Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései

Nagy András, Pesti Péter, Németh Géza, Bőhm Tamás
Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék

1 Bevezetés

A távközlési, informatikai és média technológiák napjainkban tapasztalható konvergenciájának és integrációjának köszönhetően a világ az információs társadalom létrejötte felé halad. Ebben a változásban az egyik legnagyobb hatású folyamat – természetesen a hálózatok, a mobilitás és a számítógépek fejlődése mellett – az ember-gép kapcsolat átalakulása, amiben a beszédtechnológia, így a beszédszintézis is alapvető szerepet játszik.

Az utóbbi néhány évben kezdett megfogalmazódni egy új koncepció, amit korpusz-alapú beszédszintézisnek nevezünk [1]. Az elképzelés alapötletét az az általánosan elfogadott elv adja, hogy egy hullámforma-összefűzésen alapuló beszédszintézis rendszer működésének minőségét döntően az összefűzések száma határozza meg. Minél hosszabb elemekből állítjuk elő a szintetizált szöveget - az összefűzési pontok számának csökkenése miatt – annál jobb lesz az elért minőség. Az ideális tehát az lenne, ha minden lehetséges felolvasandó szöveg, de legalábbis minden lehetséges mondat szerepelne elemként a rendszer adatbázisában. Természetesen ez a gyakorlatban kivitelezhetetlen, ezért ennél rövidebb egységeket vesznek fel az adatbázisba, de azzal a céllal, hogy nagy valószínűséggel hosszú elemekből összefűzhető legyen a kimenet. A rendszer rugalmassága érdekében pedig rögzített elemhossz helyett változó hosszúságot érdemes alkalmazni [2].

A fentiek alapján külföldön már készült néhány korpusz-alapú beszédszintetizátor a világnyelvekre [2], magyar alkalmazással azonban eddig még nem találkoztunk. Munkánk célja tehát – felhasználva a korábbi megoldások (Profivox [3][4] és számfelolvasó [5]) eredményeit és tapasztalatait – egy ilyen modern beszédszintézis rendszer kifejlesztése. Mivel egy ilyen komplex szoftver létrehozása több éves feladat, ezért először egy könnyebben tervezhető, kötött tárgyterületű, időjárás-jelentéseket felolvasó rendszert alakítunk ki, majd ezt kibővítve szeretnénk eljutni a minél szélesebb – lehetőleg kötetlen – tárgyterületű rendszerig.

Jelen cikkünk ezen kutatási-fejlesztési folyamat első fázisáról számol be. A cikkben áttekintjük a korpusz-alapú rendszerek legfontosabb tervezési kihívásait, az egyes részproblémákra megoldási javaslatokat kínálunk, illetve bemutatjuk kezdeti kísérleteinket is, majd ennek segítségével értékeljük a koncepció lehetőségeit. A cikk lezárásaként összefoglaljuk az elvégzett munkát és ismertetjük a hátralevő tervezési és fejlesztési feladatokat.

2 Korpusz-alapú rendszerek tervezési kérdései

Az alábbi alfejezetekben rövid betekintést nyújtunk a korpusz-alapú rendszerek tervezési kihívásaiba, ismertetjük az általunk elvégzett vizsgálatokat és az egyes részproblémák lehetséges megoldásait.

2.1 Bemondó kiválasztása

A szintézis szempontjából rendkívül fontos, hogy a korpusz különböző pontjairól kivágott hullámforma-darabok minél pontosabban illeszkedjenek egymáshoz. Ennek biztosítására szolgál az elem-kiválasztási algoritmus, emellett azonban nagy jelentősége van annak is, hogy a beszélő mennyire képes a szöveget végig hasonló prozódiával bemondani.

Alapvető követelmény például, hogy a bemondás közben a beszélő hangjának alapfrekvenciája ne változzon túl tág határok között. Természetesen ezt jelfeldolgozás segítésével utólag is lehet módosítani, de ez rontja a szintézis minőségét. Ezért meghatároztunk olyan követelményeket, amelyeket egy bemondótól elvárunk.

Ezek a követelmények a tiszta artikuláció, a kellemes hangszín, a konzisztencia (annak vizsgálata, hogy az egyes bemondásokon belül, illetve a különböző bemondások alkalmával mennyire hasonlóan ejtette az adott bemondó ugyanazokat a hangokat) és az elérhetőség (hozzáférhető-e az adott bemondótól megfelelő mennyiségű hanganyag) voltak. A felállított követelmények alapján a legmegfelelőbb bemondó kiválasztása több lépésben történt.

A Magyar Rádió honlapján elérhető archívumból letöltöttük az ott elérhető rádiók (Kossuth, Bartók, Petőfi) két teljes napi anyagát. A hanganyagok órás bontásban, RealAudio formátumban voltak elérhetők, azonban minőségük részletes akusztikai vizsgálatok elvégzéséhez nem volt elegendő.

A hanganyagokat többször meghallgatva kigyűjtöttük az egyes bemondókra jellemző jegyeket. Ezeket összehasonlítva egymással és a kezdetben specifikált követelményekkel, elkészítettük az általunk legjobbnak tartott bemondók listáját.

A kiválasztott bemondókhoz kértünk ezután a Magyar Rádiótól jó minőségű hanganyagokat. A kapott hangfájlok már alkalmasak voltak mélyebb vizsgálatok elvégzésére is. A két legalapvetőbb vizsgált paraméter az alapfrekvencia és az intenzitás volt. Vizsgáltuk ezek értékét az időskálán, az értékek átlagát, illetve az átlag körüli szórást. A vizsgálatok alapján javaslat született a legjobbnak ítélt bemondóra.

