Generációváltás a beszédszintézisben

FÉK MÁRK, PESTI PÉTER, NÉMETH GÉZA, ZAINKÓ CSABA

{fek,nemeth, zainko}@tmit.bme.hu, pesti@alpha.tmit.bme.hu
BME Távközlési és Médiainformatikai Tanszék

Kulcsszavak: formánsszintetizátor, elemösszefűzés, elemkiválasztás, korpusz alapú beszédszintézis, szubjektív minősítés

A cikk áttekinti a beszédszintézis rendszerek három generációjának fejlődését. Bemutatjuk a BME-TMIT-en fejlesztett magyar nyelvű kísérleti korpusz alapú, elemkiválasztásos beszédszintetizátor felépítését. Részletesen ismertetjük a hang és szóhatárok automatikus jelölésének módszerét. Feltárjuk a prozódia megvalósításának lehetséges módszereit egy korpusz alapú, elemkiválasztásos beszédszintetizátorban. Ismertetjük az elemkiválasztás működését a fejlesztés alatt álló rendszerben. A kísérleti rendszer beszédminőségét összehasonlítjuk a korábbi magyar nyelvű beszédszintézis rendszerek minőségével.

1. Bevezetés

A beszédszintézis rendszerek célja a bemeneti információ beszéddé alakítása. A bemenet legtöbbször egy felolvasandó szöveg (ekkor szövegfelolvasó (text-to-speech) rendszerről van szó), de lehet valamilyen adat (számlaegyenleg, járatinformáció, időjárási adatok stb.). Az adatjellegű bemenetet kezelő, információ-felolvasó (concept-to-speech) rendszerek első lépésként általában szöveggé alakítják a bemenetet. Egy szövegfelolvasó rendszer két alapvető részből épül fel. Az első rész a bemeneti szöveget alakítja szimbolikus információvá, a második a szimbolikus információt alakítja a beszédjelet leíró hullámformává (általában valamilyen hangfájlt állít elő). A közbenső szimbolikus információ általában a szöveg tartalmát megadó fonéma sorozatból (egy fonéma egy beszédhangot jelöl) és a beszéd prozódiai jellemzőit (hanglejtés, hangsúlyok, ritmika) leíró információkból áll össze.

A cikk első részében áttekintjük a beszédszintézis rendszerek három generációjának fejlődését. A második rész a BME-TMIT-en fejlesztett kísérleti korpusz alapú, elemkiválasztásos beszédszintetizátor működését mutatja be. A befejező részben a három bemutatott generációnak megfelelő három magyar nyelvű beszédszintetizátort hasonlítjuk össze egy szubjektív minősítési teszt segítségével.

2. Formánsszintézis

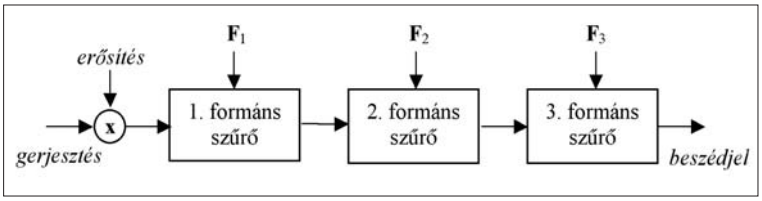
A formánsszintézis volt az első olyan beszédszintézis technológia, melynek segítségével egy szöveget automatikusan folyamatos és jól érthető beszéddé lehetett alakítani. Az elnevezés a szövegfelolvasó rendszerben alkalmazott hullámforma előállítás módszerére utal, ami egy gerjesztett szűrőrendszer kimeneteként állítja elő a beszédjelet. A formáns szintetizátor egy lehetséges megvalósítását az *1. ábra* mutatja.

A formánsszintetizátor az emberi beszédkeltést modellezi. Számítástechnikai erőforrásigénye kicsi. A gerjesztés a hangszalagok által keltett jelnek feleltethető meg: zöngés hangok esetén kvázi-periodikus, zöngétlen hangok esetén zajszerű. A gerjesztés alakja a hangszínezetet befolyásolja. Egy formáns szűrő a megadott formáns frekvencia környezetében erősíti a jelet, ezzel modellezve a garat, a gége és a szájüreg által alkotott rezonátor-rendszer erősítéseit. A formáns frekvenciák a zöngés hangokra jellemzőek, de zöngétlen hangok leírására is használhatóak.

Az első három formánsfrekvencia jól leír egy zöngés, kitartott beszédhangot. A szintézis folyamán beállítandó formánsfrekvenciákat a szövegből előállított fonémasorozat határozza meg. A formánsfrekvenciák az artikulációs mozgások függvényében szintén változnak, a spektrális szerkezet folymatosan módosul periódusról periódusra.

A beszédhangokon belül megkülönböztetünk stabil szakaszt (általában a hang közepe) és hangátmeneti részt (a hangnak a kezdete, ami az előző hanghoz kapcsolódik, illetve a vége, amelyik a következőhöz fűződik hozzá). A stabil szakaszok közötti hangátmeneteknél a formánsok mozgatását a bemeneti fonémasorozat és hangidőtartamok alapján szabályrendszer vezérli. A szabályok komplexitási szintje meghatározza a szintetizátor hangzását. Egyszerű szabályokkal csak gépies hangzás érhető el. Az újabb formánsszintetizátorokban a hangátmenetek paramétereit természetes hangátmenetekből nyerik ki. Ez valamivel jobb hangzást eredményez.

1. ábra
Soros elrendezésű formánsszintetizátor blokkvázlata



A zöngés gerjesztés alapfrekvenciájának a vezérlését az úgynevezett prozódiai modul végzi. Ennek bemenete a szövegből előállított fonémasorozat és a szimbolikus prozódia. Ez utóbbi általában a mondatok modalitását (kijelentő/kérdő) és a hangsúlyok mondaton belüli helyét és típusát adja meg. A modul kimenetei a fizikai prozódiai jellemzők, azaz az alapfrekvencia-menet, a fonémáknak megfelelő hangok időtartamai, illetve az intenzitásmenet. Ezek minősége szintén nagyban befolyásolja az előállított beszéd hangzásának természetességét.

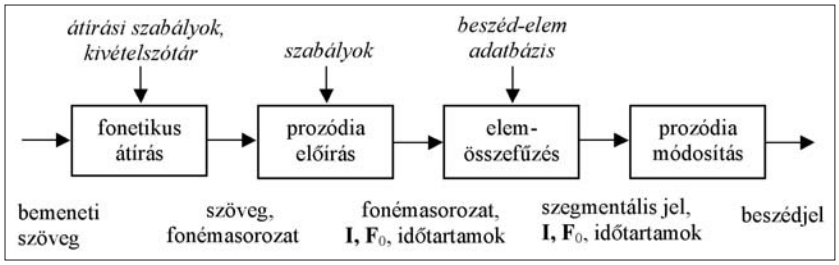
A formánsszintetizátor megfelelő vezérlésével jó minőségű, természetes beszéd állítható elő. Ugyanakkor ilyen vezérlő információt csak természetes beszédjelből, félautomatikus módszerek segítségével sikerült mindezidáig előállítani. A bemeneti szövegből kiinduló és egy szabályhalmaz segítségével előállított vezérlő információ érthető, de erősen gépies hangzású beszéd előállítását teszi csak lehetővé. Ezen minőségi korlát miatt a formánsszintetizátorok csak kis erőforrásigényű gyakorlati alkalmazásokban fordulnak elő. A módszert kutatási célra ma is használják, elsősorban azért, mert a beszédjel gerjesztése – ellentétben az újabb beszédszintézis-technológiákkal – könnyen módosítható, és így annak hatása külön vizsgálható. Másik előnye a kis tárkapacitás-, és az alacsony számításigény.

