ÚJ RENDSZERŰ, KORPUSZALAPÚ GÉPI SZÖVEGFELOLVASÓ FEJLESZTÉSE ÉS KÍSÉRLETI EREDMÉNYEI

Németh Géza – Olaszy Gábor – Fék Márk

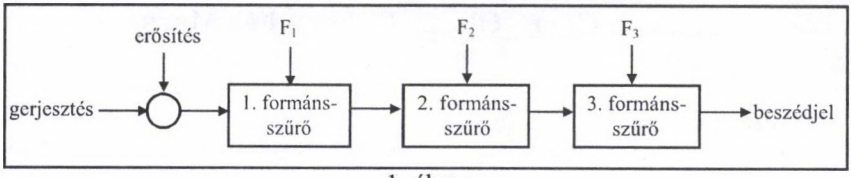
Bevezetés

A beszéd gépi előállítása ma már nem különlegesség, hiszen a mindennapi életünket körülvevő információközvetítő eszközökben, rendszerekben gyakran alkalmazzák. A géppel generált beszéd minősége változhat attól függően, hogy milyen információt kell beszéddé alakítani. Minél inkább kötött a tematika, ebből következően a szintetizálandó mondatok szerkezete, annál jobb minőséget lehet elérni. Emellett a technikai lehetőségek is behatárolhatják, hogy milyen módszert alkalmaznak a beszéd előállítására (például mobiltelefonban szól, vagy egy központi szerver generálja a beszédjelet). A kutatók célja, hogy minél jobb minőségű gépi beszédet állítsanak elő. Az eddig kifejlesztett módszerek egyre jobban érthető beszédet szolgáltattak, a minőség is közeledik az emberi hangzáshoz (nem robotos), azonban például a beszélő egyéni hangszínezetét, kiejtési stílusát még ritkán lehet felismerni a gépi szövegfelolvasóknál. Hangszínezeten értjük a személy egyéni hangját (voice timbre) normál beszédben, amely alapján például egy ismerőse felismeri a beszélőt. A korpuszalapú beszédszintézis módszere e legutóbbi hiányosságot is teljesíteni tudja, ezért várhatóan egyre jobban terjedni fog. A tanulmányban bemutatjuk az első, Magyarországon fejlesztett ilyen rendszer kísérleti eredményeit, ugyanakkor rávilágítunk arra is, hogy milyen korlátokkal kell számolni.

A beszédszintézis korábbi módszerei

Minden szövegfelolvasó rendszer elméletileg két alapvető részből épül fel. Az első rész a bemeneti szöveget értelmezi, és szimbolikus információvá alakítja, a második a szimbolikus információt alakítja át beszéd-hullámformává (általában valamilyen hangfájlt állít elő). A közbenső szimbolikus információ általában a szöveg tartalmát megadó fonémasorozatból és a beszéd prozódiai jellemzőit (hanglejtés, hangsúlyok, ritmika) leíró információkból áll össze. A megoldási módszerek között abban vannak eltérések, hogy az előbbi két részt megvalósító belső megoldások milyen elvek szerint történnek. Az alábbiakban röviden sorra vesszük az eddigi beszédszintézis-technológiák főbb jellemzőit. Megjegyezzük, hogy a technikai lehetőségek fejlődése döntően befolyásolta a módszerek kialakulását is.

A **formánsszintézis** volt az első olyan technológia, amelynek segítségével egy szöveget automatikusan beszéddé lehetett alakítani. Az eljárást napjainkban is alkalmazzák. A formánsszintézissel az ember beszédképzési folyamatát utánozzák (gerjesztő jel + toldalékcső) elektronikus formában (gerjesztett, dinamikusan hangolt szűrőrendszer). A formánsszintetizátor egy lehetséges megvalósítását az 1. ábra mutatja.



1. ábra
Soros elrendezésű formánsszintetizátor blokkvázlata

A gerjesztés zöngés hangoknál a hangszalagok által keltett kváziperiodikus jelnek feleltethető meg, illetve zöngétlen hangok esetén zajszerű. Egy-egy szűrő a megadott formánsfrekvencia környezetében erősíti a gerjesztés felharmonikusait, ezzel modellezve a garat, a gége és a szájüreg által alkotott rezonátorrendszer erősítéseit. A formánsfrekvenciák definíciószerűen a zöngés hangokat jellemzik, de zöngétlen gerjesztésnél is felhasználhatók a zörejes gerjesztésű hangok jellemző zörejgócainak az előállítására (ez műszaki egyszerűsítés). Az első három formánsfrekvencia jól leír egy-egy beszédhangot. A formánsok adatait (a szűrők rezonanciafrekvenciáit) legalább 10 ms-onként meg kell adni ahhoz, hogy érthető beszéd álljon elő. Az adatok megadását szabályrendszer vezérli, amelyet általában fonetikusok határoznak meg. A formánsszintézis során a beszéd prozódiáját is szabályrendszerrel írják le, majd a szabályok alapján változtatják az alapfrekvenciát (a beszéddallam, illetve a hangsúlyozás kialakítására), az intenzitást (a hangsúlyozás, illetve a hangerő beállítására) és a hangok időtartamát (a ritmus megvalósítására). A formánsszintetizátort tehát úgy jellemezhetjük, hogy a beszéd szegmentális és szupraszegmentális részét egyaránt szabályok alapján állítja elő. A gyakorlat azt mutatja, hogy ez a módszer ugyan érthető beszédet szolgáltat, de nem adja vissza az emberi hangszínezet finomságait (robotos a hangzás). A módszer előnye a kis tárkapacitás- és az alacsony számításigény. A BME Távközlési és Médiainformatikai Tanszékén (TMIT) kifejlesztett Multivox magyar nyelvű formánsszintetizátor (Olaszy et al. 1992) ingyenesen hozzáférhető.