2.2 Elemkiválasztás vizsgálata

A korpuszos szintézis kulcs-gondolata, hogy a szintetizálás során összefűzendő elemekből több példány is rendelkezésre áll, így lehetőség van egy adott mérték szerint megfelelőbb elem kiválasztására. Míg a diád-összefűzéses szintézis esetében az egyetlen tényező a diádok fonetikai címkéinek egyezése, addig a korpuszos megoldásnál több szempontot mérlegelhetünk, egy összetett költségfüggvény alkalmazásával.

A szintetizálandó beszédrészlet és a kiválasztott elem egyezését megadó mértéket cél-egyezési költségnek (*target cost*) nevezzük [1]. A szintetizált beszéd természetességét erősen befolyásolja az összefűzött elemek egymáshoz való illeszkedése. Ezt az összefűzési költséggel (*concatenation cost*) fejezzük ki. Két, az eredeti bemondásban is egymás mellett álló beszéddarab összefűzési költsége definíció szerint nulla, hiszen a kettévágott beszéd eredeti természetességében újra összeállítható.

Az egyezés vagy illeszkedés vizsgálatára hang, szótag, szó és prozódiai egység (pl. részmondat) szinten határozunk meg jellemzőket. Rendszerünkben a beszéd akusztikai jellemzőit (például hangmagasság, formánsszerkezet) egyelőre nem használjuk, mivel a nyelvi jellemzők (például hangsúly, mondat modalitása) eddigi vizsgálataink szerint lényegesen nagyobb diszkriminatív erőt képviselnek. A címkékkel ellátott hangkorpusz alapján megállapítható egy beszédrészlet illeszkedése a korpusz bármely részéhez, valamint a kiválasztott egymás utáni elemek összeilleszkedése, az egyes tényezők megfelelő súlyának beállítása után.

A súlyok beállítása meghallgatásos teszt és módosítás fázisokból álló iterációval lehetséges. Fontos lehetőség, hogy a hangok fonéma-egyezése nem abszolút követelmény, így lehetőség van arra, hogy azonos osztályba sorolt beszédhangok helyettesítsék egymást, ha az összefűzési költség ezzel jelentősen lecsökken. Ezen megoldás hasznosságának magyarázata, hogy a pontatlan beszédhang a környezetbe való jó illeszkedés miatt a hallgató személy számára akár észrevétlen is maradhat (például hangsúlytalan esetben).

A kiválasztást az elemek egymáshoz illeszkedése miatt nem elegendő elemenként végezni. Célunk az előálló szintetizált beszéd teljes minőségének maximalizálása, amire egy Viterbi-algoritmushoz [6] hasonló módszert alkalmazhatunk. A mondathatárokon átívelő akusztikai és nyelvi hatások elhanyagolhatók, ezért az optimálisan szintetizálni kívánt beszédrészlet a teljes mondat. A minimalizálandó költség a cél-egyezési és összefűzési költségek összege a teljes mondatra, minden lehetséges elem-választási kombinációra.

2.3 Elemméret megválasztása

A korpuszos szintézis sajátossága, hogy nem csupán a beilleszteni kívánt beszédelemről, hanem annak hosszáról is dönteni kell [6][7]. Amennyiben a használt összefűzési költségfüggvény az előző alfejezetben ismertetett elvárásoknak megfelelően zérus értékű két olyan elemre, melyek a beszédkorpuszban egymás mellett helyezkednek el (a bemondás során együtt fordultak elő), akkor a költségfüggvény minimalizálása implicit módon az elem hosszának meghatározását is jelenti.

Ez a megközelítés azonban a gyakorlatban nem alkalmazható. A beszédszintetizátort korlátozott, de nem zárt tartományra terveztük, vagyis a céltartomány ismerete nem zárja ki új szavak előfordulását (pl. tájegységek nevei). Ahhoz, hogy tetszőleges szó szintetizálása lehetséges legyen, a rendszernek képesnek kell lennie alapelemekből való építkezésre, azaz diádos és/vagy triádos elvű szintézisre. Amennyiben az elemhossz megválasztását a költségfüggvénytől várjuk, az elemek szükségszerűen az alapelemek (pl. diád vagy triád). Ekkor a keresési tér akár több millió elem méretű is lehet, így a megfelelő elem megtalálása (és végeredményben a teljes szintézis) túlságosan hosszú ideig fog tartani.

Egy lehetséges megoldás az elemek akusztikai csoportosítása (*acoustic clustering*, AC, [8]) úgy, hogy az egy csoportba sorolt elemeknek a cél-egyezési költségfüggvény által megadott távolsága minimális legyen. A csoportosítás a beszédkorpusz címkézésekor (offline módon)

elvégezhető, szintéziskor pedig a besorolás a keresési tér leszűkítésére használható. A megközelítés előnye, hogy a csoportosítást nem köti adott jellemzőkhöz.

Egy másik megközelítésben a hosszabb elemek (pl. szókapcsolat, szó, szótag) a beszéd-adatbázisban jelölve vannak és közvetlenül kiválaszthatók [9][10]. A PSM (*Phonological Structure Matching*, [8]) algoritmusban először a magasabb szinten lévő (így hosszabb) elemek közt keresünk beillesztésre alkalmasat. Sikertelenség esetén a következő alacsonyabb szintre lépünk. Legrosszabb esetben (pl. új szó esetén) a legalsó, diád- vagy triád-szint elemeiből építkezünk.

