböző helyeiről származhatnak, az ottani elemeket egymástól eltérő időpontokban ejtette a bemondó, tehát az előbbi tétel sérül. Ennek eredménye a több illesztési pont szükségszerű megjelenése is, amely a percepció számára is jól hallható hangzásingadozást okozhat.

A szintetizált beszéd előállításakor a hangsorozat kialakítása mellett a prozódiát is meg kell valósítani. A prozódia alatt a hangsúlyok helyét, a dallammenetet, a szüneteket és a tempóváltást értjük, amely fizikailag az egyes hangok hangmagasságában, energiájában és időtartamában jelenik meg.

A korlátozott tematikára fejlesztett szintetizátor nem tartalmaz külön prozódia generáló és megvalósító egységet, hanem az az elemkiválasztó algoritmusba van beépítve [2]. Mivel az adatbázis elemei természetes emberi bemondásokból származnak, tartalmazzák annak a mondatnak a prozódiáját is, amelyben szerepelnek. A prozódiai információk figyelembevétele a célegyezési költségben ($C_t$) történik. A költségben büntetve van, ha a mondat más részéből venné az elemet a válogató algoritmus. A példamondatunkban szereplő *„lesz”* szó dallammenete csak akkor megfelelő, ha szintén a mondat végéről származik. Ha mondat közepéről vagy elejéről származó *„lesz”* szót használna a szintetizátor ebben a pozícióban, akkor természetellenes hangzást kapnánk.

## 3. Mi kell egy általános korpusz-alapú szintetizátorhoz?

Az általános tematikájú szintetizáláshoz két ponton kell vizsgálnunk a korpuszos, kötött témakörű szintetizátor adatbázisát. Az egyik az, hogy a szükséges hangsorépítő elemek rendelkezésre állnak-e a beszédadatbázisban a tetszőleges mondatok előállításához. A második pedig az, hogy a korlátozott tematikájú szintetizátor algoritmusai mennyire alkalmasak arra, hogy tetszőleges mondatot állítsanak elő prozódiai szempontból.

### 3.1. Beszédadatbázis

A vizsgálathoz három különböző tematikájú, ugyanazon bemondótól rögzített beszédkorpuszt egyesítettünk. Az első rész időjárásjelentés-típusú mondatokból

| | Időtartama | Mondatok száma | Szavak száma | Hangok száma |
|---|---|---|---|---|
| **Időjárás** | 10,7 óra | 5821 | 102940 | 488093 |
| **Menetrend** | 1,1 óra | 515 | 8656 | 39027 |
| **Szám** | 14 perc | 205 | 1006 | 7042 |
| ***Összesen*** | **12 óra** | **6541** | **112602** | **534162** |

*1. táblázat*
*A vizsgálathoz felhasznált adatbázisok méretei*

állt, amely különböző időjárási internetes oldalak tartalma alapján készült. A második rész egy állomás menetrendi információit felolvasó rendszer adatbázisa, amely a járatok érkezésével és indulásával kapcsolatos bemondásokat tartalmazza. A harmadik rész egy olyan adatbázis, amely 1200 többjegyű szám felolvasását tartalmazza [4].

Az adatbázisok néhány jellemző adatát az *1. táblázat* mutatja. Mindhárom adatbázis felvételei azonos körülmények között, azonos stúdióban, azonos mikrofonnal készültek.

Látható, hogy az első – időjárásos – adatbázis a legnagyobb. Az adatbázisból a szintetizátor az aktuális prognózisokat olyan minőségben tudja felolvasni, hogy a hangzás minősége az emberi bemondásokkal közel azonos [2]. Az adatbázis tematikája a napi prognózisoknál bővebb, orvos- és közlekedésmeteorológiai témájú mondatokat is tartalmaz. A második adatbázis kisebb és a mondatok változatossága sem túl nagy, sok azonos szerkezetű és jellegű mondat is található benne. A harmadik adatbázis csak számokat tartalmaz, a három közül ez a legszűkebb tematikájú. Ez az adatbázis a többihez képest kis mérete ellenére alkalmas a számok 1 milliárdig történő emberi minőségű szintetizálására. Ez azért lehetséges, mert a felolvasott többjegyű számok a fonetikai kapcsolódások figyelembevételével, alapos tervezés után lettek meghatározva [4].