A BME-TMIT által kifejlesztett Multivox 12 nyelven beszélő formánsszintetizátor magyar nyelvű változata [1] ingyenesen hozzáférhető.

3. Elemösszefűzésen alapuló beszédszintézis

Az elemösszefűzésen alapuló beszédszintézis esetében a szövegfelolvasó rendszer két fő egysége közül (szövegfeldolgozó és hullámforma-előállító) a hullámforma-előállító rész jelent újdonságot *(2. ábra)*. Ahelyett, hogy minden egyes beszédhangra és beszédhangátmenetre előírnánk a formánsfrekvenciák és a gerjesztés alakjának időbeni menetét, természetes beszédből kivágott hullámforma elemeket fűzünk össze. Ugyanakkor a formánsszintetizátorban alkalmazott prozódiai modul általában változatlanul megmarad, és előírja az előállítandó hullámforma alapfrekvencia- és intenzitásmenetét, illetve az egyes hangok időtartamait. Emellett a szöveget fonémasorozattá, illetve szimbolikus prozódiává alakító rész is változatlan marad.

2. ábra Elemösszefűzésen alapuló beszédszintetizátor egy lehetséges felépítése



A technológia egyik alapvető kérdése, hogy melyek legyenek azok a hullámforma elemek, amelyek összefűzésével előáll a gépi beszéd. Itt több szempontot kell figyelembe venni. Egyrészt teljes fedést kell biztosítani, azaz az adott nyelv tetszőleges hangsorozatát elő kell tudni állítani. Másrészt az előállított beszédnek minél természetesebben kell szólnia. Korlátot jelenthet a hullámforma elemek száma, illetve azok együttes mérete. Az előbbi az elemek közötti keresés idejét növeli, az utóbbi pedig a szükséges tárterületet.

Alapötletként felmerülhet a fonémáknak megfelelő hangok, mint elemek használata. Ez teljes fedést biztosít, és kevés elemmel megoldható (a magyarra 38 fonémából már előállítható jó minőségű beszédszintetizátor). A fonémáknak megfelelő hangok összefűzésével előálló jel azonban nem hangzik folytonosnak, azaz az összefűzés után előálló hang minősége gyenge. A problémát az okozza, hogy a beszédjelben az egyes hangok folyamatosan mennek át egymásba, és általában csak a hang közepén tekinthetők állandósultnak. Ezt úgy is értelmezhetjük, hogy a hang elejének alakulása az őt megelőző, a vége pedig az őt követő hangtól függ [2]. A megoldás a környezetfüggő hangok használata lehetne, ahol minden egyes hangot minden lehetséges hangkörnyezetének megfelelő változatban tárolunk. Ehhez viszont nagyon sok elemet kellene tárolni, felső becslésként 38^3=54872 elemre lenne szükség. Ez azért felső becslés, mert a nyelvben nem valósul meg minden lehetséges hanghármas. Az ilyen környezetfüggő elemeket *triád*nak (angolul triphone-nak) nevezik. Megjegyezzük, hogy az elemszámot nemcsak a technológia korlátozza, hanem a rendelkezésre álló emberi erőforrások is. Az egyes elemeket ugyanis egyenként kell hangstúdióban felvenni, és félautomatikus módszerekkel feldolgozni, előkészíteni a szintézisre.

A gyakorlatban bevált kompromisszumos megoldás a két egymás utáni félhang együtteseként előálló diád (angolul diphone) alkalmazása. Ez kihasználja azt a közelítő feltételezést, hogy egy hang első fele nem függ az azt követő hangtól, második fele pedig az azt megelőzőtől. A magyar nyelvű szintézishez szükséges diád-elemek száma 38^2=1444. Megjegyezzük, hogy a kezdeti, csak diád-elemeket tartalmazó rendszereket az idők folyamán triád-elemekkel bővítették, ami némi minőség javulást eredményezett. Ezek a triád-elemek az adott hangot megelőző hang közepén kezdődtek és a hangot követő hang közepéig tartanak, azaz két hangnyi hosszúak. Ennek előnye, hogy a rendszer kevesebb vágási pontot tartalmaz, a hosszabb építőelemek miatt pedig folytonosabb lesz a beszéd hangzása. Hátránya a megnövekedett elemszám.

Hasonló utat járt be a BME-TMIT-en kifejlesztett Profivox magyar nyelvű beszédszintetizátor [3], amelynek legújabb változata az 1444 diád mellett, 6000 triád-elemet is tartalmaz. A két rendszer hangzását összehasonlító egyszerű minősítési teszt megtalálható [4]-ben.

A diád, illetve triád-elemek összefűzése után gondoskodni kell arról, hogy az előálló beszédjel kövesse a formánsszintézisnél alkalmazott módhoz hasonlóan előírt prozódiai jellemzőket (alapfrekvencia- és intenzitásmenet, hangidőtartamok). A formánsszintézissel ellentétben itt nem áll eleve rendelkezésre egy parametrikus modell, így azt vagy létre kell hozni, vagy valamilyen időtartománybeli manipuláció segítségével kell a jelet módosítani. Mindkét esetben szükséges a jel alapfrekvencia-menténének pontos ismerete. Az egyes elemek esetleges modell-paraméterei, illetve alapfrekvencia-menete előre, nem valós időben is meghatározhatóak. Általánosan igaz, hogy minél nagyobb mértékben módosul a beszédjel, annál jobban romlik a minősége. Az intenzitás-menet viszonylag szabadon módosítható, ugyanakkor ennek van a legkisebb szerepe a prozódia alakításában. A hangidőtartamok akár 50-200%-os tartományban módosíthatóak, ami a gyakorlatban elegendő. Az előírt alapfrekvencia-menet megvalósítása a legkritikusabb, ugyanis az alapfrekvencia csak körülbelül 30%-kal módosítható még elfogadható minőségben.

A technológiához hozzátartozik a hangfelvételek rögzítése és feldolgozása. Az egyes diád-, illetve triád-elemeket mesterséges szavakba, úgynevezett logatomokba (például „aboka": a „b-o" diád-elemhez) ágyazva kell felolvasni. Ezeket a bemondónak jól artikulálva, monoton hanglejtéssel kell felolvasnia. A logatomok biztosítják, hogy minél kevésbé érvényesüljön a szomszédos hangok hatása a diád-elemre [5]. Ugyanakkor ezeket összefűzve mellékhatásként „túlartikulált" lesz a beszéd hangzása. A monoton hanglejtésre azért van szükség, hogy az alapfrekvencia módosításkor ne kelljen az esetlegesen túl alacsony, vagy túl magas frekvenciájú elemeken sokat módosítani. A felvett logatomok hullámformájában félautomatikus módszerekkel kell a hanghatárokat jelölni, és az egyes diád-, illetve triád-elemeket kivágni, ellenőrizni és hibás ejtés esetén javítani. Egy diád-elemeket tartalmazó hangadatbázis felvétele körülbelül 4 órát, míg feldolgozása egy emberhónapot vesz igénybe.

A diád-, illetve triád-elem összefűzésen alapuló beszédszintézis technológiát elterjedten alkalmazzák. Erre példa a Profivox magyar nyelvű beszédszintetizátor [3], amely e-mail-felolvasó, SMS-felolvasó, számszerinti tudakozó, illetve egyéb alkalmazásokban működik. A csak diád-elemeket tartalmazó változat kis memóriaigényének köszönhetően Symbian- és Windows Mobile-alapú mobiltelefonokon is képes futni.