Az így megvalósított PSM-ben továbbra is gondot jelent a legalsó szint elemgazdagsága. Éppen ezért rendszerünkben a szegmens szint alatt az akusztikai csoportosítást (AC) alkalmazzuk diád méretű elemekre, míg a szegmens szint felett a PSM algoritmus végzi az elemméret megválasztását [8][11]. A beszédkorpuszban ennek értelmében minden lehetséges diádnak legalább egyszer szerepelnie kell. Ennek biztosítására a felolvasandó szövegtestet két részre osztjuk, melyeket eltérő szempont alapján állítunk össze. Az egyik rész biztosítja a céltartomány statisztikai jellemzői alapján megállapított gyakori szavak, szókapcsolatok fedését, és lehetővé teszi minél hosszabb elemek kiválasztását összefűzéshez. A másik rész a hangkapcsolatok fedését biztosítja a diádos szintézishez.

2.4 Adatbázis-tervezési kérdések, statisztikai vizsgálatok

A korpusz-alapú rendszerek minőségét alapvetően befolyásolja annak a beszédtestnek a megalkotása, amiből azután az elemkiválasztó algoritmus a szintézis során a szükséges, változó hosszúságú elemeket előállítja [8][12]. Ahhoz, hogy az elemkiválasztás hatékonyan megvalósulhasson, elengedhetetlen egy jól átgondolt, kellően strukturált adattárolási megoldás. Figyelni kell továbbá arra is, hogy a tervezés és megvalósítás során létrejött adatbázisban az esetlegesen szükséges későbbi bővítések a konzisztencia veszélyeztetése nélkül elvégezhetőek legyenek.

A gondos korpusztervezéshez megoldandó egy jól meghatározott, optimális méretű elemhalmaz kialakítása, amit az adatbázis tárolni fog. Az optimalitás jelen esetben azt jelenti, hogy összhangot kell találni a minőségi igényekből adódó nagy elemszám, és a teljesítményszempontok alapján elvárt kis elemszám között.

A fentiek alapján egy – méretben és összetételben – ideális elemhalmaz meghatározásának megkönnyítése érdekében statisztikai vizsgálatokat végeztünk. A vizsgálatok alapját egy folyamatosan bővülő adatbázis adja, amelyet az Interneten található időjárás-jelentések felhasználásával állítottunk elő. Az adatbázisban szóalakokat, illetve szóalak-párokat tárolunk, így az elvégzett statisztikai vizsgálatok szóalakokra, szóalak-párokra, illetve általános statisztikai tulajdonságra (például mondatok száma, modalitása) terjednek ki. Fejlesztés alatt van egy szótag-alapú adatbázis is.

Az adatbázis fő táblája tartalmazza a szóalakokat. Ebben minden szó rendelkezik egy azonosítóval, típussal (szó, szám, rövidítés, előjel, írásjel), valamint tároljuk a szót megelőző és követő szó indexét, a szóalak mondatban elfoglalt pozícióját, a szóalakot tartalmazó mondat szövegbeli helyzetét. A szóalak mondatbeli pozíciójára kétfajta – számszerű és szerkezetre utaló – információt is rögzítünk. Előbbi azt jelenti, hogy megadtuk, hogy a szó hányadik helyen van a mondatban, utóbbi pedig azt, hogy a kérdéses alak a mondat elején, végén, vessző előtt, után vagy

felsorolásban van-e. Természetesen az is lehetséges, hogy az adott szóalak egyszerre több kategóriába is tartozzon.

Mielőtt rátértünk volna a tényleges statisztikai vizsgálatokra, létrehoztunk egy rövidítéseket és azok feloldását tartalmazó adattáblát, amiben minden ilyen párhoz tároltuk annak gyakoriságát is. Elkészítettük továbbá a tipikus helyesírási hibákat tartalmazó szavak listáját, minden helytelenül írt szó mellett jelölve a helyes alakot és a hiba előfordulási gyakoriságát is. Ezen táblák létrehozása bizonyos szintű automatizálással, de döntően kézi módszerekkel történt. A gyakorlatban ez úgy történt, hogy a rövidítések esetén bizonyos szabályszerűségeket kerestünk – például hogy a három betűs, kizárólag mássalhangzókból álló szavak nagy valószínűséggel rövidítések – és ezeket felhasználva készítettünk egy listát, amit utána kézi módszerekkel pontosítottunk, a helyesírási hibák esetén pedig külső helyesírás-ellenőrző programot használtunk fel.

Vizsgálataink alapján az adatbázisban előforduló leggyakoribb hibaforma az ékezethiba volt. A rövidítés- és hibalista segítségével aztán elvégeztük a szöveg korrekcióját. Érdemes megjegyezni, hogy ezek a táblák a későbbiekben is segítséget nyújthatnak a bővítések során bekerülő új időjárás-jelentések automatikus javításában.

