Promptgenerátor – Ügyfélszolgálati hangos üzenetek automatikus gépi előállítása egy adott bemondó hangjára

Németh Géza, Zainkó Csaba, Fék Márk, Olaszy Gábor, Bartalis Mátyás

Budapesti Műszaki és Gazdaságtudományi Egyetem,
1117 Budapest Magyar Tudósok körútja 2., Magyarország
{nemeth,zainko,fek,olaszy,bartalis}@tmit.bme.hu

Kivonat: Az egyre szélesedő kommunikációs lehetőségekkel rohamosan nő a a telefonos ügyfélszolgálatok terhelése. A tájékoztatás automatizálásához egyre több hangos üzenetet kell elkészíteni, általában ugyanazzal a bemondóval. Ezt a felolvasó személy véges terhelhetősége korlátozza. A cikkben olyan gépi megoldás lehetőségéről számolunk be, amelyik leveszi a munka nagy részét a bemondó válláról, csak ellenőriznie kell a generált üzenet hangzását. A promptgenerátor olyan új beszédtechnológiai megoldás, amilyent még nem készítettek Magyarországon. Tervezése és fejlesztése mind számítógépes nyelvészeti, mind fonetikai és informatikai szempontból új megoldásokat eredményezett. A rendszer, optimális esetben olyan természetes hangminőséget szolgáltat, hogy a hallgató nem veszi észre, hogy gép beszél.

1 Bevezetés

Modern korunkban az információk beszéddel való közlése információs rendszerekben egyre szélesebb körben terjed el. A cikkben egy szűk témakörre vonatkozó célmegoldás kísérletéről számolunk be, nevezetesen a mobiltelefonok használatával és forgalmazásával kapcsolatos ügyfélszolgálati szövegek automatikus meghangosításával. A kutatás-fejlesztés újszerűsége abban áll, hogy a szintetikus beszédet teljesen természetes hangzási minőségben és egy adott bemondó hangszínezetével kell előállítani. A cél, hogy az üzenetet hallgató ne vegye észre, hogy gépileg előállított hangot hall. A célkitűzés megvalósításához korpusz alapú, elemkiválasztásos beszédszintetizálási módszer tűnt a legoptimálisabbnak. Ezt az általános módszert úgy alkalmaztuk a célfeladat megoldására, hogy a szintézis alapelemének a szó nagyságú elemeket választottuk.

Általánosságban elmondhatjuk, hogy ennél a technológiánál nagy, olykor több 10 órányi hangfelvételből választják ki a szintézis során a megfelelő hangrészleteket, melyek lehetnek hangok, hangkapcsolatok vagy akár szavak is. Az ilyen szintetizátorokat jellemzően egy-egy jól meghatározott témakörben lehet hatásosan elkészíteni. Ilyen terület például az ügyfélszolgálati üzenetek generálása is. Ezek a szövegek elég kötött nyelvezettel és viszonylag kötött szókinccsel rendelkeznek. Lássunk néhány promptszöveget:

Kérjük, válasszon a következő menüpontokból. Tudakozó, 1-es gomb. Hibabejelentés, 2-es gomb. Üzleti ügyintézés, 3-as gomb.
Samsung X510, X160, X660, E370, 1-es gomb. Samsung E900, E870, D900, 2-es gomb. Samsung Z400, P300, Z560, 3-as gomb.
A 3G a jelenlegi leggyorsabb adatátviteli megoldás, amely magában foglalja a jelen és a jövő szolgáltatásait. A 3G hálózat használható egyszerű telefonbeszélgetések lebonyolítására és multimédiás alkalmazásokra is, akár vonalkapcsolt, akár GPRS adatátvitel formájában.

A beszédszintetizátor fejlesztésének alaplépései a következők:

- szövegkorpusz adaptálása a szintézishez (szövegtisztítás és normalizálás),
- a szövegkorpuszból készített hangfelvétel feldolgozása (fonemikus átírás, címkék elhelyezése),
- a szintézis alapegységének kiválasztása,
- a kiválasztási és összefűzési költségfüggvények megtervezése,
- hibakeresési eljárások kidolgozása,
- a rendszer optimális működésének beállítása.