4. Korpusz alapú, elemkiválasztásos beszédszintézis

Az elemösszefűzéses technológia továbbfejlesztéseként jött létre a korpusz alapú, elemkiválasztásos beszédszintézis *(3. ábra)*.

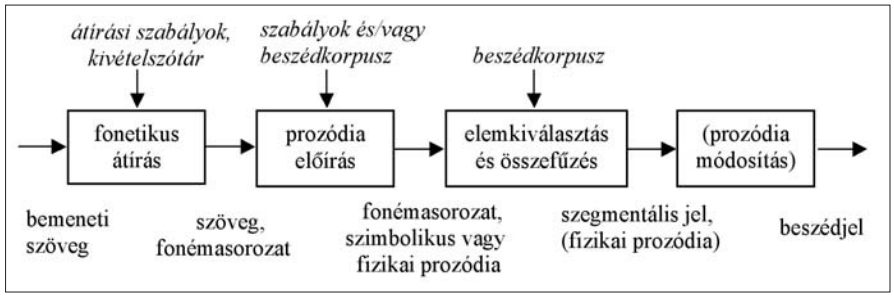
A technológia továbbra is beszédelemek összefűzésen alapul, ugyanakkor – mint a neve is mutatja – további két elvet vezet be. Az első, a korpusz alapú szintézis elve szerint a beszédszintetizátor hangadatbázisa nem a monoton prozódiájú logatomokból kivágott diád–, illetve triád-elemeket, hanem természetes hangzású teljes mondatokat tartalmaz. A mondatok egy nagyméretű szövegkorpuszból kerülnek kiválogatásra, és azok felolvasásával jön létre a több órányi beszédet tartalmazó adatbázis, azaz a beszédkorpusz.

Ellentétben a hagyományos elemösszefűzésen alapuló technológiával, a korpuszos adatbázisban egy adott hangsort tartalmazó beszédelem általában több példányban is előfordul. Ezen példányok prozódiai megvalósítása (alapfrekvencia-, és intenzitás-menet, hangidőtartamok, hangszínezet) eltérő. Másrészt a beszédkorpuszban egyszerre több különböző méretű elem is definiálható (például diád, triád, szótag, szó stb.). Ezen két ok miatt több lehetséges módon állítható elő egy adott szintetizált beszédszakasz, amelyek közül a legtermészetesebben hangzó változatot kell kiválasztani. Ezt a folyamatot elemkiválasztásnak nevezik, arra utalva, hogy a többféle lehetséges elem közül kiválasztjuk, hogy melyek kerüljenek összefűzésre egy adott bemondás előállításához. Megvalósítása a hibajavító kódolásban és a beszédfelismerésben is alkalmazott Viterbi algoritmus segítségével történik.

Megjegyezzük, hogy a hagyományos elemösszefűzésen alapuló beszédszintézis rendszernél is szükség van elemkiválasztásra, ha az adatbázis vegyesen tartalmaz diád- és triád-elemeket. Ugyanakkor – mivel minden elem csak egy változatban, egy adott prozódiával szerepel – a kiválasztás kevesebb számítással megoldható [6].

A korpusz alapú elemkiválasztásos beszédszintézis alapvetően két ok miatt eredményez jelentős minőségjavulást a hagyományos elemösszefűzéshez képest. Egyrészt kevesebb összefűzési pontot tartalmaz, mint a logatomos diád-, triád-elemekből építkező rendszer, ami folytonosabb, természetesebb hangzást eredményez [2]. Másrészt a megfelelő prozódia kialakításához kevesebb jelfeldolgozási művelet szükséges, mivel az adatbázis elemek prozódiai változatossága miatt általában ki tudunk olyan elemeket választani, amelyek közel állnak a kívánt prozódiához. Emellett a diád-elemeknél jóval hosszabb, egybefüggő beszédda-

3. ábra
Korpusz alapú, elemkiválasztásos beszédszintetizátor felépítése



rabok természetes prozódiája is megőrizhető, illetve (ha egyáltalán szükséges a prozódia módosítása) viszonylag egyszerűen eltolással biztosítható az egybefűzött darabok prozódiai illeszkedése. Ez természetesebb hangzást eredményez, mint a szabály alapon előírt, a diád-elemekből összerakott jelre kényszerített mesterséges prozódia.

A korpusz alapú szintézis velejárója a beszédadatbázis méretének jelentős növekedése. Az *1. táblázat* mutatja az adatbázis méretének változását az egyre jobb minőséget biztosító beszédszintézis technológiák esetén. A nagyobb tárigény mellett az elemkiválasztás nagyobb számításigényt is támaszt a korábbi rendszerekhez képest. Ezeket részben kompenzálja a számítástechnikai eszközök (háttértár, CPU) időközben végbement fejlődése.

A diád-, illetve triád-elemek összefűzésén alapuló szintézishez hasonlóan, itt is szükséges az elemhatárok jelölése a felvett beszédadatbázisban. Az 1. táblázatból látható, hogy az adatbázis időtartama, és ezzel arányosan a benne szereplő hangok száma több mint egy nagyságrenddel megnőtt. Az elemhatárok jelölése már nem végezhető a korábbi, sok kézi munkát igénylő módszerrel. Szerencsére a gépi beszédfelismerés technológiájának időközben bekövetkezett fejlődése lehetővé teszi az elemhatár-jelölés automatizálását.

A korábbi, formáns szintézisen, illetve a diád- és triád-elemek összefűzésén alapuló rendszerek minősége nem függött a felolvasandó szöveg tartalmától. A korpusz alapú szintézis minőségét nagyban befolyásolja, hogy a szintetizálandó szöveg mennyire van közel a szintetizátor beszédkorpuszának, azaz adatbázisának tartalmához. Minél közelebbi a szintetizálandó szöveg, annál nagyobb beszéd-darabokat lehet egyben kivenni az adatbázisból, megőrizve azok természetességét. Adott esetben előfordulhat az is, hogy a szintetizálandó szöveg egy-egy mondatát teljes egészében megtaláljuk. Ellenpéldaként viszont lehet, hogy különálló diád- vagy triád-méretű elemekből kell összerakni egy, az adatbázisban nem szereplő szövegrészt, megnövelve a vágási pontok számát. A fenti ok miatt célszerű a szintetizátor beszédkorpuszát a felolvasandó szövegekhez, azaz az adott alkalmazáshoz igazítani (például csak időjárási szövegekből álló korpusz időjárásjelentés felolvastatásához).

5. A BME-TMIT kísérleti, korpusz alapú beszédszintetizátora

A BME-TMIT-en elkészült egy magyar nyelvű korpusz alapú elemkiválasztásos beszédszintetizátor kísérleti változata. Korábban már beszámoltunk a fejlesztés kezdeti lépéseiről [4]. Itt az azóta elvégzett munka eredményeit ismertetjük.

Célunk első lépésben nem egy általános beszédszintetizátor kifejlesztése, hanem egy specifikus alkalmazás, egy időjárás-jelentés felolvasó megvalósítása volt. A későbbiekben a fejlesztés során szerzett tapasztalatainkat használjuk majd fel egy általános célú, korpusz alapú elemkiválasztásos beszédszintetizátor megvalósításához.

A kísérleti rendszer beszédkorpuszát Internetről gyűjtött időjárás-jelentés szövegekből állítottuk össze (ennek részleteit lásd [2]-ben). A szöveganyag 5400 mondatát egy fiatal színésznő olvasta fel, amit stúdió körülmények között, 44,1 kHz-es mintavételi frekvenciával, mintánként 16 biten rögzítettünk. A felvételek 4 héten keresztül zajlottak, heti 2-3 nap terjedelemben. Az 5400 mondat felvételével mintegy 11 órányi folyamatos beszédből álló hanganyag állt össze.