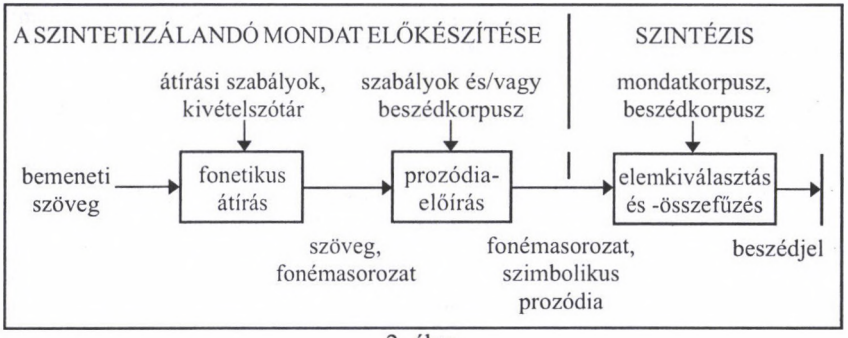
Az **elem-összefűzéses** módszert a beszéd minőségének javítására dolgozták ki. A ma működő gyakorlati alkalmazások többsége ezen az eljáráson alapul. A döntő különbség a formánsszintetizátorhoz képest az, hogy a beszédhangok hullámformájának előállítási módszere változott. A prozódiai részt itt is szabályokkal modellezik. Az eljárás lényege, hogy a formánsfrekvenciák időről időre történő megadása helyett visszatértek az emberi beszédhez. A számadatok helyett az em-

ber által felolvasott beszédből kivágott, kész hullámformákat tárolják el, majd ezeket fűzik össze. A gondolat lényege, hogy ezekben a hullámformarészletekben automatikusan benne vannak a formánsok frekvenciaértékei, azok változása a koartikulációban, továbbá még a gerjesztést sem kell parametrikusan megadni. Elvárható volt tehát, hogy ha ilyen hullámformaelemeket összefűzünk, akkor olyan hangszínezetet kapunk, amely közelebb áll az emberi beszédhez, mint a formánsszintetizátor hangja. Ezért nevezték el a módszert elem-összefűzéses technikának. A technológia egyik alapvető kérdése, hogy melyek legyenek azok a hullámformaelemek, amelyek összefűzésével létrehozzuk a gépi beszédet. Úgy kell tervezni, hogy az adott nyelv tetszőleges hangsorozatát elő tudjuk állítani. A gyakorlatban bevált kompromisszumos megoldás a két egymással összekapcsolódó félhang együtteseként előálló diád hullámformájának a tárolása. A magyarban 1600 diáddal lefedhető a teljes hang- és hangkapcsolódási állomány (Olaszy 1999). A szupraszegmentális szerkezet paramétereit ugyanakkor itt is szabályrendszerrel határozzák meg, és jelfeldolgozással végzik az alapfrekvencia-, az intenzitás- és a hangidőtartam-változtatásokat. Az eljárás előnye a szebb, emberibb hang, továbbá az is, hogy a rendszer működésének minden részlete jól kézben tartható (ez fontos a hibakeresésnél). Hátránya, hogy a hangokat a közepükön fizikailag elvágjuk, ami torzítást okoz, főként a zöngés hangok esetén. A torzítás csökkentésére alakították ki a triádos hullámformát, mint másik alapelemet, amelyben egy félhang + egy teljes hang + egy félhang jelent egy-egy tárolt elemet. Így a középső hangban nincs vágás, tehát a torzítás csökken. A legtöbb diádos, diádos és triádos hullámforma-elemtárra jellemző, hogy minden elem csak egyszer fordul elő (zárt szerkezetű elemtár), így az összefűzésnél nem lehet az elemtárban lévő elemekből válogatni (ha a hangsor több pontján van szükség ugyanarra az elemre, akkor az minden esetben ugyanaz lesz, belső tartalma nem változik). A zárt elemtárral a finom beszédrészleteket nem lehet megvalósítani, ugyan az egyénenkénti hangkarakter már részben hallható (más bemondótól készített elemtár más hangot eredményez), azonban a beszélő egyéni hangszínét továbbra sem tudják visszaadni. Ennek oka egyrészről a zárt szerkezetű elemtár, másrészről hogy az egymás mellé illesztett beszédhullámformák (diád, triád) túl rövidek ahhoz, hogy az egyéni jellegzetességek karakterisztikusan érvényesüljenek. További ok, hogy a prozódia továbbra is az egyéni jellemzőket figyelmen kívül hagyó szabályrendszer alapján áll elő. A BME TMIT-en kifejlesztett Profivox magyar nyelvű beszédszintetizátornak (Olaszy et al. 2000) létezik tiszta diádos elemtárral működő változata, illetve 1444 diádból és 6000 triádból álló nagyobb elemtára. A kétféle rendszer hangzásában (ugyanazon bemondótól származó elemtárral) a triádok alkalmazása minőségi javulást eredményez (Nagy et al. 2005).