2 A beszédszintetizátor felépítése

A korpusz alapú beszédszintetizátor két fő modult tartalmaz, egy nagyméretű, megfelelően preparált szöveg és beszédkorpuszt (adatbázis), továbbá egy elemkiválasztó kereső algoritmust, amelyik a bemeneti, szintetizálandó szöveget elemezve keresést végez az adatbázisban és kiválasztja a legoptimálisabbnak tartott hullámforma elemeket és összefűzi azokat.

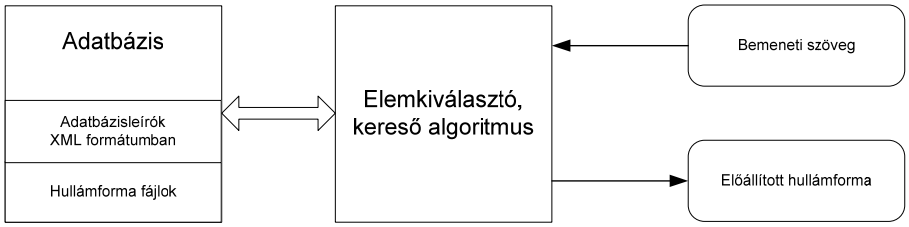

Fig. 1. A promptgenerátor beszédszintetizátor blokksémája

Az adatbázis két nagy részből áll. Egyrészt a hullámformákat tartalmazó hangfájlok gyűjteményéből, melyben minden mondat külön fájlban van tárolva, valamint az adatbázis szöveges leíróinak gyűjteményéből. Ez utóbbiakhoz tartozik a mondatok szövege, a mondatok szövegeinek fonemikus átirata, a hang- és szóhatárok pozíció jelzései a mondatokon belül, alapfrekvencia-adatok. Ezeket az információkat XML struktúrában tároljuk a rendszerben, ezzel is egységesítjük és gyorsítjuk a feldolgozást. Az elemkiválasztó algoritmus az adatbázis szöveges leírói alapján határozza

meg, hogy a bemeneti szövegnek megfelelő hangzó formát mely elemek egymás után illesztésével lehet a természetes hangzáshoz közeli minőségben előállítani.

2.1 A szintézis adatbázisa

Az adatbázis elkészítésénél több fontos szempontot vettünk figyelembe. Elsősorban a kiválasztott témakörhöz tartozó szövegtípusok, és a szintetizálandó mondat összhangját kell biztosítani. Ez esetünkben azt jelenti, hogy az ügyfélszolgálati üzenetek témakörébe tartozó szavak, mondatok tartalmát kell az adatbázisnak jól lefedni [3]. Fontos továbbá a prozódiai modellezés is (a gyakran előforduló kifejezések minden lehetséges prozódiai helyzetben előforduljanak). A cél az, hogy a szintézis során minél hosszabb beszédrészleteket találjunk meg a szintetizálandó mondat szövegéből az adatbázis szövegében és hullámformájában. A szintézishez elkészített adatbázis 3747 mondatot tartalmaz (összesen 69057 szó) szöveges és hangzó formában. Ez az állomány az elmúlt években készült, ugyanazon bemondó által felolvasott prompt-üzenetek hullámformáit tartalmazó fájlokból állt össze. Ebből készítettük el a szintézishez használható állományt, amely a megfelelő címkézéseket is tartalmazta. A címkézések a belső szinkronozást biztosítják. Az adatbázis a szövegen és a hullámformán felül tartalmazza a kettő összekapcsolását végző köztes ábrázolást, a szöveg fonemikus átiratát is. A címkéket és a fonemikus átiratot az adatbázisleíró fájl tartalmazza. A hullámformának, a fonemikus átiratnak és a címkéknek egymással szoros hangszinkronban kell lenni. Ez fontos követelmény. A szövegben nem lehetnek betűhibák, elütések, a hangban sem fordulhat elő, hogy a bemondó nem azt mondja, ami a felolvasásra készített szövegben van.

Az adatbázis elkészítésének lépései a következők:

(A1) Mondatszintű struktúra kialakítása. A szöveget és a hanganyagot mondatokra bontjuk, majd mondatonként hangszinkronba hozzuk őket egymással (a szövegben feloldjuk a rövidítéseket, a betűszavakat, a számokat stb.). Eltérés esetén (amikor a bemondó mást olvasott, mint ami a szövegben volt) a szöveget igazítjuk a hanghoz, mivel a hanganyag már nem módosítható. A számok és rövidítések feloldását előfeldolgozóval végezzük (Profivox [7] szövegfelolvasó átíró modulja) automatikus módon a manuális munka előtt. A szinkronba hozás részben manuális munkával történik, ennek eredménye, hogy a szövegben lévő betűhibák is javításra kerülnek. Ennek a munkafázisnak az eredménye, hogy olyan szövegforma áll elő, amelyben a rövidítések és számok fel vannak oldva. Ebből készül el a fonemikus átirat.