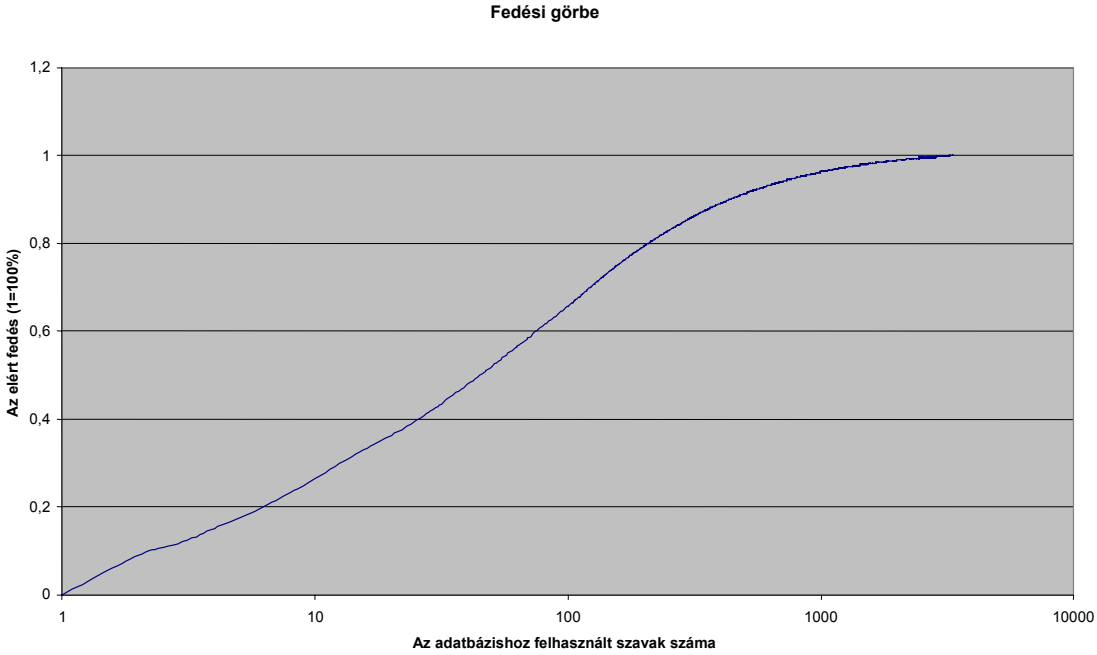
A húsz forrás (pl. *http://www.met.hu*) 2004 áprilisa és októbere közötti időjárás-jelentéseiből készített, fentiek alapján módosított táblában 14 ezer mondatban összesen 181 ezer elem (szó, szám, rövidítés, előjel[1] és írásjel) szerepel. Ebből 140 ezer szó (3300 különböző szóalak), 10 ezer szám, a többi pedig írásjel és előjel. Gyakorlatilag minden mondat kijelentő, kérdő egyáltalán nem fordult elő, felkiáltó pedig kevesebb, mint tíz. A mondatonkénti átlagos szószám (a számokat is beleszámolva) 10,7 karakter. A szavak átlagos hossza valamivel hat betű feletti, ami – figyelembe véve, hogy a leggyakoribb szavak listáját toronymagasan vezető határozott névelők hossza egy, illetve két betű – meglepőnek tűnhet. A magyarázat főként az időjárás-jelentéssel kapcsolatos gyakori kifejezések átlagosnál nagyobb hosszában keresendő (például "hőmérséklet", "várható", "csúcsértéke", "felhőzet"). A leghosszabb szó 23 betűből áll ("hőmérséklet-csökkenéssel"). Itt érdemes megjegyezni, hogy a kötőjeles szavakat egy szónak tekintettük (pl. "Dél-Dunántúl").

Készítettünk továbbá egy, a szóhosszak eloszlását tartalmazó táblát is. Vizsgálataink alapján a hat és tíz hosszúság közötti szavakból van a legtöbbféle (megjegyezzük, hogy a továbbiakban a szó és szóalak kifejezések alatt kizárólag tényleges szavakat értünk, számokat és írásjeleket nem).

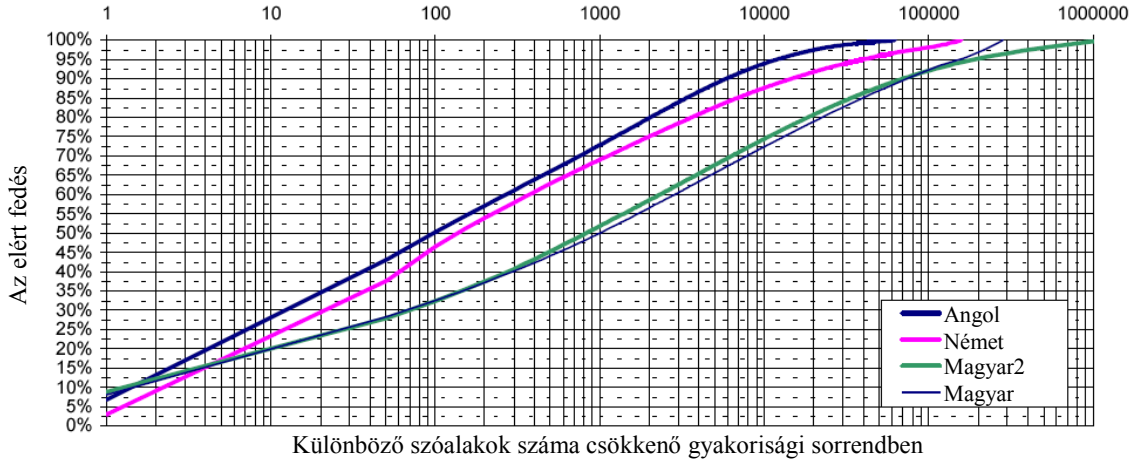
Vizsgáltuk továbbá a szavak gyakoriságát is. A készített tábla egyik oszlopa az adott alakhoz tartozó fedési százalékot is tartalmazta. A lista *k*-adik szava esetén ez azt fejezi ki, hogy ha egy olyan adatbázist készítenénk, ami gyakoriság szerint az első *k* szóalakot tartalmazná, akkor ez a kapott "korpusz" az időjárás-jelentéseink szavainak a fedési százalék szerinti részét tartalmazná. Vizsgálataink alapján arra a következtetésre jutottunk, hogy a leggyakoribb tíz szó segítségével 28%-os fedés érhető el. Ötszáz szó esetén 91%-os, míg 2000 felhasználásával 99%-os a fedési százalék. Általános tárgyterületű korpusz esetében a 90%-os szint eléréséhez 70000 szóalakra van szükség [13]. Ez a rendkívül ígéretes eredmény a kötött tárgyterületnek és a csupán félévnyi időjárás-jelentés anyagnak köszönhető (mivel nagy mennyiségben nem álltak rendelkezésre archív anyagok, ezért az adatbázis készítéséhez csak az éppen aktuálisan feltett időjárás-jelentéseket tudtuk felhasználni). További vizsgálataink egyik fontos kérdése, hogy az utóbbinak milyen mértékben. Amint rendelkezésünkre áll megfelelő mennyiségű adat, eredményeinket pontosítjuk és a kapott új értékeket összehasonlítjuk a korábbiakkal. Az alábbiakban (1. és 2. ábra)

[1] Előjel a "+" és a "-" jel, valamint a "plusz" és "mínusz" szavak.