*3.1.1. Szó-méretű elemek*

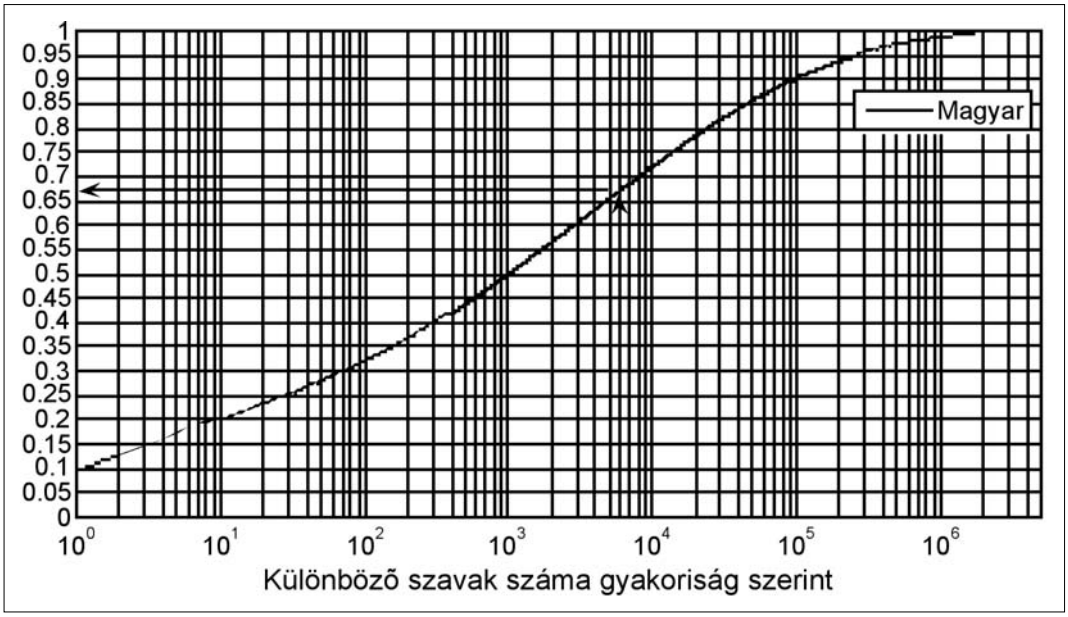
A korpuszos szintetizátor általában akkor adja a legjobb hangminőséget, amikor a leghosszabb, egybefüggő beszédrészleteket tudja felhasználni az adatbázisból. Ebben a vizsgálatban a szó az alapelem, amelyből az összesített adatbázis 112602 db-ot tartalmaz. A különböző szóalakok száma 6281. A nyelvben előforduló gyakoriságukat figyelembe véve meghatározhatjuk, hogy ezek a szavak a szintetizálandó mondatok szavainak hány százalékát teszik ki. A szavak statisztikai elemzéséhez egy saját gyűjtésű, korábbi szövegadatbázis adatait használtuk fel [5]: Digitális Irodalmi Akadémia, internetes újságok cikkei és a Magyar Nemzeti Szövegtár (1999), összesen 80 millió szó. A *2. ábrán* látható, hogy a nyelv leggyakoribb szavai a teljes nyelv szavainak hány százalékát fedik le. Amennyiben tehát a leggyakoribb 6000 szó állna rendelkezésünkre, akkor csak a 67%-ot tudnánk lefedni (nyíllal jelezve az ábrán).