(A2) A fonemikus átirat elkészítése. A leírt szöveg nem definiálja egyértelműen a megvalósuló hangsorozatot. Figyelembe kell venni a hasonulásokat és más kiejtési sajátosságokat (hangkiesés, hangbetoldás). A fonemikus átírást a beszédfelismerőhöz tartózó átíróval készítjük (ez a felismerő jelöli ki később a hanghatárokat is).

(A3) Hullámforma előkészítés. Az adatbázisban minden mondatnak van egy hullámforma reprezentációja. A hullámformát címkékkel kell ellátni. Kétfajta alapjelölést alkalmazunk. Az egyik a hanghatár jelzése, vagyis a hang fonemikus szimbóluma és a hozzá tartozó hullámforma részlet összekapcsolása. A másik a szóhatár jelölése, vagyis a szövegben szereplő szó kezdőpontjának meghatározása. Ez a két jelölés a legfontosabb, mivel ezek alapján fogjuk kiválasztani a szintézis során az összekap-

csolandó hullámforma részleteket. Mindezeken felül címkézzük még a gerjesztési formákat (zöngés/zöngétlen), valamint a zöngés hangperiódusok kezdőpontjait (zöngejelzők). A zöngeperiódusok jelölése az alapfrekvencia pillanatnyi értékének a meghatározásához, illetve az alapfrekvencia-menet esetleges módosításához szükséges. Az alapfrekvencia pillanatnyi értékét az elemkiválasztás folyamán használjuk fel.

Az eszközrendszer a fenti címkézésekhez a következő. A zöngeperiódus-határok bejelöléséhez a Praat 4.6 fonetikai és beszéd-analizátor szoftverben implementált ablakfüggvénnyel korrigált autókorrelációs módszeren alapuló alapfrekvencia-detektálót használjuk [1]. 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 [2] segítségével oldjuk meg. A felismerőt kényszerített módban használjuk, ami azt jelenti, hogy a beszédet tartalmazó hangfájl mellett bemenetként megadjuk annak fonemikus változatát is, ami segít abban, hogy milyen hangsorozatokat keressen a felismerő. A felismerő 30 ms-os keretekkel és 10 ms-os kereteltolással dolgozik, azaz a hanghatárokat elvileg is csak 10 ms-os pontossággal határozza meg. A gyakorlatban a hanghatárjelölésekre 20 ms-os átlagos hiba a jellemző. A szóhatárok jelölését a beszédfelismerő által visszaadott, a beszédjelhez legjobban illeszkedő fonémasorozaton végezzük. A szavakon átívelő hangegybeolvadások (például: „ideig ködös”) miatt előfordulhat, hogy egy hang egyszerre két szóhoz is tartozik. Ennek kezelésére külön jelölést alkalmazunk a szavak kezdetére és végére. Így lehetővé válik a szavak közötti szóhatár-átfedést kezelése (például: „<idei<k>ödös>”, ahol a „<” jel a szó kezdetét, a „>” jel a szó végét jelöli). A szóhatárok jelölését teljesen automatikusan végezzük, olymódon, hogy a beszédfelismerő által visszaadott fonémasorozatot illesztjük a szintén a beszédfelismerő által előállított, (a szóhatároknál elágazó) összes lehetséges fonetikus átírást megadó irányított gráfhoz [4]. Az illesztést egy állapotgéppel végezzük, amely a fonémasorozat és a gráf alapján követi az elágazásokat és ennek megfelelően helyezi el a szóhatárt jelölő címkét.

(A4) Építőelemek. A szintetizátorban kétféle építőelem-típust definiáltunk, ezek a szó és a beszédhang. Az adatbázis belső címkézése illeszkedik a szintézis alapegységeihez. A szó szintű összeállítás során szavakból illetve szókapcsolatokból állítjuk össze a mondatot. Ez jó hangminőséget szolgáltat, ha elég nagy az adatbázis, és a keresési algoritmus is jól működik (figyelemmel kell lenni a prozódia helyes megvalósítására is). A szó szint továbbá gyors keresést tesz lehetővé. A beszédhang képviseli a tartalékot, vagyis azt az esetet, amikor az adatbázisban nem találjuk meg a szintetizálandó szövegrésznek megfelelő szót, szókapcsolatot. Ilyenkor hangokból, hangkapcsolatokból állítja elő az adott szót a rendszer, természetesen ennek a hangminősége gyengébb lesz, mint a szó szintű összeállításé.