5.1. Hangfelvételek címkézése

Az adatbázis hanganyagát mondatokra bontottuk szét. Minden egyes mondathoz tartozik egy szöveges átírás. Ezt a mondat felolvasásához használt szövegből származtattuk. A mondatokban az elemhatárok, illetve a zöngeperiódus-határok jelölése automatikusan

Szintetizátor technológia	Adatbázis tartalma	Mintavételi frekvencia	Adatbázis mérete	Hangfelvétellel előálló adatbázis időtartama	Hangfelvétel időigénye
Multivox formáns szintetizátor *(1985)*	12 paraméterrel meghatározott hangszeletek (4 formáns)	11,025 kHz	1,2 K	parametrikus szintézis, nincs hangfelvétel	parametrikus szintézis, nincs hangfelvétel
Profivox diád-elemes szintetizátor *(1997)*	1444 diád-elem	8 kHz/ 22,05 kHz	1,5 MB/ 7 MB	3 perc	4 óra
Profivox diád- és triád-elemes szintetizátor *(2002)*	1444 diád- és 6000 triád-elem	22,05 kHz	92 MB	35 perc	10 óra
Kísérleti korpusz alapú szintetizátor *(2006)*	5400 mondat	22,05 kHz	1,7 GB	11 óra	50 óra

1. táblázat

A BME-TMIT-en fejlesztett beszédszintetizátorok adatbázisméretének növekedése

történt. A zöngeperiódusok a beszéd zöngés (kvázi-periodikus) részének egy-egy periódusát jelentik. A zöngétlen részeken a beszéd nem periodikus, ezért ott 5 ms-onkénti jelöléseket alkalmaztunk.

A zöngeperiódus-határok jelölése egyrészt az alapfrekvenvcia pillanatnyi értékének a meghatározásához kell, másrészt az alapfrekvencia-menet esetleges módosításához szükséges. A jel alapfrekvencia-menete kiszámítható a periódusidők reciprokaként. Az alapfrekvencia-menetet az elemkiválasztás folyamán is felhasználjuk. A zöngeperiódus-határok bejelöléséhez a Praat fonetikai és beszéd-analizátor szoftverben implementált ablakfüggvénnyel korrigált autokorrelációs módszeren alapuló alapfrekvencia-detektálást használtuk fel [7].

A szintetizátorban kétféle építőelemtípust definiáltunk, ezek a szó és a beszédhang. A beszédhang biztosítja a teljes fedést, azaz, hogy tetszőleges tartalmú szöveget elő lehessen állítani. A szó szint gyorsabb keresést tesz lehetővé az adatbázisban és biztosítja a bemondó hangszínetéhez közel álló hangzás elérését. Ennek megfelelően a felvett beszédkorpuszban jelölni kell a hang és szóhatárokat. A hanghatárok jelölése a beszédjel szintjén történik, azaz megadjuk, hogy hányadik mintán kezdődnek az egyes hangok. Ezt a feladatot egy a BME-TMIT-en kifejlesztett magyar nyelvű beszédfelismerő [8] segítségével oldjuk meg. A felismerőt kényszerített módban használjuk, ami azt jelenti, hogy a beszédet tartalmazó hangfájl mellett bemenetként megadjuk annak szöveges tartalmát is, ami meghatározza, hogy milyen hangsorozatokat keressen a felismerő.

Ehhez először a szöveg fonetikus átírását kell elvégezni, ami automatikusan történik, a magyar nyelv hasonulási szabályainak figyelembe vételével. Megjegyezzük, hogy a szövegben szereplő rövidítések, illetve az esetleges speciális jelek (pl. mínusz jel) átírását kézzel kell elvégezni és ellenőrizni. A leírt szöveg nem definiálja egyértelműen a megvalósuló hangsorozatot. Egyrészt előfordulhatnak kiejtési variációk (de ezek inkább csak a spontán beszédre jellemzőek), másrészt a szóhatároknál csak opcionálisan tartunk szünetet. Az is előfordul, hogy szóhatáron átívelő hasonulás, vagy egybeolvadás jön létre. Ezekre példa a „hűvös záporok”, ami legtöbbször „hűvözs záporokként” kerül kimondásra, illetve az „ideig ködös” szókapcsolat, ahol általában egy darab hosszú „k” hang valósul meg a szóhatáron. Ennek kezelésére a felismerő a fonetikus átírás többféle változatát állítja elő, amit egy irányított gráf ír le.

Jelen változatban a gráf csak a szóhatároknál ágazik ketté egy, az adott szóhatáron szünetet tartalmazó, illetve egy szünetet nem tartalmazó hangsorozattá. A szünetet nem tartalmazó változatban működnek a hasonulási és egybeolvadási szabályok. Ezek az alternatív útvonalak itt nem részletezett módon beépülnek a rejtett Markov modelleket használó beszédfelismerő keresési terébe, és ezek közül a beszédjel alapján a legvalószínűbb hang kerül kiválasztásra a Viterbi algoritmus felhasználásával. A kényszerített felismerés nem csak a hanghatárokat, hanem a beszédjelben megvalósult hangsorozatot is megadja kimenetként.

A felismerő 20 ms-os keretekkel és 10 ms-os kereteltolással dolgozik, azaz a hanghatárokat elvileg is csak 10 ms-os pontossággal határozza meg. A hanghatárokat (zöngés hangok esetén) a legközelebbi zöngeperiódushoz igazítjuk, ezáltal biztosítjuk, hogy szintetizált mondatban egymás mellé kerülő beszéd-darabok azonos fázisban legyenek, azaz teljes periódusokból álljanak. Így nem lesz pattogó, vagy rekedtes hangzású az előállított beszéd.

A hanghatár-jelölés ellenőrzéséhez minden egyes hangra egy hanghossz eloszlás hisztogrammot készítettünk. Ennek segítségével meghatároztuk azokat a hangokat, amelyek hossza jelentősen eltért a velük azonos hangok átlagolt hosszaitól. Az ilyen hangokat tartalmazó mondatokat külön-külön manuálisan megvizsgáltuk. A tapasztalt hibákból sorolunk fel néhányat.

Az abnormális hanghosszak egy része átírási hibákból származott, azaz a hangfájlok és a hozzájuk tartozó szöveges fájlok tartalma nem minden esetben felelt meg egymásnak. Az is előfordult, hogy a szövegben olyan rövidítések maradtak, amelyeket nem tudott megfelelően feloldani a fonetikus átíró.

A hibák másik részét az automatikus hanghatár-jelölés okozta. Tipikus hiba például a szóvégi réshangok (“f”, „sz”) és a szavak közötti esetleges levegővételek egymásba csúszása. Hasonlóan a „c” hang határa is sok esetben a szomszédos hangra csúszott. Ennek valószínű oka, hogy a beszédfelismerő tanítása egy telefonon keresztül felvett adatbázissal történt, amelyben az átviteli tulajdonságok miatt a „c” hang spektrumának nagy része elveszett. Ennek megoldására a felismerőt magával a felcímkézendő hangadatbázissal kell betanítani, ami várhatóan pontosabb hanghatárokat fog eredményezni.

A korábbi tapasztaltok alapján a 11 órányi folytonos beszédkorpusz elegendő hosszú a felismerő megfelelő tanításához.