Méréseink szerint a rendelkezésünkre álló – nem a leggyakoribb – 6281 szóval a mért szövegadatbázis 45%-a fedhető le, ami a szintetizálás szempontjából azt jelenti, hogy hozzávetőlegesen minden második szó esetén tudunk szó-méretű elemet felhasználni, a közbenső szavakat kisebb egységekből kell előállítani. Ez összehasonlítva a korlátozott tematikájú rendszerekkel, lényegesen rosszabb minőséget prognosztizál. A szűk tematika esetén átlagosan csak minden 15. szót állítunk elő kisebb elemekből, ami biztosítja a jó minőséget.

*3.1.2. Szónál kisebb méretű elemek*

A 45%-os szófedési adatból következik, hogy a kisebb elemekre gyakran lenne szükség a szintézis során. A kisebb elemek közül az egyedi hangok, a hangkapcsolatok, és a hanghármasok előfordulását vizsgáltuk meg. A magyar nyelvű szintézishez minimálisan 33 különböző hang szükséges a szünetet – mint a hang induló és befejező szakaszát – is beleértve. Vizsgálatainknál a „dz", „dzs" hangokat ritka előfordulásuk miatt, valamint a rövid-hosszú oppozíciót nem vettük figyelembe. Az így mért adatbázisban 534162 hang szerepel. Egyedi hangok összefűzéséből azonban nem lehet jó minőségű beszédet előállítani, figyelembe kell venni a hangkörnyezetet is.

Az egymásra hatások miatt az adatbázis fonetikai gazdagságáról jobb képet ad, ha a hangkapcsolatokat vizsgáljuk meg. Ilyen hangkapcsolatoknak nevezzük a kettős hangkapcsolatokat (diádok), amelyek más szintézistechnikákban rendszeresen használt elemek. Egy diád egy hangkapcsolatban szereplő két egymás melletti félhangból áll. Az összesített adatbázisban több,




*2. ábra Leggyakoribb szavak fedése*

mint félmillió diád szerepel. A matematikailag lehetséges 1089 (33*33) darab különböző diádból csupán 855 dbot találtunk meg az összesített adatbázisban. Ha csak azokat az diádokat számoljuk, amelyek legalább 15-ször előfordultak, akkor csak 703 különböző diád áll a rendelkezésünkre. A nagyon ritkán előforduló diádokkal az lehet a probléma a szintetizáláskor, hogy a kevés jelölt miatt, nagyon korlátozott azoknak az utaknak a száma, amelyből a szintetizátor kiválaszthatja a legjobbat, így a minőség várhatólag rosszabb lesz.

Az 1089 különböző diád élő nyelvben nem létezik, mert a nyelvtani és fonológiai szabályok miatt bizonyos kapcsolatok nem valósulhatnak meg. Például kizárólag a mássalhangzókat vizsgálva, a gyakorlatban csak 423 ilyen kettős kapcsolat van jelen a beszédben [6], amennyiben az abszolút hangsorkezdő-záró állapotot is ide számoljuk. Annak megállapítására, hogy melyek azok a diádok, amelyek tetszőleges szöveg szintetizálásakor szükségesek lehetnek, a szavaknál mutatott statisztikai módszerhez hasonlót használtunk. A szószablya [7,8] magyar webkorpusz (mint független adatbázis) mondatait a szintetizátor betű-hang átalakító rendszerével átírattuk fonetikus formába, majd előállítottuk ezekből ugyanazokat az adatokat, amelyeket az összesített adatbázisból is.

A szószablya korpusz adatai a *2. táblázatban* láthatók. Egy mondat átlagosan 83 diádból épül fel.

A különböző diádok száma itt már nagyobb, mint az összesített adatbázisban. A gyakorisági adatok szerint azok a diádok amelyek a szószablya webkorpuszban szerepelnek, de az összesített adatbázisban nem, az összes diád 1%-át teszik ki, ami azt jelenti, hogy átlagosan minden századik felhasználandó diád hiányozni fog. Ha csak diádokból építenénk fel a mondatot, akkor átlagosan 1,2 mondatonként lenne hiányzó diádunk, ami – ha csak ebből a szempontból vizsgáljuk – jó minőséget eredményezhetne.

