REJTETT MARKOV-MODELL ALKALMAZÁSA MAGYAR NYELVŰ GÉPI SZÖVEGFELOLVASÁSHOZ

Tóth Bálint – Németh Géza

Bevezetés

Napjainkban már számos automatikus szövegfelolvasási módszer létezik: a beszédkeltés mechanizmusát modellező formánsszintézistől kezdve a diádos és triádos hullámforma-összefűzéses szintézisen át az elemkiválasztó (korpusz) szintézisig. A beszédszintetizátor által kiadott hangot érthetőség és természetesség szempontjából szokták minősíteni, magát a megoldást pedig műszaki paraméterekkel jellemzik. A formánsalapú szintetizátorok érthetősége jó, viszont hangzása nem természetes. A hullámforma-összefűzéses technológia kissé jobb eredményt nyújt, a legjobban érthető és a legtermészetesebbnek ítélt beszédet a korpuszalapú elemkiválasztásos módszerrel érték el. A műszaki jellemzők szoros kapcsolatban vannak a generált hang minőségével. Például, a legjobb minőséget nyújtó korpuszalapú szövegfelolvasó rendszerek adatbázisának a mérete igen nagy (gigabyte-os nagyságrendbe esik), az elemkiválasztás is sok számítási kapacitást igényel. A beszélő hangját az adatbázis meghatározza, azon változtatni nem lehet, viszont az adott hangon a mesterségesen előállított hangszínezet egészen természetes. A formánsszintetizátorok kisméretűek (2–10 kilobyte) és gyorsak, tehát olyan helyen is használhatók, ahol kis memóriakapacitás áll rendelkezésre. A technológia megengedi, hogy többféle hangon is megszólalhatnak. A hullámforma-összefűzéses eljárás ma a leggyakrabban alkalmazott módszer. Memóriaigénye viszonylag kicsi (2–20 megabyte), ezért sokféle gyakorlati alkalmazásban használják. Az ilyen rendszerek hangszínezete már közelít az emberihez, bizonyos korlátok között változtatható is. A jelen tanulmány olyan technológiát ismertet, amellyel növelni lehet a generált beszéd természetességét, és érthetősége is magas. Ilyen szempontból a tanulmányban bemutatott magyar nyelvű rejtett Markov-modellekkel (Hidden Markov Model, HMM) megvalósított beszédszintetizátor a hullámforma-összefűzéses eljárás és a korpuszalapú szintézis közötti helyet foglalhatja el. A beszédépítési eljárása lényegesen különbözik az előbbi technológiáktól, mivel nem közvetlenül a hullámformával dolgozik, hanem a hullámformából spektrális és prozódiai jellemzők sokaságának kinyerése után (tanító fázis) ezekből válogatva alakítja ki a szintézishez szükséges adatsorozatot. A válogatást rejtett Markov-modellek végzik. A HMM-alapú beszédszintézis számos előnye miatt lehet a jövő sikeres mesterséges beszédkeltési technológiája, mivel kisméretű (1,5–2 mega-

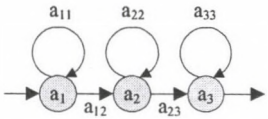
byte) beszédadatbázisból képes jó minőségű, érthető beszédet előállítani, amely hordozza a beszélő hangszínezeti tulajdonságait is (Yoshimura 2002). Magyar nyelvre eddig még nem alkalmazták az eljárást, az első, kísérleti eredményekről számolunk be.