A szóhatárok jelölését a beszédfelismerő által visszaadott, a beszédjelhez legjobban illeszkedő fonémasorozaton végeztük. A szavakon átívelő hangegybeolvadások (például: „ideig ködös”) miatt előfordulhat, hogy egy hang egyszerre két szóhoz is tartozik. Ennek kezelésére külön jelölést alkalmaztunk a szavak kezdetére és végére, így lehetővé tettük a szavak közötti átfedést (például: „<idei<k>ödös>”, ahol „<” a szókezdetét, „>” a szó végét jelöli).

A szóhatárok jelölését teljesen automatikusan végeztük, olymódon, hogy a beszédfelismerő által visszaadott fonémasorozatot illesztettük a szintén a beszédfelismerő által előállított, (a szóhatároknál elágazó) összes lehetséges fonetikus átírást megadó irányított gráfhoz. Az illesztést egy állapotgéppel végeztük, amely a fonémasorozat és a gráf alapján követte, hogy mikor merre ágazott el a felismerő, és az elágazásoknál a választásnak megfelelően szúrta be a szóhatárokat.

5.2. Prozódia

A kísérleti korpusz alapú szintetizátor első lépésben a szöveg fonetikus átírását végzi, ami nem változott a korábbi szintetizátorokhoz képest. A fonetikus átírást követően a prozódia előírására kerül sor. Az alábbiakban ennek lehetséges megvalósítási módszereit elemezzük.

Az elemösszefűzésen alapuló beszédszintetizátor szabályalapú prozódiája alkalmazható a korpusz alapú, elemkiválasztásos rendszerben is. A megközelítés előnye, hogy a szabályok átvehetők a korábbi rendszerből, hátránya viszont, hogy a megvalósított prozódia ugyan elfogadható, de nem ad természetes hangzást. A szabályalapon meghatározott prozódia először szimbolikus, majd fizikai szinten áll elő. A korábbi, diádos, triádos elemösszefűzésen alapuló technológiánál a monoton elemek egymásután illesztésével előálló beszédjelre kellett a fizikai-szintű prozódiát ráültetni. A korpusz alapú rendszernél azonban lehetőség nyílik arra, hogy az előírt prozódiát a jel összefűzése előtt, az elemkiválasztás folyamán vegyük figyelembe. Ez történhet fizikai szinten, de akár a szimbolikus prozódia szintjén is. Az előbbi azt jelenti, hogy például az előírt alapfrekvencia-menethez minél közelebbi elemeket választunk ki. Az utóbbi pedig azt jelenti például, hogy ha a bemeneti szöveg alapján előírt szimbolikus prozódia szerint hangsúlyos egy szó, akkor a beszédkorpuszban olyan elemet, vagy elemeket keresünk, amely az előírás szerinti hangsúllyal rendelkezik. Ennek fizikai megvalósítása számos problémát vet fel, amiket itt nem részletezünk.

A szimbolikus prozódia alapján történő elemkiválasztás előnye, hogy megőrződik a kiválasztott elemek természetessége, hátránya, hogy a vágási pontoknál megtörhet a természetes prozódia, ezért utólagos prozódia-simításra van szükség. Ez azt jelenti, hogy az előállított mondatban szereplő beszédelemek alapfrekvencia-, és intenzitás-menetét, illetve hangidőtartamait jelfeldolgozási módszerekkel úgy módosítjuk, hogy az egymás melletti elemek között folytonos legyen az átmenet. Megjegyezzük, hogy abban az esetben is szükség lehet ilyen simításra, ha a fizikai szintű prozódiát használjuk az elemkiválasztás folyamán. Továbbá ekkor sem feltétlenül kell jelfeldolgozással pontosan a jelre kényszerítenünk az előírt fizikai prozódiát.

A szimbolikus prozódia használata esetében probléma, hogy mind a szintetizátor beszédkorpuszában, mind a szintetizálandó szövegben jelölni kell a szimbolikus információt (például a hangsúlyokat). Ezt a jelenlegi automatikus módszerek csak pontatlanul tudják megtenni. A kézi jelölés a beszédkorpusz mérete miatt nem praktikus megoldás. Ráadásul nem garantált a konzisztencia a korpuszban kézileg jelölt hangsúlyok, illetve a bemeneti, szintetizálandó szövegből előre jelzett hangsúlyok között. A szimbolikus szinthez képest még egy szinttel magasabb információkat is használhatunk az elemkiválasztás során. A szimbolikus prozódiát a bemeneti szöveg (leegyszerűsített felszíni) nyelvi elemzése alapján határozzuk meg. Ennek a nyelvi elemzésnek a kimenete alapján is kereshetünk a beszédkorpuszban. A módszer használata esetén szintén szükséges az összefűzött elemek prozódiai simítása.

További lehetőség a korpusz alapú fizikai prozódia generálása. A korpusz mondatainak fizikai prozódiáját ki lehet nyerni a korpuszból. Ez azt jelenti, hogy a szimbolikus prozódiából nem csak szabályok segítségével tudjuk előállítani a fizikai prozódiát, hanem a szimbolikus prozódia alapján a beszédkorpuszban, mint fizikai prozódia tárban is kereshetünk. A keresés történhet a beszédkorpusz mondataihoz tartozó (előre meghatározott), illetve a szintetizálandó szöveghez meghatározott szimbolikus prozódia egyezése alapján. Természetesen itt is elképzelhető a nyelvi elemzés szintjére történő visszalépés, azaz a nyelvi elemzés által megadott információk alapján történő keresés. Az adatbázisból a keresés során kinyert fizikai prozódiát előírhatjuk az előállítandó mondat cél-prozódiájaként.

A kísérleti rendszerben leegyszerűsített szimbolikus prozódiát használunk, szópozíció jellegű információ formájában. Az elemkiválasztás előtt a bemeneti szöveget prozódiai egységek szerint tagoljuk. A prozódiai egység a szintetizátor jelenlegi megvalósításában egy írásjelekkel határolt, tagmondat-jellegű szövegrészt jelent. Minden egyes prozódiai egységet megcímkézünk aszerint, hogy a mondaton belül milyen pozícióban van (első-, utolsó-, közbenső szó).

A szópozíció százalékos formában megadja az adott szó helyét az azt tartalmazó prozódia egységben. Ezt a prozódiai egység fonémákban megadott hossza és a szó első/utolsó fonémájának prozódiai egységen belüli pozíciója alapján megállapított százalékos értékek definiálják. Az elemkiválasztás folyamán megpróbálunk a bemeneti szövegben szereplő szavakhoz hasonló pozíciójú szavakat kiválasztani. Természetesen a pozíció jellegű információ nem határozza meg egyértelműen sem a hangsúlyokat, sem a hangidőtartamokat. A módszer előnye az egyszerűsége, hátránya, hogy sok esetben nem biztosít megfelelő prozódiát. A kísérleti rendszer jelenlegi implementációja nem tartalmaz utólagos prozódia simítást.

5.3. Elemkiválasztás és összefűzés

A prozódia előírása után, a kísérleti korpusz alapú szintézis rendszer harmadik lépése az elemkiválasztás. Az elemkiválasztás alapelve, hogy a rendszer a bemeneti szöveget (elviekben) az összes lehetséges módon összerakja a beszédkorpusz elemeiből, és azok közül a legtermészetesebben hangzót választja ki. A természetesség automatikus megállapításához kétféle költséget vezetünk be.