összehasonlításképpen közöljük a kötött és az általános tárgyterületű rendszer esetében kapott fedési diagramokat.



1. ábra. Időjárás-jelentés fedési görbe



2. ábra. Általános tárgyterületű fedési görbék (forrás: [13])

Figyelembe véve, hogy egy adott szóalakból többfajta környezetnek megfelelő változatra is szükség van, ennél a számnál a megvalósítandó korpuszban lényegesen több elem lesz, hiszen például máshogy kell ejteni egy szót a környező hangoktól vagy a mondatbeli pozíciótól függően. Éppen ilyen megfontolások miatt tartottuk fontosnak egy olyan statisztika elkészítését, amely a szavak gyakoriságát a jobb- és baloldali környezetnek megfelelően vizsgálja. A vizsgálat gyakorlati

jelentősége abban áll, hogy egy gyakori szót (például hőmérséklet) érdemes a leggyakoribb előfordulási helyzeteiben felvenni.

Természetesen ehhez figyelembe kell venni például a mondatbeli pozíciót is, ezért is tartottuk fontosnak a fő táblát úgy elkészíteni, hogy az ezzel kapcsolatban is tartalmazzon információt. A gyakorlati megvalósításnál nem a számszerű mondatbeli pozíció, sokkal inkább az általunk szerkezetinek nevezett információ az érdekes. Másképpen kell ugyanis a szavakat ejteni mondat elején, mondat végén, vessző előtt és után (azaz tagmondat határán vagy felsorolás elemei között). A beszédtest szövegének megállapításakor ezt is figyelembe fogjuk venni. A statisztika készítésekor nem csak a különálló szavakat érdemes vizsgálni, hanem a szópárokat, általános esetben szó *n*-eseket is, hogy a gyakori szókapcsolatokat, szófordulatokat hatékonyan tudjuk kezelni.

Végezetül a fentiek mellett elkészítettük az időjárás-jelentésekben található idegen szavak (például "Dubrovnik") jegyzékét is, mivel ezek helyes szintetizálására külön kell figyelni. Egy lehetséges megoldás egy kivételszótár felvétele, ami tárolja, hogy a kérdéses idegen szót hogyan kell magyarul ejteni ("Dubrovnik" esetében a szótárban "dubrovnyik" ejtési alak szerepelne).

Áttérve a korpusz kialakításának kérdésére, a tervezés során a rendelkezésre álló nagyméretű szövegtestből kell kiválasztani olyan kisméretű halmazt, amely jól fedi a teljes szöveget. Ehhez a kutatók általában a mohó (*greedy*) algoritmust használják [12]. Ez egy egyszerű iteratív megoldás, amely egy nagyméretű szövegből választ ki úgy mondatokat, hogy a kapott kisméretű halmaz minél jobban lefedje a teljeset. Minden lépésben olyan mondat kerül bele a halmazba, amelyik a legtöbb még lefedetlen elemet tartalmazza. Egy adott elem akkor nincs lefedve, ha az épülő halmazban nincs azonos tulajdonságvektorral rendelkező korpuszelem. A tulajdonságvektor a figyelni kívánt paraméterekből (például hangsúlyosság, intenzitás) álló vektor. Az egyes értékek azt fejezik ki, hogy egy adott tulajdonság mennyire teljesül az adott elemre. Az iteratív eljárás akkor áll meg, ha valamilyen peremfeltételt már teljesít az épülő halmaz (például elér egy adott fedési százalékot).

Az algoritmus sikerének egyik sarkalatos pontja a tulajdonságvektor méretének és összetételének meghatározása. Túl nagy méret esetén szinte minden előforduló elem lefedetlen, míg túl kis méret esetén a döntő többség akár többszörösen is lefedett lesz. Optimális megoldást nem ismerünk, bár számos javaslat történt a tulajdonságtér összetételére (például a hangsúlyosságot, jobb- és baloldali környezetet érdemes figyelni). A hagyományos megoldás minden tulajdonság teljesüléséhez bináris értéket rendel (teljesül vagy nem), létezik azonban olyan megvalósítás is, amely bizonyos tulajdonságok esetében megenged 0 és 1 közötti tetszőleges értéket is. Ilyen tulajdonság lehet például a baloldali környezet egyezősége. A hangsúlyosság vizsgálatakor ugyanakkor két lehetséges érték megengedése ésszerű (hangsúlyos vagy nem). A módosított megoldás működését döntően befolyásolja a nem bináris értékek megfelelő definiálása.

Az algoritmust általában a teljes diád-fedés megvalósítására használják, azonban kiegészíthető úgy is, hogy emellett olyan mondatokat is kiválasszon, amelyekben az elemek (szavak, szópárok) az általunk végzett statisztikai vizsgálatok alapján gyakoriak és ezért érdemes szerepelniük a korpuszban. A vizsgálatnál, mint ahogy azt már korábban is említettük, a puszta elemgyakoriság mellett érdemes figyelembe venni egyéb tényezőket, például a környezetet is.