A hanghármasok vizsgálatára azért van szükség, mert a korpuszos szintetizátor hang-alapú működése során akkor lehet a legjobb a kiválasztott hang minősége, ha a szintetizálandó mondat minden hangját (a környezetével együtt) megtaláljuk a beszédadatbázisban is. Ezt úgy biztosíthatjuk a keresésnél, hogy egy hang bal és jobb oldali szomszédját is figyelembe vesszük a célegyezési költség számításakor. Akkor optimális a helyzet, ha a szomszédos hangok ugyanazok, mint a szintetizált mondatban. Az adatbázis vizsgálatakor tehát most azt nézzük, hogy az ott megtalálható hanghármasok mennyire fedik le a magyar nyelvben használtakat. Az összesített adatbázisban 8727 db különböző hanghármas található, amiből 5748 db fordult elő legalább ötször.

A hanghármasok statisztikai vizsgálatához a – diádoknál is használt – szószablya webkorpuszt használtuk. Az elkészített fonetikus átiratban megvizsgáltuk, hogy milyen hanghármasok fordultak elő a webkorpuszban. Összesen 27982 különböző hanghármast találtunk, melyek közül 16643 fordult elő gyakran (legalább ezerszer).

Abban az esetben, ha az összesített adatbázisban előforduló összes hanghármas fedését vizsgáljuk, akkor az ott találtak a webkorpusz 96%-át fedik le. Ha a 15 vagy többször előforduló hanghármasokat vesszük csak figyelembe, akkor a fedés csak 82%-os. Ezt az adatot annak függvényében kell vizsgálni, hogy jó minőségű beszédet abban az esetben is elő lehet állítani, ha az adott hanghármas nincs meg pontosan, csak a hang artikulációs pozíciója egyezik. Az azonos képzési helyű mássalhangzók (consonant-C) hatása a hozzájuk kapcsolódó magánhangzókra (vowel-V) hasonló [9]. Tehát ha egy VCV kapcsolatban a C-re csak azonos képzési helyű C1 helyettesítőt találunk, akkor a C1-hez kapcsolódó magánhangzó ugyanolyan akusztikai szerkezettel fog rendelkezni, mint a VCV kapcsolatban, a helyettesítés tehát nem rontja az akusztikai eredményt.

Az összesített adatbázisról általánosságban elmondhatjuk tehát, hogy hang-szinten alkalmas tetszőleges beszéd előállítására, hosszabb elemek szintjén azonban túl hiányos.

### 3.2. Prozódia

Az emberi minőséghez közelítő szintetizált beszéd előállításához nem elég az, ha az adatbázisban megtalálhatók az előállítandó hangsornak megfelelő hangsorépítő elemek, hanem szükség van arra is, hogy a szintetizált mondat megfelelő prozódiával is rendelkezzen. Ha a prozódia nem megfelelő, a hallgató nem fogadja el természetes hangzású beszédnek a mondatot. A prozódia helyes előállítása legalább olyan nehéz feladat, mint a hangsorépítő elemek biztosítása.

A vizsgált, korlátozott tematikájú szintetizátorok adatbázisa csak kijelentő mondatokat tartalmaz. Mivel a kérdő mondat prozódiája jelentősen eltér ezektől, ezért a kérdő mondatokat az aktuális algoritmusok ezzel az adatbázissal nem képesek előállítani. A kérdő mondatok előállításához vagy olyan adatbázis kell, amely nagy számban tartalmaz kérdő mondatokat is, vagy olyan prozódia kiválasztó és megvalósító algoritmusok szükségesek, amelyek ezeket meg tudják valósítani. A továbbiakban már csak azt vizsgáljuk, hogy kijelentő mondatok esetében milyen esély van a helyes prozódia megvalósítására.