A korpuszalapú beszédszintézis

A korpuszalapú beszédkeltés adja napjainkban a legjobb minőségű szintetizált beszédet (Schweitzer et al. 2003). Ennél a módszernél a beszélő hangszínezete, kiejtési stílusa egyértelműen felismerhető. A technológia nevéből adódik, hogy

egy adott, a beszédhangok változatossága szempontjából elméletileg nyitottnak tekinthető **beszédkorpusz** helyettesíti a korábban alkalmazott zárt szerkezetű, diádos, triádos elemtárat. A nagy korpusz sok órányi beszédet tartalmaz, annak hullámformaszintű tárolásával és címkézésével. A beszédkorpusszal párhuzamosan tároljuk annak szöveges formáját is, ezt nevezzük **mondatkorpusz**nak. Ebben végezzük a keresést, majd, ha megtaláltuk a megfelelő szövegrészt, azt kiemeljük a beszédkorpuszból. Az így kiemelt hullámformákat összefűzzük, így áll elő a szintetizált mondat hullámformája. A korábbi rendszerek módszertanához képest ez nagy változás. A formánsszintetizátoroknál döntően fonetikai tudásra volt szükség a vezérlő paraméterek meghatározására mind szegmentális, mind szupraszegmentális szinten. Az elem-összefűzéses technológiánál a hangsorszintű összeállításnál már visszakanyarodtak a kutatók az emberi beszédhez (diádok, triádok tárolása), és csak a prozódia megvalósításához használtak konkrét fonetikai ismereteket. A korpuszalapú rendszereknél a hangsor-összeállításhoz igen nagy méretű, nyitott beszéd- és mondatkorpuszt használnak, és a szegmentális és szupraszegmentális szerkezetre vonatkoztatott összegzett fonetikai tudást áttételesen építik be az elemkereső és összefűző költségfüggvénybe. A módszer mögötti mérnöki gondolat a következő. Egy bemondó által felolvasott mondat tartalmazza a hangsor hullámformáját, és ebben benne foglaltatik a prozódia is. Ha sok mondatot tárolunk az egyébként nyitott beszédkorpuszban, akkor nagy a valószínűsége annak, hogy egy szintetizálandó mondat előállítására a korpuszban találunk olyan szavakat, szókapcsolatokat, mondatrészeket, amelyek a legjobban illeszkednek (hangsorilag és prozódiailag) a készítendő beszéd adott pontjához. A mondatot így viszonylag hosszú egységekből, szavak, szókapcsolatok hullámformáinak sorozatából össze lehet állítani (2. ábra).



2. ábra
Korpuszalapú beszédszintetizátor blokksémája

Ha a szavak kiválasztásánál a prozódiai tartalmat is figyeljük, és csak a megfelelő prozódiájú szót engedjük kiválasztani a beszédkorpuszból, akkor a szintetizált mondatban a prozódia is meg fogja közelíteni az optimálist. A két feltétel

teljesítéséhez – elsősorban a szókapcsolások folytonosságának biztosításához – a szóvégi és szókezdő hangok szerinti fonetikai szabályrendszert kell alkalmazni. Ez megmondja a válogatásnál, hogy két összekapcsolandó szó találkozási pontján a hangok spektrálisan (pl. formánsmenetben) illeszkednek-e. Modellezni kell ezenfelül a mondat prozódiáját is. Ha ugyanazt a modellt használjuk a beszédkorpuszban, a mondatkorpuszban és a szintetizálandó mondatban is, akkor a prozódiai tartalom szempontjából is tudunk keresni a beszédkorpuszban, el tudjuk dönteni, hogy a hangszinten összekapcsolhatónak ítélt szó, szókapcsolat prozódiailag is megfelelő-e. Ha igen, akkor alkalmazzuk, ha nem, akkor tovább keresünk. (A beszédkorpusz mondatainak sokasága elméletileg nincs korlátozva, azonban a feldolgozási sebesség és az optimális keresési módszer behatárolja ezt a szabadságot.) A hipotézis tehát az, hogy a beszédkorpuszból való gondos összeválogatás és egymáshoz kapcsolás esetén az előállított szintetizált beszéd minősége igen közel lesz a korpusz eredeti beszédének a minőségéhez. A fentiek megvalósítását lehetővé tette a számítástechnikai eszközök rohamos fejlődése memóriakapacitásban és feldolgozási sebességben. Jó példa erre egy Japánban készített rendszer, amelynek beszédkorpuszát 80 órás hangfelvételből alakították ki (Kawai et al. 2004). Az ilyen rendszerek legfőbb előnye, hogy a hangminőség nagyon jó lehet (a szintetizált mondatot össze lehet téveszteni az eredeti bemondó hangjával). Hátránya, hogy nehéz az összefűzendő elemeket kiválasztó függvényt optimálisra megtervezni, hogy hatalmas adatbázisokkal kell dolgozni, hogy nehéz a hibakeresés, valamint a javítás. Ezek miatt első lépésben csak kötött, jól körülhatárolt témában van esély az optimális eredmény elérésére. A fentieket figyelembe véve kísérleti rendszerünkkel csak időjárás-jelentések jó minőségű előállítását tűztük ki célul. A módszerrel előállított beszédjel természetessége, így hangminősége ugrásszerű javulást mutat a korábbi módszerekhez képest. Ezt az teszi lehetővé, hogy a diád és triád elemeknél hosszabb, folytonos beszéddarabok kerülnek egymás mellé, megőrizve azok eredeti hangszínezetét, ritmikai tartalmát, hanglejtését. Továbbá a nagy beszédkorpusz lehetővé teszi, hogy spektrálisan is, és prozódialiag is egymáshoz jobban illeszkedő beszédszakaszokat válogassunk és fűzzünk össze, ami szintén a természetes hangzást biztosítja. Ez a módszer közelíti meg legjobban a biológiai beszédkeltést, azt, hogy **az emberi beszéd egyedi és egyszeri produktum.** Az egyedi jelző az egyén saját hangjára vonatkozik, az egyszeri pedig azt fejezi ki, hogy a beszédjelet az adott időpillanatra jellemző biológiai rendszer hozza létre (mindkettő pillanatnyi állapotától függően). A beszédprodukció akusztikuma tehát a megvalósulási időtengelyhez kapcsolódik (gondoljunk arra, hogy sokórás hangfelvétel képezi a beszédkorpuszt, ennek az időtengelyén kell keresni). Ha ebből az időtengelyből viszonylag hosszú szakaszokat ragadunk ki (szó, szókapcsolat), akkor a hangzás hordozni fogja a beszélő hangszínezetét. A hosszabb beszédrészlet jobban megközelíti az egyszeri és egyedi ejtésre jellemző hangzást, mint a rövid. A korpuszalapú szintézisnél az elérhető minőségjavulást a válogatási tér, azaz a korpusz nagysága befolyásolja. Az ideális az lenne, ha a korpusz tartalmazná a szinteti-

zálandó mondatok mindegyikét. Ezt nem lehet megvalósítani, ezért a fejlesztők arra törekszenek, hogy a műszaki lehetőségekhez optimalizálják a beszédkorpusz nagyságát és tartalmát.