Az egyezési költség megadja, hogy egy adott elem mennyire felel meg a szintetizálandó beszédszakasznak. A jelenlegi megvalósításban a beszédszakaszt annak betűsorozatként megadott szöveges tartalma, illetve hangsorozatként megadott fonetikus átírása határozza meg. Ehhez járulnak még a szintetizálandó szövegből meghatározott prozódiai előírások, ami jelenleg

a szó-, illetve hangpozíciókat jelenti. Az összefűzési költség azt adja meg, hogy a leendő szomszédos elemek mennyire folytonosan illeszkednének egymáshoz. Az elemkiválasztás folyamata azt a mondatot választja ki az összes lehetséges közül, amelyre az egyezési és összefűzési költségek összege a legkisebb.

A kísérleti szintetizátor kétféle elemet, szavakat és beszédhangokat kezel. Szószintű keresés esetén az adott szót alkotó betűsorozat alapján azonosítjuk az elemeket, míg beszédhangszintű keresés esetén a hangot megadó fonéma alapján. Az elemkiválasztás hierarchikusan történik. Első menetben csak a szószintű elemek között keres a rendszer. Ha a bemeneti szövegnek vannak olyan szavai, amit nem sikerült szóalapon lefedni, akkor a hiányzó szavakat beszédhangokból rakja össze a rendszer. A szószinten már megtalált elemeket nem próbáljuk kisebb elemekből előállítani, még akkor sem ha ez esetleg prozódiai szempontból célszerű lenne.

A szószint bevezetésének alapvető előnye a gyorsabb keresés. Ennek hatékonysága függ a szintetizátor beszédkorpusza és az előállítandó szöveg közötti hasonlóságtól. Egy általános célú beszédszintetizátornál ez kevésbé hatékony megoldás – különösen a ragozó magyar nyelv esetében – ugyanakkor korlátozott tematikájú alkalmazások, például időjárásjelentés-felolvasás esetén jól működik. Általános célú alkalmazásnál is gyorsítható a keresés a beszédhangnál hosszabb elemek, például szótagok bevezetésével.

Az elemkiválasztást mondatonként végezzük. A keresés folyamán az adott mondatban szereplő egy-egy szóhoz, vagy hanghoz többféle lehetséges jelöltet is kiválasztunk. Egy-egy elemhez implementációs és hatékonysági okokból maximáltuk a lehetséges jelöltek számát. Ha a kiválasztás folyamán egy adott elemhez tartozó jelöltek száma elérte a megadott maximumot, akkor minden egyes további jelölt hozzávétele után a legmagasabb egyezési költségű elemet eldobjuk. Az összefűzés során elvileg minden elem esetében tetszőleges jelöltet kiválaszthatunk. Ebből következőleg a különböző előállítható lehetséges mondatok számát a jelöltek számának szorzata adja meg. Az optimális mondat kiválasztását a dinamikus programozáson alapuló Viterbi-algoritmus segítségével végezzük. Az algoritmus minden egyes lehetséges útra (mondatra) meghatározza az egyezési és összefűzési költségek összegét, és a minimális költségű utat választja ki.

Az egyezési költségben a következők szerepelnek:

1. A jelöltet (szó vagy beszédhang) megelőző és követő fonéma egyezése az előírt célelemet megelőző és követő fonémával. A legkisebb költséget a teljes egyezés jelenti. Emellett definiáltunk egymással helyettesíthető fonéma-kategóriákat [2]. Az azonos kategóriába eső hangok egyezési költsége kisebb, mint az eltérő kategóriákba esőké. Eltérő kategóriák esetén az egyezési költséget egy költségmátrix definiálja, amelynek értékei minden esetben nagyobbak, mint azonos kategória esetén.
2. Prozódiai egység mondaton belüli pozíciójának egyezése. Ezt csak szavak esetében vesszük figyelembe.
3. Prozódiai egységen belül előírt pozíciótól való eltérés. Ezt csak szavak esetében vesszük figyelembe.

Az összefűzési költség a következők szerint alakul:

1. Ha a vizsgált két jelölt a beszédkorpuszban egymás után következett, akkor az összefűzési költség mindig 0.
2. Ha a vizsgált két jelölt a beszédkorpuszban azonos mondatból származott, akkor kisebb összefűzési költséget rendelünk hozzá, mintha eltérő mondatokból származott volna.
3. Alapfrekvencia-menet folytonossági költsége, amit az első elem végső és a második elem kezdő alapfrekvenciájának eltérésével arányosan számolunk.

Az egyes költségek értékeit számos hangminta meghallgatása során, ad hoc módon állítottuk be. Ezek optimalizálása még tovább javíthatja az előállított beszéd minőségét.

6. Szintézis rendszerek három generációjának összehasonlítása szubjektív minősítéssel

A következőkben leírt vizsgálatunk célja annak a meghatározása volt, hogy mekkora a minőségi ugrás a különböző generációjú beszédszintézis rendszerek között. Emellett alaposabban akartuk vizsgálni a kísérleti korpusz alapú, elem-összefűzéses szintetizátorral előállított mondatok minőségét.

Korábban már ismertettünk egy hasonló tesztet a diád- és triád-elemek összefűzésén alapuló technológia, és a korpuszos, elemkiválasztáson alapuló szintetizátor összehasonlítására [4]. Az ott ismertetett tesztben a korpuszos rendszer működését kézzel összevágott mondatok szimulálták, mivel akkor még nem állt rendelkezésre működő kísérleti rendszer.

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 teszt során a tesztelők véletlenszerű sorrendben hallgatják meg a különböző szintetizátorokból származó mondatokat, és azok minőségét egyenként osztályozzák egy ötfokozatú skálán *(2. táblázat)*. Az osztályzatok összes tesztelőre vonatkoztatott átlaga adja meg a MOS értékét.

Megjegyezzük; a beszédszintetizátoroknál hagyományosan az érthetőséget szokás vizsgálni, ez azonban az újabb rendszereknél kevésbé okoz problémát.

5	kiváló
4	jó
3	közepes
2	gyenge
1	rossz

2. táblázat
Ötfokozatú szubjektív hangminősítő skála

Továbbá létezik a beszédszintetizátorok minősítésére vonatkozó P.85-ös ITU-T szabvány [9], amely más szempontokat, például a szintetizált szöveg megértéshez szükséges koncentráció mértékét is vizsgálja. Ez a gyakorlatban nem terjedt el. Ennek lehetséges magyarázatát kereste egy tanulmány [10], amely kimutatta, hogy a különböző vizsgált szempontokra vonatkozó minősítések nagy korrelációt mutatnak, azaz valójában nem adnak plusz információt az egyszerű minősítéshez képest, viszont fölöslegesen növelik a teszt idejét és költségét.

A jelen tesztben három, a BME-TMIT-en fejlesztett magyar nyelvű beszédszintetizátort hasonlítottunk össze. Az első a Multivox formánsszintetizátor női hangú változata, a második a Profivox szintetizátor diád- és triád-elemeket tartalmazó változata volt. Ennek adatbázisa ugyanattól a női bemondótól származott, mint a vizsgálatban szereplő harmadik, kísérleti korpusz-alapú, elem-összefűzéses beszédszintetizátoré. (Megjegyezzük, hogy a Multivox és a Profivox szintetizátor – tapasztalataink szerint – férfi-hangú adatbázissal némileg jobb minőségű beszédet ad, mint a női adatbázisokkal. Ez megegyezik a nemzetközi tapasztalatokkal, és a hangok eltérő fizikai jellegzetességeiből adódik. Például a magasabb frekvenciájú hang minősége jobban romlik prozódia-módosítás esetén.) A szintetizált mondatok mellett felvettük azok természetes változatát is, a beszédkorpusz hangját adó ugyanazon bemondó közreműködésével.