Az adatbázis tervezése során tekintettel kell lenni továbbá az LNRE-szabályra (*Large Number of Rare Events*, [14]), mely szerint bár a lehetséges, szintézis során felhasználható beszédelemek közül a döntő többség nagyon ritka, így gyakorlatilag külön-külön alig fordulnak elő

egy szintetizálandó szövegben, ezen ritka elemek együttes száma már olyan jelentős, hogy nagy valószínűséggel szükség van közülük néhányra egy adott szöveg szintetizálása során.

Mivel a beszédtestet úgy célszerű kialakítani, hogy az a leggyakoribb szótagokat, szavakat, szókapcsolatokat, mondatokat tartalmazza, ezért a fentiek alapján gyakorlatilag minden szintetizálandó szövegben előfordulnak olyan részek, amelyekre nincs megfelelő elem az adatbázisban. Ezért célszerű a beszédtestet úgy felépíteni, hogy az minden lehetséges diádot tartalmazzon, a leggyakoribbakat több környezetben is. Ezzel biztosítható az, hogy minden lehetséges szintetizálandó szövegrész előállítható legyen, legrosszabb esetben diádok segítségével [11]. A lehetséges kapcsolatok száma minden esetben a hangok száma plusz egynek (mivel a szünet is benne lehet a hangkapcsolatban) a négyzete. A lehetséges diádok közül azonban nem szükséges mindegyik: a teljes diád-fedés kialakításához az európai nyelvek esetében becslések szerint legfeljebb néhány ezer hangkapcsolatra van szükség [15].

Amennyiben a fent említetteknek megfelelően kialakítottuk a korpuszt, gondoskodni kell annak jól strukturált, hatékony elemkiválasztást lehetővé tevő adatstruktúrában való tárolásáról. A megtervezett struktúra alapvetően három lényegileg különböző részből áll.

Az első a konkrét hullámformákat tartalmazó fájlok halmaza. Minden fájlban egy mondat felcímkézett hanganyagát tároljuk. Ezzel a megoldással egyrészt a fájlok mérete kellően kicsi lesz, másrészt mivel ritkán van szükség a szintézis során olyan elemre, ami két mondat határán van, ezért egy beszédelem betöltéséhez általában elegendő egyetlen fájl használata is.

A struktúra második része a diádokat tartalmazza. Ezeket a diádokat, a rájuk jellemző vektorral, illetve az őt tároló fájlra vonatkozó hivatkozással egy fában tároljuk. A fában a diádok a beszédtestnek megfelelő sorrendben vannak, vagyis a fa inorder bejárása segítségével pont a korpuszt kapjuk vissza [16]. Ennek jelentősége abban áll, hogy így lehetőség van az általunk kívánt változó elemhossz egyszerű megvalósítására. Ehhez mindössze a fában az adott elemtől inorder bejárással indulva kell bővíteni egy adott diádot hosszabb elemmé az út közben érintett csúcsokhoz tartozó diádok összefűzésével.

A struktúra harmadik része a fában való keresést könnyíti. Ehhez egy szófát alakítottunk ki [17], ami tárolja a lehetséges diádokat úgy, hogy a fa csúcsai a diádoknak lehetséges előtagjai (prefixei), a levelei maguk a diádok. A levelekben tároltuk továbbá az adott elem másik fában elfoglalt helyét.

Amikor szükség van egy elemre, ami egy adott diáddal kezdődik, akkor a szófában megkeressük ezt a diádot tartalmazó levelet, ami mutatót tartalmaz a kérdéses diád fában elfoglalt pozíciójára (amennyiben több ilyen diád is van, akkor a lehetséges pozíciók láncolt listában szerepelnek). A fában aztán az előbb ismertetett bejárással megkereshető a kívánt hosszabb elem.

A fenti struktúra továbbfejleszthető azzal, hogy nem csak diádokról készítünk fát és szófát, hanem szavakról, szópárokról és mondatokról is. A struktúra kialakítása során figyelni kell a konzisztencia könnyű fenntarthatóságára. Mivel az adatbázis bővítése a szintézis előtt már megtörténik, ezért ilyenkor lehetőség van a hullámformákat tároló fájlok mellett a keresést segítő struktúra frissítésére is, így biztosítható, hogy nem lesz nem kívánt inkonzisztencia.

2.5 Meghallgatásos kísérletek

A Magyar Rádió online archívumából összegyűjtött időjárás-jelentésekre építve elvégeztük a rendszer működési elvének tesztelését. Bár a közelítőleg kétnapnyi időtartam alatt a Kossuth, Petőfi és Bartók rádiókban beolvasott időjárás-jelentések szövegét nem mi terveztük, mégis képet kaphatunk a kész rendszer várható beszédminőségéről, ha az elveket figyelembe véve kézzel szintetizálunk időjárás-jelentés részleteket.

A kétnapos periódusból összesen 149 időjárás-jelentést gyűjtöttünk össze 22 bemondótól. Mivel egy szintetizált beszédrészlethez csak egy bemondótól származó hangfájlok használhatók, így a kézi szintézis során nehezebb helyzetben voltunk, mint a készülő rendszer, hiszen egy bemondótól nagyon kicsi beszédkorpuszunk volt. Sok szó csak egyszer fordul elő az egy bemondótól rendelkezésre álló hangfájlokban. Ugyanakkor a beolvasások egy rövid időszakból származnak, így szintén sok szó szinte mindegyik hangfájlban megtalálható (ugyanarra az időszakra nagyjából ugyanazt a prognózist közlik).