A szintézis alapegysége és a prozódia megvalósításának egy lehetséges modellje

A korábbiakból már láthattuk, hogy a beszédszintézisnél a beszédjelet minden esetben valamilyen jól definiált részegységekből rakjuk össze. A beszédkorpusz készítésénél is az első alapkérdés, hogy mi legyen ez a jól definiált alapelem. Mivel nyitott szerkezetű beszédkorpuszról van szó, elvileg bármilyen hosszúságú hangelem lehet (hang, diád, triád, szótag, szó, mondat stb.) a jelölt, azt a tervező határozza meg. A választástól függ a további korpusztervezés menete. Szónál hosszabb alapelemet nem célszerű használni, mert nehéz a komplett lefedést biztosítani (hogy minden lehetséges hangsort elő lehessen állítani). A szónál rövidebb elemeket pedig már kipróbáltuk a korábban fejlesztett rendszerekben, azok előnyeit, hátrányait ismerjük. Marad a szó. A szó választása természetes hangminőséget jelenthet, hiszen olyan nyelvi egységet határoz meg (legyen rövid vagy hosszú), amelynek észlelésére percepciós rendszerünk is fel van készülve. A szó kiejtésében automatikusan benne van a szegmentális és a pillanatnyi szupraszegmentális szerkezet is (még ha a szavak összefolynak is a folyamatos beszédben). A jó hangminőség még fokozható azzal is, ha a szintetizálandó mondatot sikerül szókapcsolatokból felépíteni (találunk olyan, több szóból álló mondatrészeket a beszédkorpuszban, amelyeket közvetlenül felhasználhatunk a szintetizálandó mondat felépítéséhez). Ilyenkor még hosszabb hangsorrészre vonatkozik az előbbi állítás. Mindezekből adódóan a most bemutatott rendszerben a keresés alapelemének a szó nagyságú egységet választottuk. Emellett azonban gondoskodni kell egy alacsonyabb szintű építőelemről is, arra az esetre, ha nem találunk megfelelő szót a keresés során. Ez a tartalék elem pedig a hang (hangonként is össze lehet állítani a szükséges szót). Tehát a hangok is építőelemeknek számítanak a rendszerben, csak igen ritkán használjuk őket. Mindezekből következik, hogy a beszédkorpuszban minden mondat hullámformájába be kell jelölni minden hang határát és minden szóhatárt (lásd később). Vegyünk egy példát. Egy 10 szavas mondathoz például a legrosszabb esetben 10 szót kell keresni a beszédkorpuszban. Emlékeztetünk arra, hogy a formánsszintetizátoroknál az alapegység a 10 ms-onként megadott adatsorozat volt, a hullámforma-szintézisnél pedig a diád, illetve a triád. Mindkét esetben sokkal több elemből kellett az előbbi mondatot összeállítani, mint a korpuszszintézisnél. A beszédkorpusz készítésének második alapkérdése a prozódiai modell. A kiindulási gondolat fontos alapeleme, hogy a korpuszalapú szintézisnél a prozódia előállítására lehetőleg ne alkalmazzunk mesterséges dallam-, intenzitás- és időtartam-módosítást (mivel az torzítja a hangot), hanem keressük a beszédkorpuszban azt az optimális szót, amelyik prozódiai szempontból is illik a szintetizálandó mondat adott részéhez. A felolvasandó szöveganyag mondatait (mondatkorpusz) tehát úgy kell összeál-

lítani, hogy ugyanabból a szóból prozódiailag többfajta is legyen a mondatokban. A kérdés az, hogy az előbbi általános megállapítást hogyan lehet modellezni. Hányfajta szóvariánst kell a szövegbe ágyazva tárolnunk egy-egy szóból? A prozódia modellezésében alapegységnek a mondatot tekintjük. A modell szorosan összefügg a szintetizálandó szöveg szerkezetével, jelen esetben csak kijelentő mondatokat modellezünk. A kijelentő mondat prozódiai szerkezete jól körülhatárolható, ismert egységekből áll. Ezeket az egységeket a mondaton belüli hely szerinti pozicionálással (hol van a szó a mondatban), valamint a központozással (vesszők, gondolatjelek stb.) kapcsolatba lehet hozni. Ez a modell lényege. Ugyanazt a modellt alkalmazzuk a mondatkorpuszban, a beszédkorpuszban és a szintetizálandó mondatban is. A mondat- és beszédkorpuszban felcímkézzük a mondatokat a modell szerint, a szintetizálandó mondatra pedig alkalmazzuk a modellt. Így prozódiai vonatkozásban is ki lehet alakítani keresési kritériumokat.