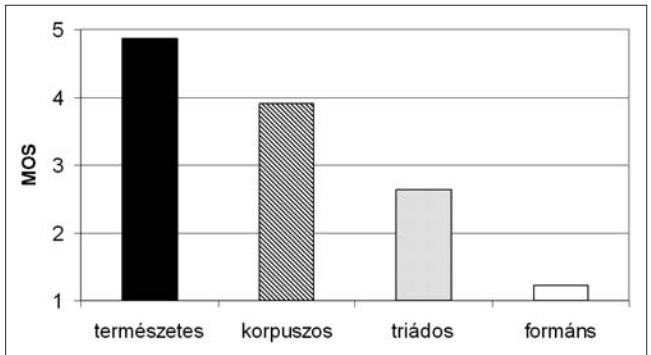
Korábbi tapasztalataink alapján egy 30 perces meghallgatásos teszt folyamán az átlagos motivációjú tesztelő elveszíti érdeklődését. Ennek elkerülésére egy körülbelül 10 perces tesztet állítottunk össze, melyben kizárólag a szintetizált beszéd minőségét kellett értékelniük a tesztelőknek. A tesztben szereplő szövegek tartalmát a kísérleti korpusz-alapú szintetizátor által megcélzott időjárás-jelentés felolvasáshoz igazítottuk. A tesztanyag tíz időjárás-jelentésekből származó mondatot tartalmazott *(5. táblázat)*. A mondatokat egy időjárásjelentés-portálról véletlenszerűen választottuk. Ugyanakkor a portálon megjelent korábbi anyagokat felhasználtunk a kísérleti szintetizátor beszédkorpuszának összeállításához, így azok stílusa „ismerős" volt a szintetizátor számára. Egy tesztelőnek összesen 40 mondatot kellett meghallgatnia (ebből 10 mondat természetes, 10 a Multivox szintetizátorral, 10 a Profivox szintetizátorral és 10 a kísérleti korpusz-alapú szintetizátorral lett előállítva).

A tesztet az Interneten keresztül, egy web-es felület segítségével kellett elvégezni. Ez lehetővé tette nagyszámú tesztelő részvételét. A 248 tesztelő nagy része egyetemi hallgatók közül került ki. A résztvevők átlagkora 22,9 év (a legfiatalabb 12, a legidősebb 63 éves volt). A kiértékelésre került adatok 185 férfitől és 36 nőtől származnak. A tesztelők 15%-a (27 fő) által adott értékelést – a későbbiekben részletezett okok miatt – nem használtuk fel. A 40 mondat meghallgatása előtt egy, a tesztben nem szereplő (de szintén időjárás-jelentésből származó) mondatot kellett a három szintetizátorral előállítva, illetve természetes változatban meghallgatni. Így minden egyes tesztelő kialakíthatta saját szubjektív rangsorát a későbbi teszteléshez. Egy ilyen előzetes „ismerkedési" fázis általában része a MOS-típusú teszteknek. A teszt folyamán minden mondatot csak egyszer lehetett meghallgatni. Ez csökkentette a teszt elvégzéséhez szükséges időt. A tesztelők által a meghallgatáshoz használt eszközök minőségét érthető okok miatt nem tudtuk szabályozni. Ugyanakkor a teszt elején kértünk erre vonatkozó adatokat. Ezek szerint a tesztelők többsége átlagos, otthoni minőségű eszközön végezte a tesztet, csendes környezetben.

A kiértékelésből kizártuk azokat a tesztelőket, akik a tízből legalább két esetben közepes (3), vagy rosszabb minősítést adtak a természetes beszédből származó mondatokra. Feltételeztük, hogy ezek a tesztelők vagy komolytalanul „össze-vissza" válaszoltak, vagy technikai problémák miatt akadozással hallhatták az egyes felvételeket. Az utóbbi problémával mi is találkoztunk előzetes tesztjeink során. A probléma lassabb Internetkapcsolat esetén jelentkezett, és a hangfájl letöltés akadozása miatt keletkezett. Ezért a 22 kHz-es monó hangfájlokat egy 32 kbps sebességű MPEG1-LIII kódolóval tömörítettük, változó bitsebességű módban. Az esetek többségében sikerült megszűntetni a lejátszás korábbi akadozását. Stúdió minőségű fejhallgatóval végzett informális meghallgatás során nem találtunk észrevehető különbséget az eredeti és a tömörített felvételek hangzása között. A fentieken kívül azokat a tesztelőket is kizártuk, akiknek az eredményét webszerver túlterhelés miatt nem tudtuk helyesen rögzíteni. Összességében a tesztelők 15%-át zártuk ki.

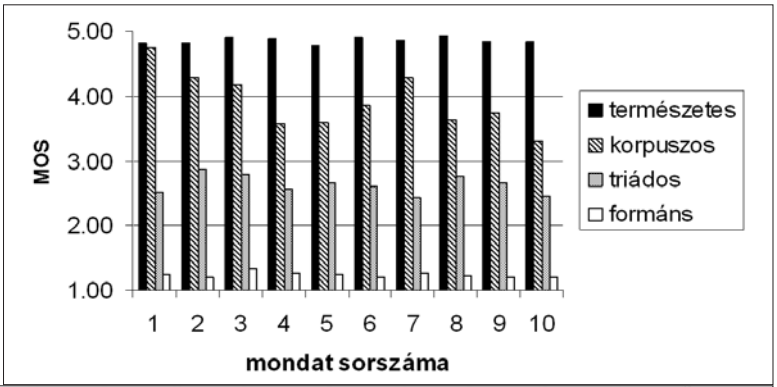
A *4. ábra* és a *3. táblázat* mutatja a teszt összesített eredményét. Az eredményekből jól érzékelhető, hogy a beszédszintetizátorok generációváltása minden esetben jelentős minőségjavulást eredményezett, ugyanakkor a jelenlegi legjobb rendszerek sem érik el a teljesen természetes hangzást. A generációk közötti ugrásszerű javulás a rendszerek szélesebb-körű használatát, új alkalmazások bevezetését tette, teszi lehetővé.

4. ábra és 3. táblázat
Szubjektív minősítés átlagai
az egyes szintézis technológiákra



	természetes	korpuszos	triádos	formáns
MOS	**4,86**	**3,92**	**2,63**	**1,24**
konfidencia (α = 0,05)	**0,02**	**0,03**	**0,03**	**0,02**

5. ábra és 4. táblázat
Szubjektív minősítés átlagai mondatonként. A MOS értékek konfidenciája (α=0,05) 0,04 és 0,09 közötti.



sorszám	1	2	3	4	5	6	7	8	9	10	szórás
természetes	**4,83**	**4,83**	**4,91**	**4,88**	**4,79**	**4,90**	**4,86**	**4,92**	**4,84**	**4,84**	0,04
korpuszos	**4,75**	**4,28**	**4,16**	**3,56**	**3,59**	**3,85**	**4,29**	**3,63**	**3,75**	**3,30**	0,44
triádos	**2,52**	**2,88**	**2,79**	**2,56**	**2,66**	**2,60**	**2,44**	**2,76**	**2,65**	**2,46**	0,15
formáns	**1,25**	**1,20**	**1,33**	**1,26**	**1,25**	**1,20**	**1,26**	**1,21**	**1,21**	**1,20**	0,04

Az *5. ábra* és a *4. táblázat* mutatja a mondatonkénti átlagos minősítéseket. A táblázatban szereplő szórás értékeket összehasonlítva látható, hogy a korpusz-alapú, elemkiválasztásos szintetizátor minőségének a legnagyobb az ingadozása.