(A5) Tesztelés, hibakeresés, utólagos hibajavítás emberi és gépi erővel. Az adatbázis fejlesztése során több lépcsős tesztelésre van szükség, mivel a jelölések és a hullámforma között sok esetben nincs szoros összhang. Az első ilyen szükséges teszt, amikor azt ellenőrizzük, hogy a hanghatárok közötti időtartamok mennyire jellemzőek a jelöléshez tartozó fonemikus hangra. A hanghatár-jelölés ellenőrzéséhez minden 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 és javítottuk. A tesztelés következő fázisában a spektrális tartalom és az adott hang összevetését végeztük el automatikusan. Az így feltárt hibákból kiderült, hogy sok esetben a szövegben olyan rövidítések maradtak, amelyeket nem tudott megfelelően feloldani az előfeldolgozó program. Az elemzések során feltárt hibákat manuálisan javította egy fonetikai szakképzettségű informatikus. A munka az egész adatbázisra vonatkozóan 3 hónapot vett igénybe.

2.2 A vágási pontok meghatározása

A beszédszintetizátor optimális működése szempontjából fontos, hogy olyan ponton vágjuk el a hullámformát (vegyük ki az adatbázisból), amely a legkevesebb torzítással jár a későbbi összeillesztésnél. A döntést két tényező befolyásolja: milyen hangkapcsolódások vannak az adott ponton és, hogy milyen a prozódiai szerkezet. A fizikai vágási pont kialakításához ismerni kell a beszédhangok artikulációs és spektrális belső szerkezetét, valamint tisztában kell lenni a hangkapcsolódások megvalósulásakor létrejövő hangszerkezeti és spektrális módosulások fajtáival. A vágást akkor végezhetjük sikeresen, ha tudjuk, hogy a beszédhangoknak milyen az egymásra hatása, a belső akusztikai szerkezete, hol milyen változás zajlik le a hang frekvencia-, illetve intenzitásszerkezetében a folyamatos artikuláció következtében, melyek azok a hangrészek, amelyek esetleg egymással megegyeznek, illetve nagyon hasonlóak egymáshoz. Úgy kell kiválasztani a kivágandó elemet, hogy ne sértsük meg a spektrális folyamatosság elvét. Az optimális vágási pontok a következők: a hangsor minden olyan pontja, ahol gerjesztés váltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitás minimum keletkezik), továbbá a hangok belsejében lévő néma fázisok, fojtott zönge szakaszok (ez a zár- és zár-rés hangok sajátja). Az optimális vágási pont kijelöléséhez 5-10 ms pontosságú helymeghatározásra, általában zöngeszinkron jelölésre van szükség. A hangsor összeállításánál ezek után a kivágott beszédrészletek egymáshoz való illesztését hangszerkezeti és artikulációs fonetikai szabályok alkalmazásával tehetjük torzításmentessé. Az illesztés akkor lesz sikeres, ha a beszédjelen nincs hallható akusztikai torzulás a beavatkozás után [6].

2.3 A prozódia modellezése

Az adatbázisban szereplő mondatok prozódiája adott. A szintetizálandó mondat prozódiáját meg kell jósolni. Az elemkiválasztó algoritmusnak figyelembe kell venni a jósolt prozódiát és annak megfelelő szót, szókapcsolatot kell keresni az adatbázisban. Az adatbázisban a prozódia modellezése bonyolult feladat, hiszen sok paraméter (hangsúly, ritmus, dallam, tempó stb.) jelölésére, kezelésére van szükség. A bemeneti mondat esetében pedig a prozódia jóslását kell elvégezni. Ez bonyolult nyelvi elemzéssel érhető el. Jelenleg egyik megoldásra sincsenek eszközök, ezért más utat kerestünk. Modellünk kialakítását egy mondatszerkezeti vizsgálatra épülő szópozíció meghatározásra alapoztuk. Másik kiindulási pontunk az volt, hogy a magyar kijelentő mondatokban a hangsúlyozás többnyire a frázis első tartalmas szaván van [5]. A modell lényege a következő. 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 megcímkéztünk aszerint, hogy a mondaton belül milyen pozícióban van (Me első, Mk közbenső, Mu utolsó). Ugyanilyen címkézéssel láttuk el a prozódiai egységek szavait is (Sze első, Szk közbenső, Szu utolsó). Így egy kétszintű ábrázoláshoz kötöttük a prozódiát. Példaként lássunk egy prozódiai címkékkel ellátott mondatot.