A mondatkorpusz kialakítása időjárás-jelentések automatikus felolvasásához

A mondatkorpusz kialakításához meg kell határozni, hogy az milyen mondatokat tartalmazzon. Belátható, hogy ehhez olyan nyers szöveganyagot kell összeállítani, amelyiknek a szóállománya lefedi a majdan szintetizálandó időjárás-jelentéses mondatok szóállományát. Első lépésben ennek a mondatállománynak a gyűjtését végeztük el. Egy éven keresztül gyűjtöttünk (saját, automatikus szoftverrel) magyar időjárásjelentés-szövegeket 20 különböző weboldalról. Az eredmény 56 000 mondat, bennük 493 000 szó és 43 000 szám. A teljes szöveg 5 200 különböző szóalakot tartalmazott (a szóalak akkor különbözik, ha a szó betűkarakteres formájában legalább egy karakter különbözik). A statisztikai analízis azt mutatta, hogy a leggyakoribb 500 szóalak lefedte a mondatállomány 92%-át, 2 300 szóalak pedig a 99%-át (prozódiai szempontok nélkül). Ez a kis szám abból ered, hogy a témakört limitáltuk az időjárásos mondatokra. (Megjegyezzük, hogy korlátozás nélküli szöveg hasonló fedéséhez a tárolt mondatállománynak mintegy 70 000 szóalakot kellene tartalmaznia.) A második lépésben alakítottuk ki az 56 000-es mondatállományból a későbbiekben felolvasásra kerülő mondatkorpuszt (5 260 mondat), amely tartalmazta az 5 200 szóalakot és azok prozódiai variánsait (összesen 82 000 szó). Ez a mondatkorpusz képezi a beszédszintetizátor elsődleges (szóalapú) keresési terét.

A beszédkorpusz elkészítése

A beszédkorpusz a mondatkorpusz felolvasásából jött létre. A hivatásos női bemondó 4 héten át, heti 2-3 alkalommal, naponta 4-5 órát olvasva mondta fel a mondatkorpusz 5 260 mondatát. Az eredmény 11 órás folyamatos beszédanyag – ez képezi a beszédkorpuszt. A folyamatos beszédanyagot mondatokra daraboltuk, minden mondat kapott egy azonosítót. Ezután a mondatkorpusz szöveges formáját (minden mondatát) manuálisan össze kellett vetni a beszédkorpusz tartalmával, kijavítottuk az esetleges felolvasási hibákat (a bemondó néha tudat

alatt átformálta az írott szöveget, ilyenkor a szöveget a felolvasott formához igazítottuk, továbbá a felolvasott számokat és rövidítéseket is szövegesen ki kellett fejteni (például *4-6 °C = négy, hat Celsius-fok*). A cél az volt, hogy teljes szinkronba hozzuk a hangot és annak szöveges formáját. A következő lépésben minden mondat hullámformáját elláttuk szóhatárokkal, hanghatárokkal és a hangok jeleivel (fonetikusan átírtuk a szövegeket). Ezt a BME TMIT automatikus beszédfelismerő szoftverének (Mihajlik et al. 2002) támogatásával végeztük. Az automatikusan átírt és címkézett anyagot – a szó- és hanghatárokat – félautomata módszerrel ellenőriztük. Itt felhasználtuk például a magyar beszédre kidolgozott időtartammodellt (Olaszy 2006), minden hangra jósoltunk egy időtartamot, és összehasonlítottuk a bejelölt értékkel. Nagy eltérés esetén manuálisan megkerestük a hiba helyét, és korrigáltuk a rossz jelzést. A szóhatár jelölése sok esetben nem végezhető el egyértelműen. Ezért külön jelet használtunk a szó kezdetének jelzésére (<), és külön jelet a befejezésére (>). Hangösszeolvadás esetén a szóhatár nem jelölhető ki egyértelműen a fonetikus változatban (például: a *Balatonnál legalább* átírása: *<balatoná<l>egaláb>*). Ilyenkor az „előző szó vége" jelzést megelőzi a „következő szó eleje" jelzés, és ezt az algoritmus értelmezi. Mindezekből látható, hogy a korpuszalapú szintézis beszédkorpuszának végleges formára hozása többszintű, kitartó munkát igényel.

Elemkiválasztás

Az elemkiválasztás algoritmusa a korpuszalapú szintézis legproblematikusabb eleme. A bemeneten rendelkezésre áll a szintetizálandó mondat szövege a magyar helyesírás szabályainak megfelelően leírva. Ezzel szemben áll a mondatkorpusz mint keresési tér a szöveges és fonemikus formákkal. Az elemkiválasztó veszi a szintetizálandó mondat első szavát, és elkezd keresni a mondatkorpusz szövegében. A legtöbb esetben sok jelöltet fog találni a 82 000 szóból. Az eljárásnak ezekből a jelöltekből kell kiválasztania a legmegfelelőbbet, amihez felhasználjuk a fonetikus átiratot is. A feldolgozás szavanként folytatódik, míg a teljes mondat minden szavára nem kaptunk sok-sok jelöltet. Ezután az elemkiválasztó kiválasztja a jelöltek halmazából a legjobban illeszkedőnek vélt szavakat, megkeresi a nekik megfelelőket a beszédkorpuszban, onnan kiveszi azokat, és összefűzi őket a megszólaltatáshoz. Hogyan működik az elemkiválasztó? Kétféle költségfüggvény (KLTS–1 és KLTS–2) határozza meg, hogy az éppen vizsgált szó mennyire felel meg a kívánt követelménynek. A költségfüggvények értéke alapján történik a végleges költségszint kiszámítása. Ha ennek értéke nulla, akkor az elem 100%-osan a legoptimálisabb, ha magas szám, akkor az elem nem illeszthető, el kell vetni. A KLTS–1 határozza meg, hogy a hangsor és a hangkörnyezet szempontjából, azaz szegmentálisan mennyire felel meg a szó. Itt a szó hangsorát kell azonosítani, továbbá azt, hogy mennyire illeszkedik az őt követő, illetve megelőző szóhoz. Mivel a kiejtés folyamatos, a szóhatárokon törekedni kell arra, hogy a spektrális illeszkedés (pl. formánsmenet) is folyamatos legyen. A szó első és utolsó hangjának illeszkedését vizsgáljuk, és az illeszkedés