A vizsgált kötött témájú korpusz-alapú szintetizátorban a prozódia modellezése úgy történik, hogy figyeljük a szavak mondatbeli pozícióját [3]. A mondatokat első lé-

*2. táblázat A szószablya korpusz főbb adatai*

| | weblapok száma | mondatok száma | szavak száma | hangok száma |
|---|---|---|---|---|
| **Szószablya webkorpusz** | 1,2 millió | 42 milló | 589 milló | 3,5 milliárd |

pésben tagmondatokra bontjuk, majd ezen belül is meghatározzuk a szó helyzetét. A talált szóalakok vizsgálata során láthattuk, hogy azok átlagosan 45%-ban fedik le a magyar nyelvet, tehát a helyes prozódia is ilyen arányban állhat elő a szavakból a legjobb esetben. Az adatbázisban kis számban előforduló szavak esetén az is előfordulhat, hogy a szó ugyan egészben szerepel az adatbázisban, de nem a megfelelő mondatbeli pozícióban, ezért nem a megfelelő prozódiai információt hordozza.

Abban az esetben, ha kisebb elemekből, építi fel a mondatot a szintetizátor, akkor már nem veszi figyelembe ezeket a mondatbeli pozíció információkat. Előfordulhat tehát, hogy egy hangsúlyos szót olyan szavak elemeiből állít elő, amelyek hangsúlytalanok, ezért a kimenet is hangsúlytalan lesz.

A prozódia megvalósításáról tehát összegezve azt mondhatjuk el, hogy csak akkor várható el viszonylag elfogadható hangzás, ha a szintézis szó szinten tudja biztosítani a hangsorépítő elemeket és ezekből is elég számú van a beszédadatbázisban, amelyek a prozódiai változatosságot biztosítják.

## 4. Meghallgatásos tesztek

A beszédszintézis rendszerek minőségét meghallgatásos tesztek során végzett szubjektív minősítéssel lehet összehasonlítani. Ennek egyik módja a MOS (Mean Opinion Score – átlagos szubjektív osztályzat) teszt alkalmazása. A tesztekhez mondatokat válogattunk két témakörből. Az elsőben hírolvasásból, a másodikban egy meséből származtak a mondatok. Az előállított tesztanyag 5-5 szintetizált mondatot tartalmazott, amelyek eltérő hosszúságúak voltak. A mondatokat meghallgató és értékelő személyek számára az volt az utasítás, hogy egy 5-ös skálán értékeljék a minőséget (5-ös a legjobb érték). A tesztben továbbá szerepeltek a korpuszos szintetizátor eredeti mondatai is, amelyek a tematikának megfelelő időjárás jelentések voltak. A teszt internetes elérhetőségű volt, a tesztelők a mondatokat véletlen sorrendben hallgatták meg. A teszt tartalmazott egy bevezető részt is, amely azt a célt szolgálta, hogy a tényleges értékelés előtt már képet kapjanak arról, hogy milyen minőségű mondatokat fognak hallani. A teszt során a tesztelők nyilatkoztak arról is, hogy milyen eszközön, milyen környezetben hallgatják a mondatokat.

A tesztet 10 személy értékelte ki; 3 nő és 7 férfi. Az átlagéletkor 32 év volt. A tesztelők mindegyike csendes környezetben hallgatta meg a mondatokat, a legtöbben átlagos minőségű eszközökön. A tesztelők fele-fele arányban használtak hangszórót és fejhallgatót.

A *3. ábrán* az első oszlop mutatja a korpuszos szintetizátorral előállított, a témakörbe vágó mondatok értékelését. A második, vonalazott oszlop a hír és mese témakörökből válogatott mondatok átlaga. Az utolsó két oszlopon a két témakör külön-külön számított átlaga látható. A tematikán kívüli mondatok érthetősége roszszabb és kevésbé természetesek, mint az adatbázisnak megfelelő tematikájú korábban szintetizált mondatok. A különbség a két átlag között több mint 2, ami azt jelenti, hogy a minőségromlás jelentős. Az eredményekből az is megállapítható, hogy az eredeti tematikához közelebb álló hírjellegű mondatok jobbak, mint a tematikától messze álló meserészlet, bár ezek eltérése kicsi, ha a témakörbe vágó mondatokhoz viszonyítjuk.