✓Me,(Sze)A (Szk)3G (Szk)hálózat (Szk)használható (Szk)egyszerű (Szk)telefonbeszélgetések (Szu)lebonyolítására Me☛✓Mk1,(Sze)és (Szk)multimédiás (Szk)alkalmazásokra (Szu)is Mk1☛,✓Mk2,(Sze)akár (Szu)vonalkapcsolt Mk2☛, ✓Mu,(Sze)akár (Szk)GPRS (Szk)adatátvitel (Szu)formájában Mu☛.

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, azonban a percepciós tesztek szerint közelíti a természetes prozódiát. A módszer előnye az egyszerűsége és gyorsasága, hátránya, hogy vannak esetek, amikor nem biztosít megfelelő prozódiát. A kísérleti rendszer jelenlegi implementációja nem tartalmaz utólagos jelfeldolgozást a prozódia simítására.

2.4 Elemkiválasztás és összefűzés

A prozódia meghatározása után, a szintézis következő 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 vezettünk be: Az **egyezési költség** megadja, hogy egy adott elem mennyire felel meg a szintetizálandó beszédszakasznak. A jelenlegi megvalósításban a beszédszakaszt az annak betűsorozatként megadott szöveges tartalma, illetve a hangsorozatként megadott fonetikus átírása határozza meg. Ehhez járulnak még a szintetizálandó szövegből meghatározott prozódiai címkék. Az **összefűzési költség** azt adja meg, hogy a leendő szomszédos elemek hangilleszkedési szempontból 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 rendszer a költségfüggvények alapján automatikusan osztályozza is a mondatok minőségét egy ötfokozatú skálán (jó minőségűnek tekinthető a 4-es osztályzat feletti). Az elemkiválasztás hierarchikusan történik. Először csak szószintű elemeket keres az elemkiválasztó. Ha a bemeneti szövegnek vannak olyan szavai, amit nem sikerült szóalapon megtalálni, akkor a hiányzó szavakat beszédhangokból rakjuk össze. A szószinten már megtalált elemeket csak abban az esetben próbálja meg a rendszer kisebb egységekből felépíteni, ha annak az egyezési költsége nagyobb, mint egy előre meghatározott érték. 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 nyolcvanban. Ha a kiválasztás folyamán egy adott elemhez tartozó jelöltek száma elérte a

nyolcvanat, 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 (szavakat illetve szófüzéreket) választja ki.

Az **egyezési költség** kialakításának paraméterei

1. A jelöltet (szó vagy beszédhang) megelőző és követő fonéma egyezése az előírt célelemet megelőző és követő beszédhanggal. A legoptimálisabb keresési eredmény a teljes egyezés. Ehhez a legkisebb az egyezési költség értéke (nulla). Arra az esetre, ha nem biztosítható a teljes egyezés, definiáltunk egymással helyettesíthető hangkategóriákat is. Az azonos kategóriába eső hangok egyezési költsége is kicsi. Abban az esetben, ha egyik korábbi eset sem áll fenn az fennmaradó elemekből választ a rendszer. Ez adja a legnagyobb költséget. A fennmaradó elemek esetére az egyezési költséget egy költségmátrix definiálja.

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 következő paraméterek szerint alakul ki:

1. Ha a vizsgált két jelölt a beszédkorpuszban egymás után következett, akkor az összefűzési költség mindig 0. Ez a legoptimálisabb eset.

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 két elem összekapcsolásakor az első elem végső és a második elem kezdő zöngés hangjából számított átlagos alapfrekvencia eltéréséből arányosan származtatunk.

Az egyes költségek értékeit számos hangminta meghallgatása során, tapasztalati úton állítottuk be. Ezek további optimalizálása még jelentősen javíthatja az előállított beszéd minőségét.