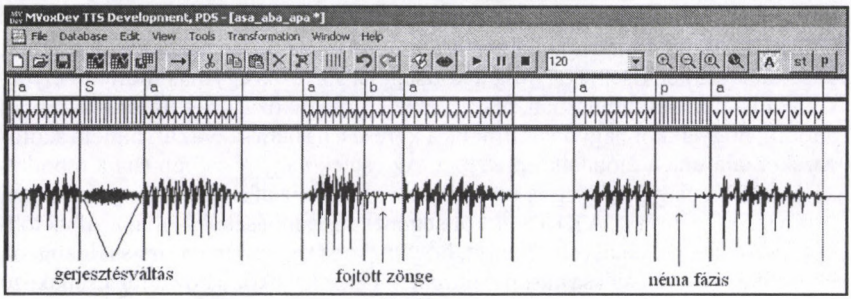
költségét több szempont alapján számítjuk ki. Magas költségű például, ha a szóhatáron magánhangzók találkoznak (*dunántúli áramlások*). Az ilyen szavak magas költséget képviselnek. Nulla a költség, ha a két szó egymás mellett helyezkedik el a mondatkorpuszban, hiszen ekkor a csatlakozásuk is optimális. Ebből adódik, hogy akkor nagyon optimális a keresés, ha nem szavakat, hanem szófüzéreket találunk a mondatkorpuszban. Az esetek nagy részében (ha a mondatkorpusz elég nagy) ez meg is valósul, így a szintetizált szöveg hangzása közel lesz a természeteshez. A KLTS–1 költségének kiszámításához felhasználjuk többek között azt a kutatási eredményt, hogy azonos képzési helyű mássalhangzók akusztikai megvalósulása hasonló átmeneti fázisokat okoz a hozzájuk csatlakozó magánhangzóban (Olaszy 2003), továbbá a mássalhangzók képzési módjának osztályozását és a gerjesztés fajtáját (zöngésség-zöngétlenség). A mássalhangzók képzési helyéből adódó azonos akusztikai vetületeket az 1. táblázat mutatja be.

1. táblázat: A magyar mássalhangzók képzési hely és mód szerinti csoportosítása
(Az egy sorban lévő mássalhangzók hasonló akusztikai vetületet hoznak létre a hozzájuk csatlakozó magánhangzóban.)

Képzésmód / Képzéshely	Zárhangok								Zár-rés hangok				Réshangok							Nazálisok					
	b	p	d	t	gy	ty	g	k	c	dz	cs	dzs	v	f	z	sz	zs	s	h	m	n	ny	j	l	r
Két ajak	☒	☒																		☒					
Ajak-fog													☒	☒											
Fog-fogmeder			☒	☒					☒	☒					☒	☒					☒			☒	☒
Fogmeder											☒	☒					☒	☒							
Kemény szájpad					☒	☒																☒	☒		
Lágy szájpad							☒	☒																	
Gége																			☒						

Melyek tehát az optimális összefűzési pontok szóhatárok esetében? Ezt elsősorban az 1. táblázat szerinti 7 artikulációs vetületi sor, illetve a beszédjel nagysága dönti el. Nem célszerű összeillesztést végezni nagy energiájú jelszakaszban (például magánhangzóban), a kis energiájú helyeket kell előnyben részesíteni. Szabad illeszteni a hangsor minden olyan pontján, ahol gerjesztésváltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitásminimum keletkezik), továbbá a hangok belsejében lévő néma fázisokban, illetve zöngeszakaszokban (3. ábra).

Ha tehát az akusztikai vetület ugyanaz, és például gerjesztésváltás van a két szó határán, akkor az összeillesztési költség értéke kicsi lesz, hiszen a spektrális folytonosság biztosított, és az illesztésnél kicsi az energia. Hasonló elvek alapján kialakítható az a fonetikai szabályrendszer, amellyel ki lehet jelölni a vágás konkrét helyét (a vágási pontot) az összeillesztendő szavakon belül. Erre mutat példát a 2. táblázat.



3. ábra

Az optimális illesztési pontok bemutatása az *asa, aba, apa* hangkapcsolódásoknál

(A függőleges vonalak a hanghatárokat jelzik, a zöngés hangok periódushatárait a „v" jelzésű vonalmarkerek jelzik a hullámforma felett.)

2. táblázat: Példa a fonetikai szabályrendszerből az alacsony költségű vágási pontok kijelölésére

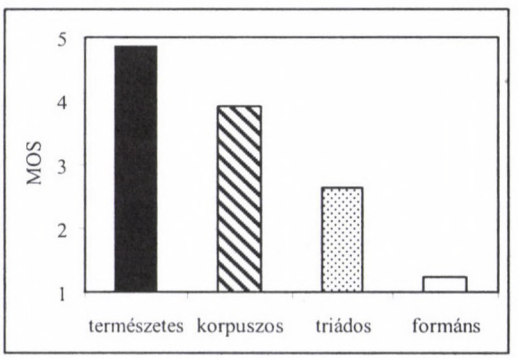
(A csatlakozó második hangot a következő szimbólumok jelölik: C = bármely mássalhangzó; V = bármely magánhangzó; C_1 = *p, t, k, ty, h, f, s, sz, c, cs;* C_2 = *v, j, l, r;* C_3 = *m, n, ny.* A hangokat a betűjelükkel adjuk meg.)