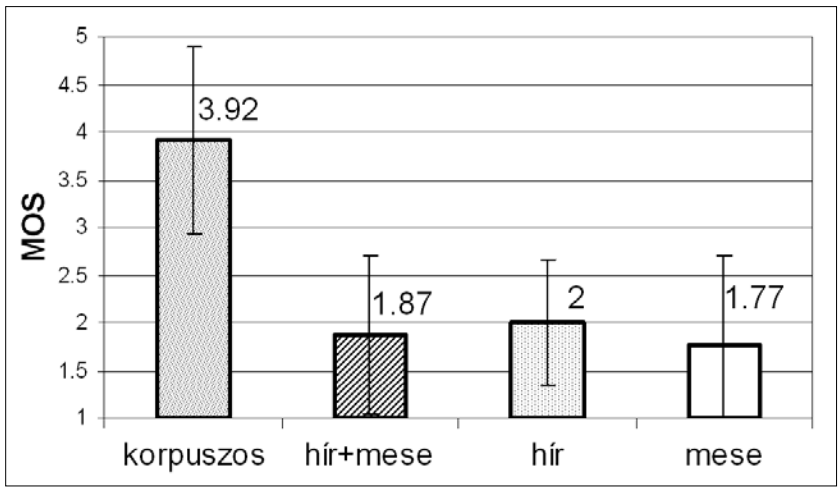
A meghallgatás utáni szabad véleményalkotás során kiderült, hogy a tesztelők szerint a mondatok egyes részei mind prozódiában, mind akusztikai szerkezetben lényegesen különböztek egymástól. Voltak részek, amelyek sokkal jobb osztályzatot kaptak volna, de a mondat többi része lehúzta az értékelést. A legtöbbet említett jelenség az egyenetlen minőség volt.

## 5. Fejlesztési lehetőségek

Az adatbázisok elemzéséből látható, hogy méretük növelése egyértelműen javíthatja a generálandó szintetizált beszéd minőségét. Ezt a kötött témakörű rendszerek fejlesztése során már többször alkalmaztuk. Ha újabb mondatok szintetizálásának igénye jelent meg és a szintézis hangminősége nem volt megfelelő, akkor egy jól megtervezett hangfelvétellel az adatbázist úgy bővítettük, hogy ezután ezeket az újabb mondatokat is jó minőségben tudta előállítani a rendszer. Amennyiben viszont azt szeretnénk, hogy tetszőleges tematikájú mondatot is szintetizálni tudjunk megfelelő minőségben, akkor az adatbázist olyan mértékben kellene bővíteni ezzel a módszerrel, amely nehezen vagy gyakorlati szempontból egyáltalán nem megoldható.

A jelenleg használt adatbázis 6281 különböző szót tartalmaz. Ha azt szeretnénk elérni, hogy az adatbázisban a magyar szavak 95%-a szerepeljen, akkor a 2. ábrából leolvashatjuk, hogy ehhez hozzávetőlegesen 150

*3. ábra*
*Szubjektív minősítés átlagai az egyes tematikákra*



ezer szót kellene felvenni legalább ötfajta mondatbeli pozícióban. Ez azt jelenti, hogy a meglévő adatbázishoz képest körülbelül 700 ezer szót tartalmazó mondatkorpuszt kellene a bemondóval bemondatni és feldolgozni. Ez a meglévő adatbázis 10 órájához képest, nagyságrendileg újabb 100 óra felvételt jelentene, ha sikerülne egyáltalán ezeket a mondatokat megalkotni. Ennek teljesítése irreális követelmény.