A rejtett Markov-modell

A Markov-láncokat gyakran használják fizikai folyamatok modellezésére, ahol különböző megfigyelések alapján kell a folyamatot szimulálni. Ha egy adott megfigyelés egyértelműen azonosítja, hogy a folyamat milyen állapotban van, akkor a használt modellt Markov-láncnak nevezzük. Azonban számos olyan folyamat létezik, mint a beszéd is, mely esetben az állapotok jól definiálhatók, de rájuk a megfigyelések alapján mégsem tudunk egyértelműen következtetni. Ezeket modellezzük úgynevezett rejtett Markov-modellekkel. A beszédtechnológiában használt rejtett Markov-modell alkalmazásakor a beszédre jellemző, abból kinyert paramétereket kell eltárolni, mely jelentősen hatékonyabb, mint a hangmintaalapú rendszerek esetén a minták tárolása, hiszen a paraméterek jóval kevesebb helyet foglalnak el, mint az eredeti hullámformák. A paraméterek (például spektrális jellemzők) kinyeréséhez úgynevezett akusztikus modelleket alkalmaznak. Régebben fonémánkénti (ún. monofón) akusztikus modellt alkalmaztak, legújabban már a hangkörnyezetet is figyelembe vevő akusztikus modellek (pl. hanghármasok, ún. trifónok) a leggyakoribbak (Mihajlik et al. 2006).

Napjainkban a beszédtechnológia területén a rejtett Markov-modelleket nagyrészt a beszédfelismerésben használják. A modell működését egy egyszerű példán keresztül mutatjuk be. A szavakat úgy tekintjük, hogy azok beszédhangok sorozataként állnak elő. Minden beszédhangra három állapotot feltételezünk: a hang eleje, közepe, vége. Az egyes állapotok között és az egyes állapotokból saját magukra mutató, úgynevezett élek határozzák meg, hogy az adott állapotból mely következő állapotokba lehet lépni (1. ábra). Az ábrán az a_1, a_2, a_3 jelöli a beszédhang három belső állapotát, az a_{12}, a_{23} élek a három belső állapot közötti átmeneti valószínűségeket, végül az a_{11}, a_{22}, a_{33} pedig azt jelzi, hogy milyen valószínűséggel maradunk az adott belső állapotban. A modell betanítása során az élekhez valószínűségek rendelhetők, melyek a helyben maradás (a_{11}, a_{22}, a_{33}), illetve továbblépés (a_{12}, a_{23}) valószínűségét határozzák meg.

Az egyes állapotok tartalmazzák az akusztikus modellek készítése során becsült sokdimenziós Gauss-eloszlások paramétereit. Általában egy adott környezetben lévő beszédhang többször előfordul a tanító adatbázisban, a tanítás során pedig az ehhez tartozó spektrális paraméterhalmazt próbáljuk becsülni Gauss-eloszlással. A mintaillesztő eljárás ezen akusztikus modellekhez illeszti a bejövő paramétereket, hogy eldöntse, megegyezik-e az a felismerendő szóval. A témakörről részletesen olvashatunk például Lawrence Rabiner (1989) klasszikus cikkében.



1. ábra
Három állapotú rejtett Markov-modell sematikus ábrája egy beszédhang három belső állapotának leírására

A rejtett Markov-modell alkalmazása a beszédszintézis területén az elmúlt évtizedben merült fel. Az erre kidolgozott eljárás két lényegi ponton tér el a beszédfelismerésre kidolgozott megoldástól. A legjelentősebb különbség az, hogy a két eljárás esetében a bemeneti és a kimeneti paraméterek felcserélődnek, tehát a végső lépésnél a mintaillesztés helyett (amikor felismerésnél a beszédhangokat jellemző paramétereket hasonlítjuk össze az akusztikus modell paramétereivel, tehát a hangok felismerésére adunk javaslatot) mintaválogatást hajtunk végre (azt jósoljuk, hogy melyik a legjobban jellemző mintasorozat a felolvasandó beszédhangra), majd a kiválasztott jellemző paraméterhalmazból a modell egy beszédkódoló eljárással beszédhangot állít elő, és így jön létre a szintetizált beszédhullám. A második fontos különbség, hogy a prozódia jellemző komponenseit (például hangmagasság, hangidőtartam) is modellezni kell a beszédszintézis esetében, és ezeket a feladatokat szintén végezhetik rejtett Markov-modellek.

A jelen kutatásban a spektrális és az alapfrekvencia-paramétereket, valamint az állapot-időtartamokat is HMM-ekkel modellezzük.