A megelőző hang	**A következő, kapcsolódó hang**	**A vágási pont kijelölésének szabálya**	**Szöveges példa (a csatlakozó hangok félkövér kiemeléssel)**
V	*a)* V	*a)* A hanghatár be van jelölve, ennek ellenére nem célszerű elvágni a hanghatárnál, hanem megfelelő vágási pontot kell keresni visszafelé vagy előre a hangsorban.	*éjszakai esőzésre*
	b) C	*b)* A hanghatárnál kell vágni.	*nyári záporok*
a) b, d, g, gy	*a)* V, C_2, C_3	*a)* A hanghatárnál kell vágni.	*vad vihar, nagy meleg*
b) b, d, g, gy	*b)* önmagával csatlakozik	*b)* A hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne.	*vad dörrenés*
c) p, t, k, ty	*c)* C_1, kivéve *d)*	*c)* A hanghatáránál kell vágni.	*szép sereg*
	d) önmagával csatlakozik	*d)* A hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne.	*sok kis*
	e) V, C_2, C_3	*e)* A hanghatárnál kell vágni.	*szép felhők*
m	*a)* C, kivéve *m*	*a)* A hanghatárnál kell vágni.	*nem volt*
	b) önmagával	*b)* A hang 70%-ánál kell elvágni.	*nem marad*

A KLTS–2 határozza meg, hogy a hangsor és hangkörnyezet szempontjából kiválasztott szó, szófüzér mennyire felel meg a prozódiai követelményeknek. Itt szempont az is, hogy a kiválasztott szó a mondatkorpusz ugyanazon mondatában szerepel-e, mint az előző. Ha igen, akkor a költséget ez a tény is csökkenti. A prozódiai költség meghatározásánál – az időtengelyi pozíción felül – felhasználjuk az F_0 értékének a változását is. Ha nagy F_0-ugrás van a két szó között, akkor a költség magas lesz, tehát a két elem nem illeszthető össze.

A működő rendszer minősítése

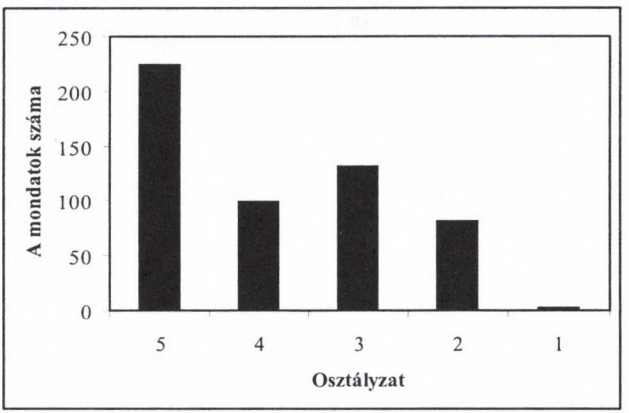
A kísérleti, időjárás-jelentést felolvasó, korpuszalapú beszédszintetizáló rendszer minőségét percepciós teszttel vizsgáltuk. Három magyar rendszert hasonlítottunk össze, a Multivox formánsszintetizátort, a Profivox diád-triád elemösszefűzéses rendszert és a korpuszos felolvasót. A tesztelőktől a hangminőségre vonatkozó ítéletet kértünk 5 pontos skála szerint: nagyon jó (5), jó (4), átlagos (3), gyenge (2), elfogadhatatlan (1). A teszt anyagát a webről kiválasztott időjárás-jelentés 10 mondata alkotta. Ezeket állítottuk elő a fenti rendszerekkel, valamint felolvastattuk őket a beszédkorpusz eredeti bemondójával is. A tesztelőknek tehát összesen 40 mondatot kellett meghallgatni véletlen sorrendben. Minden mondatra egy ítélet született. A tesztet egy interaktív honlap segítségével bonyolítottuk le. A mondatokat 221 személy (egyetemi hallgatók, 185 férfi és 36 nő) hallgatta meg. A teszt elején ismerkedésképpen minden mondattípusból egy-egy mondatot meghallgathattak. A tesztben a mondatokat csak egyszer hallották, ismétlésre nem volt mód. A meghallgatások csendes, otthoni környezetben, átlagos (nem professzionális) hangszórókon, illetve fejhallgatókon történtek (a tesztelők a meghallgatás körülményeire vonatkozóan is kitöltöttek adatokat a teszt előtt). A teszt eredményeit a 4. ábrán mutatjuk be. A tesztelés eredményét a MOS (Mean Opinion Score) értéke fejezi ki.



4. ábra

A szubjektív minősítés átlagai az egyes szintézistechnológiákra

Az értékelésekből látható, hogy a korpuszalapú szintézis hangminősége magasan kiemelkedik a másik két technológiával szemben. A szó-, szófüzéralapú összeillesztéssel tehát átléptünk a percepciós megítélésben egy olyan határt, amelyet a hullámforma-összefűzéses rendszereknél még nem tudtunk elérni, annak ellenére, hogy ott is emberi beszéd részleteit fűztük össze. Feltételezzük, hogy a szó képviseli azt a mondatépítő elemet, amelynek szintjén már elégséges egyéni hangjellegzetesség van jelen a hullámformában, hogy a hallgató a beszélő hangszínezetét, egyéni stílusát is felismerje, és ennek folytán értékítéletével megközelítse a jó (4) szintet. Természetesen a korpuszalapon szintetizált mondatokban is vannak egyenetlenségek a hullámforma folytonosságát illetően (dallamugrások, hangszínezet-változások stb.), de úgy tűnik, ha kevés van ezekből, akkor az összegzett ítéletek meghozásakor ezeket a percepciós mechanizmusunk ugyanúgy tűri, feldolgozza, mint az olvasásnál a felolvasási mechanizmusunk a betűkimaradásokat, betűhibákat. Elvégeztünk egy másik percepciós értékelést is, amelyben a rendszer belső működését, a mondat-összeállítás hatásosságát vizsgáltuk. Kíváncsiak voltunk, hogy nagyszámú időjárás-jelentéses mondat előállítása esetén hogyan alakul a mondatok hangminőségi eloszlása. A teszthez 540 mondatot szintetizáltunk (olyanokat, amelyek nem szerepelnek a rendszer korpuszában), és ezek hangminőségét értékeltettük 4 fővel (30–60 éves férfiak). A mondatokat a tesztelők fejhallgatón hallgatták. Egy tesztelő egy alkalommal 135 mondatot hallgatott meg. Egy mondatot többször is meghallgathattak. Az értékelést az előző teszthez hasonlóan egy 1–5 osztályzatú skálán kértük, a következők szerint: 5 = nagyon jól érthető, mintha bemondó olvasta volna; 4 = jól érthető; 3 = közepesen érthető; 2 = nehezen érthető, 1 = nem érthető. Az eredményeket az 5. ábrán mutatjuk be.