A kézi szintézist az összehasonlíthatóság érdekében olyan mondatokra végeztük el, melyek egy más bemondó általi realizációban már rendelkezésre álltak a hangfájlokban. Éppen ezért, valamint a bemondónkénti beszédkorpusz erős korlátozottsága miatt nem választhattunk tetszőleges mondatokat. Első lépésként azon bemondók szövegeinek átiratát készítettük el, akiktől a legtöbb bemondásunk volt. Az így elkészült 54 időjárásjelentés-szövegben ezután olyan mondatokat kerestünk, melyek szavai (lehetőleg a megfelelő szövegkörnyezetben) egy másik bemondónál is megtalálhatók voltak. Ezek alapján öt mondat szintézisét végeztük el. Az 1. táblázatban látható, hogy a szintetizált mondatok számos vágási pontot tartalmaznak. A vágási pontokat a mondatokban függőleges vonallal jeleztük.

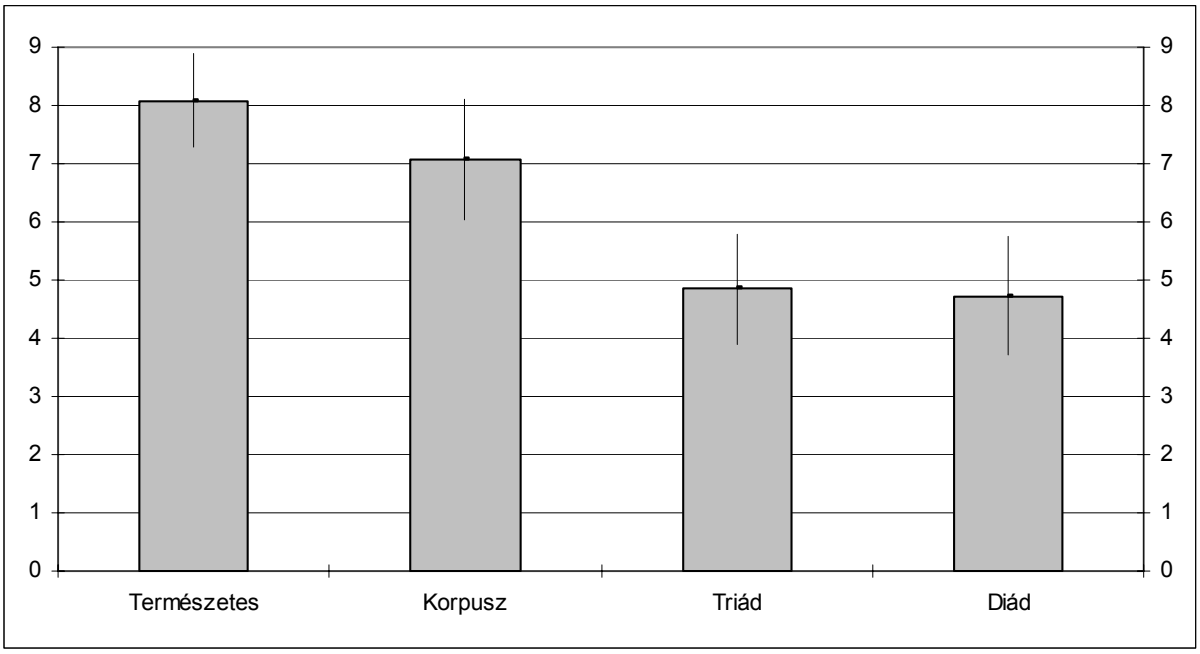
Mondat	Eredeti bemondó	Szintetizált bemondás bemondója	Összefűzött elemek száma
A hőmérséklet hajnalban mínusz | egy, | mínusz | hat, | holnap | napközben | mínusz | egy, | plusz négy fok | között alakul.	András	Adrienn	10
Napközben | országszerte várható csapadék, északon, | északnyugaton | havazás, | délkeleten eső, | másutt | havas | eső, | ónos | eső.	Erika	Klára	10
Végül az időjárásról: | mindenütt beborul | az ég, | reggelig egyre többfelé | lehet gyenge havazás, | hószállingózás.	Zsuzsa	András	6
A hőmérséklet | kora délután | kettő | és | hét | fok | között | alakul.	Zsuzsa	András	8
A nyugati, | északnyugati | szelet | sokfelé erős, | a Dunántúlon | helyenként | viharos | lökések | kísérik	Zsuzsa	István	9

1. táblázat. Szintetizált mondatok tulajdonságai

A mondatokat szintetizáltuk a Profivox szintetizátor diádos, valamint készülő triádos szintetizátorával is, férfi és női hang alkalmazásával. Ezután a kapott diádos és a triádos szintézisekből is kiválasztottunk ötöt-ötöt.

A meghallgatásos tesztben 5 magyar anyanyelvű személyt kértünk meg arra, hogy értékelje a véletlenszerű sorrendben lejátszott 20 hangfájl minőségét egy 1-től 9-ig terjedő skálán. A különböző beszéd-előállítási módonkénti átlagot és szórást az 3. ábra mutatja.

A teszteredményekből látható, hogy a triádos rendszer a diádoshoz képest csak enyhe minőségi javulást jelentett; ezt részben az is okozhatja, hogy a triádos implementáció még fejlesztés alatt áll (például új jelfeldolgozási algoritmusok integrálása van folyamatban), valamint a Profivox-szal szintetizált mondatok prozódiája általános szövegekre készült, így érthető az időjárás-jelentésre optimalizált kézi szintézistől jelentősen rosszabb eredmény. A kézileg előállított korpuszos mondatok több mint két ponttal jobb eredményt értek el a kis számú, fix hosszúságú elemekből építkező diádos és triádos megoldásoknál, ugyanakkor egy ponttal a valódi bemondók teljesítménye alatt maradtak. Minden csoportnál 1-hez közeli a szórás, ami az eredmények általánosíthatóságát támasztja alá. Az eredeti bemondások megítélésében voltak a legbiztosabbak a meghallgatást végző személyek, ami egyezik várakozásainkkal, hiszen természetes beszéd hallgatásához vagyunk szokva.