A másik megközelítés lehet a minőség javítására, hogy a korábbi szintetizátortechnikáknál használt prozódiai modulok kimeneti információit használjuk fel az általános korpuszos szintetizátorban. Tapasztalatból tudható azonban, hogy az emberi hangminőség – amelyet a szintetizátor akkor nyújt, amikor a saját tematikájának megfelelő mondatokat állít elő – nem érhető el ezzel a technikával. Ezzel a módszerrel azonban ki lehet egyenlíteni azokat a minőségbeli durva egyenetlenségeket, amelyek a meghallgatásos teszt során az észlelők kifogásoltak. Egy korábbi, elemösszefűzéses technikájú szintetizátor 2,5-es szubjektív minősítést ért el egy hasonló meghallgatásos teszt során[2]. Tehát ha ennek a szintetizátornak a prozódiai információt és a korpuszos szintetizátor bővebb hangadatbázisát egyesíteni tudjuk, akkor várhatóan a mostani 2 körüli minősítést a régebbi technikájú szintetizátor 2,5-es minősége fölé tudjuk vinni.

## 6. Összefoglalás

A korlátozott tematikára tervezett beszédadatbázis és a hozzá kapcsolódó korpuszos beszédszintetizátor változtatás nélkül nem alkalmas tetszőleges tematikájú mondatok előállítására. Amennyiben mégis ilyen irányú fejlesztést kívánunk elindítani, akkor a szintetizátor minőségének egyik javítási megoldása lehet az adatbázis növelése. Ez a jelentős mennyiségű adatbővülés miatt nehezen megvalósítható.

A másik megoldás a prozódiai modul fejlesztése, amellyel az érthetőség jól javítható. Ennek a hátránya, hogy további jelfeldolgozást kíván meg, amely a természetes hangzást ronthatja, de elkerülhető vele az egyenetlen minőség a hangzásban.

**Köszönetnyilvánítás**

Köszönöm a BME TMIT Beszédtechnológiai laboratórium munkatársainak segítségét, bátorítását.
A kutatást részben az NKFP 2. programja támogatta (szerződésszám: 2/034/2004).

**A szerzőről**

**Zainkó Csaba** 1999-ben végzett a BME Villamosmérnöki és Informatikai Kar Médiainformatika szakirányon és azóta a Távközlési és Médiainformatikai Tanszék Beszédtechnológiai laboratóriumában dialógusrendszerek és az ahhoz kapcsolódó komponensek kutatásával és fejlesztésével foglalkozik. Részt vett az első magyar nyelvű elektronikus levél felolvasó és a számszerinti tudakozó fejlesztésében. Jelenleg a korpusz-alapú beszédszintézis technológiájának vizsgálata áll kutatási témájának középpontjában.

**Irodalom**

[1] Németh Géza, Olaszy Gábor, Fék Márk:
Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei.
Beszédkutatás 2006. Szerk.: Gósy Mária.
MTA Nyelvtudományi Intézet, 2006, pp.183–196.

[2] Fék M., Pesti P., Németh G., Zainkó Cs.:
Generációváltás a beszédszintézisben.
Híradástechnika, 2006/3. pp.21–30.

[3] Olaszy Gábor:
A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései.
Híradástechnika 2006/3. pp.43–50.

[4] Olaszy G., Németh G.:
IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method.
In: Human Factors and Voice Interactive Systems, Ed.: Daryle Gardner-Bonneau.
Kluwer Academic Publishers, 1999, pp.237–256.

[5] G. Németh, Cs. Zainkó:
Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation,
Acta Linguistica Hungarica, Vol. 49. (3-4), 2002, Akadémiai Kiadó, pp.385–405.

[6] Olaszy Gábor:
Mássalhangzó-kapcsolódások a magyar beszédben.
Tinta Kiadó, Budapest, 2007.

[7] Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor:
Creating open language resources for Hungarian,
In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC) 2004.

[8] Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.:
Web-based frequency dictionaries for medium density languages,
In: Proc. of the 2nd Int. Workshop on Web as Corpus, Ed.: Adam Kilgarriff, Marco Baroni, ACL-06, 2006, pp.1–9.

[9] Olaszy Gábor:
Az artikuláció akusztikai vetülete – a hangsebészet elmélete és gyakorlata.
KIFLAF 2003, Szerk.: Hunyadi László.
Debreceni Egyetem, pp.241–254.