5. ábra
A mondatok hangminőségi eloszlása 540 időjárás-jelentéses mondaton mérve

A tesztelők 225 mondatra adtak nagyon jó értékelést, 99-re jót és 132-re közepesen érthetőt. Az összes mondatra számolva ez 90%-os lefedést jelent. Mindössze 81 mondatnál volt gondjuk a minőséggel, itt nehezen érthető ítéletet adtak. 3 mondat esetében nem lehetett megérteni a mondat tartalmát. Ezek az eredmények azt mutatják, hogy a jelen korpuszalapú szintetizáló rendszernél az esetek többségében igen jó hangminőség érhető el. Előfordulhatnak azonban igen kis számban olyan generált mondatok, amelyeknél a költségfüggvény gyakorlatilag abszolút rossz döntéseket hoz. Ilyen mondat volt például a következő: *Vasárnap számottevő eső csak Északkelet-Magyarországon esik, hajnalban és délelőtt.* Mindezekből az is látszik, hogy a meggyőzően jó MOS eredmények mellett számolni kell azzal is, hogy rossz válogatás esetén a korpuszalapú szintetizálás is adhat igen rossz minőségű beszédet. A jelen kísérleti rendszerben ez csak 0,6%-ban fordult elő. A hibák eredhetnek abból, hogy a mondatkorpusz és a neki teljes mértékben megfeleltetett beszédkorpusz között nem teljes az egyezés, azaz hanghatár, illetve átírási hibák vannak. Továbbá a prozódiai modell döntései is hibásak lehetnek. A hibák kijavítására a konkrét hibakeresésen és hibajavításon túl az egyik megoldás az lehet, hogy az eredeti bemondóval felolvastatjuk a rossznak ítélt mondatokat, és hozzákapcsoljuk a korpuszhoz, kihasználva, hogy ez egy nyitott beszédkorpusz.

Összefoglalás

A tanulmányban bemutattuk az első magyar nyelvű, korpuszos technológián alapuló szövegszintetizáló kísérleti változatát. A rendszerben alkalmazott módszer gyökeresen eltér a korábbi beszédszintézis-technológiákra jellemző módszerektől. A rendszert kötött témakörre készítettük el, időjárás-jelentéseket tud felolvasni. Az eddigi teszteredmények azt mutatják, hogy ez a legújabb technológia igen jó minőségű beszédet biztosíthat: az esetek nagy részében nem lehet megkülönböztetni, hogy szintetizált-e a mondat, vagy egy bemondó olvasta-e fel.

Irodalom

Kawai, Hisashi – Toda, Tomoki – Ni, Jinfu – Minoru, Tsuzaki – Tokuda, Keiichi 2004. Ximera: a new TTS from ATR based on corpus-based technologies. In: *Proceedings of the 5th ISCA Speech Synthesis Workshop (SSW5).* Pittsburg, 642–645.

Mihajlik, Péter – Révész, Tibor – Tatai, Péter 2002. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica 49/3–4.* 407–425.

Nagy András – Pesti Péter – Németh Géza – Bőhm Tamás 2005. Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései. *Híradástechnika 2005. január,* 18–24.

Olaszy, Gábor – Gordos, Géza – Németh, Géza 1992. The MULTIVOX multilingual text-to-speech converter. In Bailly, G. – Benoit, C. – Sawallis, T. (eds.): *Talking machines: Theories, models and applications.* Elsevier, Amsterdam, 385–411.

Olaszy Gábor 1999. Beszédadatbázisok készítése gépi beszédelőállításhoz. *Beszédkutatás'99.* 68–89.

Olaszy Gábor 2003. Az artikuláció akusztikai vetülete – a hangsebészet elmélete és gyakorlata. In Hunyadi László (szerk.): *Kísérleti fonetika, laboratóriumi fonológia a gyakorlatban.* Debreceni Egyetem Kossuth Egyetemi Kiadója, Debrecen, 241–254.

Olaszy Gábor 2006. *Hangidőtartamok és időszerkezeti elemek a magyar beszédben.* Nyelvtudományi értekezések 155. Akadémiai Kiadó, Budapest.

Olaszy, Gábor – Németh, Géza – Olaszi, Péter – Kiss, Géza – Zainkó, Csaba – Gordos, Géza 2000. Profivox – a Hungarian TTS system for telecommunications applications. *International Journal of Speech Technology 3/3–4.* 201–215.

Schweitzer, Antje – Braunschweiler, Norbert – Klankert, Tanja – Möbius, Bernd – Sauberlich, Bettina 2003. Restricted unlimited domain synthesis. *Proceedings of Eurospeech 2003.* Geneve, 1321–1324.

A szerzők köszönetüket fejezik ki Pesti Péternek, aki a rendszer kódját programozta, és Mihajlik Péternek, aki rendelkezésükre bocsátotta a jelen munkában használt beszédfelismerési eszközöket, és segítséget nyújtott azok használatában.
Ez a kutatás az NKFP 2/034/2004 szerződése alapján támogatásban részesült.