3. ábra. A meghallgatásos teszt eredményei: átlag és szórás

A teszteredmények alátámasztják egy korpuszos elven működő rendszer létjogosultságát és jelentős minőségi javulást jeleznek elő. A rögzített elemhosszúságú rendszerek esetén a működés elvéből következően a jövőben nem várható olyan mértékű minőségi ugrás, mint ami a korpuszos rendszerrel elérhető. Természetesen ez bizonyos kompromisszumokkal jár együtt: nagyobb adatbázis felvétele, feldolgozása, tárolása és használata szükséges, a tematika kötött, és a szintézis számításigénye is nagyobb.

3 Összefoglaló

A korpusz-alapú megközelítés új, Magyarországon eddig nem alkalmazott koncepció, ami rugalmasabb, jobb minőséget nyújtó szintézisre ad lehetőséget. Vázoltuk a módszer alapvető elveit, részletesen foglalkoztunk a BME TMIT Beszédkutatási Laboratóriumban fejlesztés alatt álló, korpusz-elven működő, magyar nyelvű, kötött tárgyterületű rendszer fejlesztési kérdéseivel. Ismertettük többek között az időjárás-jelentésekkel kapcsolatos statisztikai vizsgálatainkat, a bemondó kiválasztásának szempontjait és menetét, valamint vizsgáltuk és értékeltük a korpusz-alapú rendszerek egyéb tervezési kérdéseit. Meghallgatásos teszteket végeztünk a korpuszos elven működő rendszer várható minőségének előrejelzésére.

A biztató eredményekre építve következő lépésünk a megtervezett rendszer implementálása lesz. A meteorológiai tárgyterületre elvégzett statisztikai vizsgálatok alapján összeállított és felolvasott beszédkorpuszhoz elkészítjük a több elemméret szinten választást lehetővé tevő algoritmust. Kezdeti megvalósításunkban szó és szókapcsolat szintek megkülönböztetését tervezzük, miközben kizárólag a céltartományra koncentrálunk, így a tetszőleges szó szintézisét lehetővé tevő, akusztikai csoportosításon alapuló diádos szintézist az implementáció második fázisára tervezzük. A cél-egyezési és összefűzési költségekben szerepet játszó jellemzők egyezési mértékének súlyozását meghallgatásos teszteket magába foglaló iterációk sorozatával kívánjuk megvalósítani.

Köszönetnyilvánítás

A szerzők a munka elvégzéséhez sok segítséget kaptak a BME TMIT Beszédtechnológiai Laboratórium munkatársaitól. Külön köszönjük a Magyar Rádiónak, hogy hozzáférést adott jó minőségű időjárás-jelentés felvételeihez.

Irodalomjegyzék

[1] Bernd Möbius, “Corpus-Based Speech Synthesis: Methods and Challenges”, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87-116., 2000.

[2] Yi, J.R.W., Glass, J.R., “Natural-Sounding Speech Synthesis using Variable-Length Units”, *Proc. ICSLP-98*, Sydney Australia, Vol. 4, pp. 1167-1170, 1998.

[3] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G., “PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications”, *International Journal of Speech Technology*, Volume 3, Numbers 3/4, pp. 201-216., December 2000.

[4] Olaszi Péter, “Magyar nyelvű beszéd-szöveg átalakítás: nyelvi modellek, algoritmusok és megvalósításuk”, 5-15.o., doktori értekezés, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2002.

[5] G. Olaszy, G. Németh, “IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method", in D. Gardner-Bonneau ed., *Human Factors and Interactive Voice Response Systems*, Kluwer, pp. 237-255., 1999.

[6] Jon Rong-Wei Yi, “Natural-Sounding Speech Synthesis Using Variable-Length Units”, Master of Engineering Thesis, Massachusetts Institute of Technology, 1997.
[7] S. P. Kishore and Alan W. Black, “Unit Size in Unit Selection Speech Synthesis”, *Eurospeech 2003*, pp. 1317-1320., 2003.
[8] Antje Schweitzer, Norbert Braunschweiler, Tanja Klankert, Bernd Möbius, Bettina Sauberlich, “Restricted Unlimited Domain Synthesis”, *Eurospeech 2003*, pp. 1321-1324., 2003.
[9] Eric Lewis and Mark Tatham, “Word and Syllable Concatenation in Text-to-Speech Synthesis”, *Eurospeech 2001*, vol. 2, pp. 615-618., 1999.
[10] Eric Lewis and Mark Tatham, “Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis”, *Eurospeech 2001*, pp. 1703-1706., 2001.
[11] Michael Pucher, Friedrich Neubarth, Erhard Rank, Georg Niklfeld, Qi Guan, “Combining Non-uniform Unit Selection with Diphone Based Synthesis”, *Eurospeech 2003*, pp. 1329-1332., 2003.
[12] Baris Bozkurt, Ozlem Ozturk, Thierry Dutoit, “Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection”, *Eurospeech 2003*, pp. 277-280., 2003.
[13] G. Németh, Cs. Zainkó, “Word Unit Based Multilingual Comparative Analysis of Text Corpora”, *Eurospeech 2001*, pp. 2035-2038., 2001.
[14] Ove Andersen, Charles Hoequist, “Keeping Rare Events Rare”, *Eurospeech 2003,* vol. 2., pp. 1337-1340, 2003.
[15] Dr. Gordos Géza, Takács György, “Digitális beszédfeldolgozás”, *Műszaki Könyvkiadó*, pp. 191-197, 1983.
[16] Rónyai L., Ivanyos G., Szabó R., “Algoritmusok”, *Typotex*, p. 60., 1999.
[17] Knuth, D. E., “A számítógép-programozás művészete”, *Műszaki Könyvkiadó*, Budapest, p. 503., 1988.