A rendszer legjobban sikerült mondata elérte a természetes minőséget, a legrosszabb mondat pedig 1,5 jeggyel rosszabb minősítést kapott. Ugyanakkor az összes mondat jelentősen meghaladja a korábbi rendszerek minőségét. Ez az ingadozás egyrészt a technológia velejárója, másrészt a kísérleti rendszer hiányosságaiból, például a leegyszerűsített prozódia-előállításból származik. Várakozásaink szerint a rendszer továbbfejlesztett változatában a rosszabban hangzó mondatok minősége javulni fog, csökkentve a minőség ingadozását.

A következőkben azt vizsgáltuk, hogyan függ össze a korpuszos szintetizátorral előállított mondatok minősége az azokban szereplő vágási pontok számával. Az *5. táblázatban* megadjuk a tesztben szereplő mondatokat, szóközzel jelölve a vágási pontokat, aláhúzással jelölve a beszédszüneteket.

A *6. táblázat* számszerűleg is összesíti a vágási pontokat. A mondatokat a bennük szereplő vágási pontok száma szerint rendeztük növekvő sorrendbe. A legjobb minősítést a legkevesebb, azaz három vágási pontot tartalmazó mondat érte el, míg a legrosszabb minősége a legtöbb, 24 vágási pontot tartalmazó mondatnak volt. Ugyanakkor a minőség nem függ konzisztensen a vágási pontok számától. Például a 7. (12 vágási pontot tartalmazó) mondat minősége a második legjobb, megelőzve sok kevesebb vágást tartalmazó

5. táblázat
A teszteléshez használt mondatok tartalma. A szóközök a korpuszos szintetizátor vágási pontjait jelölik, míg az aláhúzások a beszédszüneteket.

1	éjszakatöbbfelépárássáválika levegő_foltokban ködis képződhet
2	a párás_ néhol ködösreggeltkövetőentöbbórára kisütanap
3	akövetkező napokbanfolytatódik az igazitéli időjárás
4	felhős égboltra _ jelentős esőzések kialakulására kellszámítani
5	azészakkeleti szél helyenként megélénkül _ csütörtökön a tiszántúlon néhol megerősödik
6	aszél mérsékelt északkeleti lesz _a hőmérséklet tizenegy _ tizenhat fok körülalakul
7	budapesten eleinte többnyireerősen felhősleszazég _ majd csökkena felhőzet_ és csütörtökön néhány órára kisütanap
8	péntektől ismét csapadékosra fordulazidő _ többfelé várható havaseső _ havazás _és egyúttal lehűlés is kezdődik
9	csütörtökön északkeleten többórárakisüta nap_ délnyugaton azonban méggyakran leszerősen felhősazég_ ésott helyenként kisebbeső _zápor továbbra is valószínű
10	akövetkezőnapokban a cs ila gá sz a ti tél az átlagosnál jóvalenyhébb idővelin d ít _ sok felhőreszámíthatunk_ és főleg szerdán kell számítani sokfelé esőre

6. táblázat
Szubjektív minősítés (-0,68-as) korrelációja a vágási pontok számával, korpusz alapú elem-összefűzéses kísérleti beszédszintetizátorban.

mondat sorszáma	1	2	3	4	5	6	7	8	9	10
MOS (korpuszos)	**4,75**	**4,28**	**4,16**	**3,56**	**3,59**	**3,85**	**4,29**	**3,63**	**3,75**	**3,30**
vágási pontok száma	**3**	**4**	**4**	**6**	**9**	**10**	**12**	**14**	**15**	**24**
szavak száma	10	11	8	7	10	12	9	15	25	22
elemekből összefűzött szavak száma	0	0	0	0	0	0	0	0	0	2

mondatot. A vágási pontok száma és a minősítés között -0,68 a korreláció, ami szintén a jelentős, de nem teljesen konzisztens összefüggésre utal. A 6. táblázatból az is látszik, hogy csak egy mondatban került sor hangszintű elemek használatára. A mondatokat meghallgatva megállapítható, hogy minőségi problémák leginkább a helytelen, nem illeszkedő prozódiából adódnak.

Ugyanakkor látható, hogy az egyszerű modell is tízből négy-öt esetben jónak minősített prozódiát ad. A prozódia-simítás, illetve a pozíció alapú megközelítés helyett komplexebb prozódiai modell megvalósításával a probléma várhatóan csökkenthető.

7. Összefoglalás

A beszédszintetizátorok generációváltása minden esetben jelentős minőségjavulást eredményezett, ugyanakkor a jelenlegi rendszerek sem érik el a teljesen természetes hangzást.

A generációk közötti ugrásszerű javulás a rendszerek szélesebb körű használatát, új alkalmazások bevezetését tette, teszi lehetővé. Látható, hogy a generációk közötti váltás nem a teljes rendszert, hanem csak egyes részeinek lecserélését érinti. Míg a korábbi rendszerek viszonylag függetlenül fejlődtek a beszédfelismeréstől, a legutóbbi, korpusz alapú, elemkiválasztásos rendszer nagymértékben támaszkodik az automatikus beszédfelismerés technológiájára.

A cikkben bemutatott kísérleti korpusz alapú elemkiválasztásos beszédszintetizátor már jelen állapotában is meghaladja a korábbi rendszerek minőségét, korlátozott tematikájú területen. A rendszer minősége ugyanakkor egyenetlen, de ez várhatóan további fejlesztéssel csökkenthető.

A korlátozott tematikájú rendszer fejlesztése során nyert tapasztalatok alapján egy általános elemkiválasztásos szintetizátor készítése is elérhető távlatba került.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani a szubjektív kiértékelésben résztvevő nagyszámú tesztelőnek. Külön köszönet illeti Bartalis Mátyást a webes tesztfelület elkészítéséért és Mihajlik Pétert a magyar nyelvű beszédfelismerő eszközök használatához nyújtott segítségéért.

A kutatást az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Irodalom

[1] Olaszy, G., G. Gordos, G. Németh:
„The MULTIVOX multilingual text-to-speech converter”, In: G. Bailly, C. Benoit and T. Sawallis (eds.): Talking machines: Theories, Models and Applications, Elsevier, 1992. pp.385–411.

[2] Olaszy Gábor:
„A Korpusz alapú beszédszintézis nyelvi, fonetikai kérdései”, jelen számban.

[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, Vol. 3, Numbers 3/4, December 2000, pp.201–216.

[4] Nagy András, Pesti Péter, Németh Géza, Bőhm Tamás:
„Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései”, Híradástechnika, 2005/1, pp.18–24.

[5] Olaszy Gábor:
Beszédadatbázisok készítése gépi beszédelőállításhoz. Beszédkutatás`99 (Szerk.: Gósy Mária – MTA Nyelvtudományi Intézet), Budapest 1999. pp.68–89.

[6] Olaszi Péter:
„Magyar nyelvű szöveg-beszéd átalakítás: nyelvi modellek, algoritmusok és megvalósításuk”, (Ph.D. értekezés), BME 2002.

[7] Paul Boersma:
„Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”, IFA Proceedings 17: pp.97–110.

[8] Mihajlik P., Révész T., Tatai P.:
„Phonetic Transcription in Automatic Speech Recognition”, Acta Linguistica Hungarica, Vol. 49 (3-4), 2002. pp.407–425.

[9] ITU-T:
„A method for subjective performance assessment of the quality of speech voice output devices”, Draft ITU-T, Recom. P.85, COM 12-R 6, 1993.

[10] Alvarez, Y. and M. Huckvale:
„The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems”, Proc. International Conference on Spoken Language Processing (ICSLP), Denver 2002, Vol. 1, pp.329–332.