3 A prompt generátor tesztelése, az optimális működés beállítása

A fejlesztés során többször hajtottunk végre meghallgatásos tesztet. A géppel generált hangüzeneteket 1-5 közötti MOS skálán (Mean Opinion Score – átlagos szubjektív osztályzat) osztályoztuk. Az ilyen belső tesztekben 3 fő vett részt (a fejlesztői gárdából). Ezen értékelések adatait felhasználtuk a szintetizátor belső költségfüggvényeinek javításához. A fejlesztés során háromhavonta szubjektív tesztelésen, szövegesen is értékeltük a hibákat, majd hibacsoportokat állítottunk fel a következők szerint: hangerő problémák; zaj hallatszik; a hangsúlyozás nem megfelelő; kevés a szünet a tagolási helyeken; túlságosan tagoltan beszél; túl gyorsan mondja a szöveget; adatbá-

zis hiba (hanghatár rossz helyen van, rossz a hangdefiníció). A hibacsoportok elemzése alapján folyamatosan optimalizáltuk a költségfüggvények paramétereit.

A rendszer stabilitása érdekében terheléses teszteket végeztünk. Vizsgáltuk a program futási idejét, a memóriahasználatot, a futási stabilitást. A hibák és felmerülő problémák folyamatos javításával elértük, hogy a rendszer stabilitása fokozatosan növekedett. Jelen állapotában megbízhatóan működik folyamatos felhasználói tesztelésre alkalmas.

A rendszerben a komponensek folyamatosan naplózzák az egyes eseményeket. Ilyen naplóbejegyzések pl.:

- A szintetizálandó prompt szövege.
- A szintetizálandó prompt szövegének fonetikus átirata.
- A szintetizálandó prompt szövegének megfelelő korpuszelemek, melyek közül a program válogat.
- A végleges korpuszelemek, melyből a prompt hangzó formája előáll.
- A program működésével kapcsolatos bejegyzések (indítás, leállítás, beállítások…).

A részletes naplófájlok használata a fejlesztési szakaszban és utána is megkönnyíti a hibák keresését, illetve azok okának kiderítését, ezáltal javításukat is.

A rendszer kész állapotú hangminőségének ellenőrzésére 120 prompt (312 mondat) automatikus generálását végeztük el. A költségfüggvények alapján automatikusan adott osztályzatok átlaga 4,12. A meghallgatások során 3 személy értékelte ugyanezeket a mondatokat, ítéleteik átlaga 4,02 volt. Ez azt mutatja, hogy a költségfüggvények alapján adott gépi értékelés jól közelíti a szintetizált mondat hangminőségének jellemzését.

4 Összegzés

A beszédszintézissel szemben támasztott új követelmények eredménye a bemutatott beszédtechnológiai megoldás. A kutatás-fejlesztés újszerűsége abban áll, hogy a szintetikus beszédet teljesen természetes hangzási minőségben és egy adott bemondó hangszínén kell előállítani adott témájú, ügyfélszolgálati üzenetek szövegeiből kiindulva. A célkitűzés megvalósításához a korpusz alapú, elemkiválasztásos beszédszintetizálási módszert választottuk. Magyarországon ez az első ilyen beszédelőállító rendszer. Az eredmények azt mutatják, hogy ilyen technológiával, optimális esetben el lehet érni a követelményként megadott hangminőséget. A rendszer legérzékenyebb pontja az elemkiválasztás költségfüggvényének a behangolása. Valószínűsítjük, hogy az optimálisabb működéshez több-szintű költségfüggvény kombinációt kell majd alkalmazni.

Bibliográfia

1. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, IFA Proceedings 17: 97–110
2. Mihajlik, P., Révész, T., Tatai, P.:Phonetic Transcription in Automatic Speech Recognition, Acta Linguistica Hungarica, Vol. 49 (3-4), (2002) 407–425
3. Nagy, A., Pesti, P., Németh, G., Bőhm, T.: Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései, Híradástechnika, (2005/1) 18–24
4. Németh, G. – Olaszy, G. – Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás - 2006. MTA Nyelvtudományi Intézet, (2006) 183–196
5. Olaszy, G.: A Korpusz alapú beszédszintézis nyelvi, fonetikai kérdései, Híradástechnika (2006/3) 43–50
6. Olaszy, G.: Az artikuláció akusztikai vetülete, a hangsebészet elmélete és gyakorlata. Kif-LAF 2003. Szerk.: Hunyadi László. Debreceni Egyetem, (2003) 241–254
7. 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, (2000) 201–216

Ezt a kutatás-fejlesztést az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.