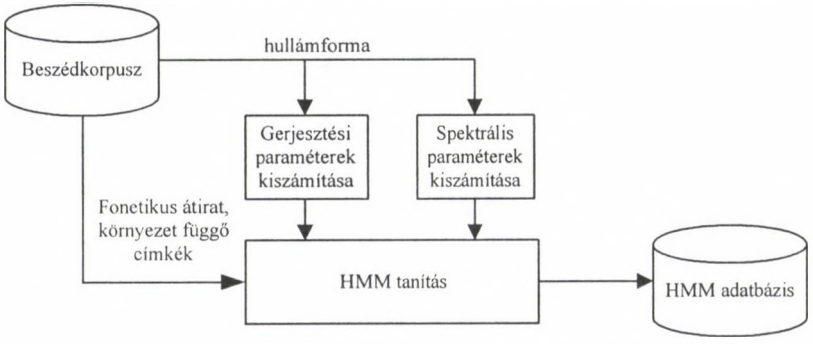
A HMM-alapú beszédszintézis

A HMM-alapú beszédszintetizátor elkészítése két részre bontható: a tanulási és a szintetizálási fázisra. A tanulási fázisban készítjük fel a szintetizátort a későbbi működésre. A tanulás során (2. ábra) tanítjuk be a rejtett Markov-modelleket egy nagy, gondosan megtervezett és felcímkézett beszédadatbázis segítségével. A tanítási folyamat végére egy kisméretű HMM modelladatbázis áll elő, amelyben a betanított beszédkorpuszra jellemző HMM-paraméterek találhatóak. Ezekből válogatja majd ki a szintetizátor a szintetikus beszéd generálásához szükséges paramétereket.

A működési fázisban már csak a tanítás eredményét használjuk, tehát ez a rész a tényleges beszédszintetizátor. A bemeneti szöveg alapján meghatározzuk, hogy milyen hangsorozatot kell generálni, és a HMM modelladatbázisban tárolt paraméterekből kiválogatjuk azt a paramétersorozatot, amelyik legjobban reprezentálja az előállítani kívánt hangsorozatot. Ezekből állítjuk vissza a spektrális jellemzőket, az időtartamokat, a szüneteket és az alapfrekvenciát, majd ezek alapján a beszédkódoló eljárással a szintetizált hullámformát.

A HMM-ek tanítása

A tanításhoz több órányi beszédet tartalmazó, nagyméretű beszédkorpuszt kell használni. A tanítás típusától függően ez lehet egy vagy több beszélőtől.



2. ábra
A HMM-alapú szövegfelolvasó tanítása

A korpusz tartalma a következő: a hullámforma digitalizált változata, a felolvasott szöveg fonetikai átirata és a hang- és szóhatárok pontos pozíciója. A feldolgozás egysége a mondat. A hullámformából 25 ms-os ablakolással mintákat veszünk, és azok paramétereit tároljuk, összesítjük, optimalizáljuk. Minden ablakolt részből kinyert valamennyi paraméterhez tartozik legalább egy HMM. A HMM-alapú beszédszintetizátort egyszer kell tanítani, mely eredményeként egy kisméretű paraméteradatbázis áll elő (HMM modelladatbázis), amelyet a beszéd szintetizálásakor fog használni a rendszer.

A rejtett Markov-modellek tanításához a beszédkorpuszból származtatott paraméterek sokaságára van szükség. Ezek a következők: a hullámforma spektrális tartalmára utaló, úgynevezett MFCC (Mel Frequency Cepstrum Coefficients) adatok, ezek első és második deriváltjai, továbbá az alapfrekvencia (F_0), valamint annak első és második deriváltjai.

Ezen túl még szükség van a beszédkorpusz fonetikai átiratából képzett hangkörnyezetfüggő címkékre. A környezetfüggő címkék írják le egy adott beszédhang környezetét (pl. előtte és utána következő beszédhangok, hangsúlyos szótagok jelölése, szótagszámok stb.). A környezetfüggő címkékről bővebben alább szólunk. A fenti paramétereket jelfeldolgozási és matematikai szoftverekkel automatikusan lehet előállítani. A tanító beszédadatbázis címkézési pontosságát célszerű kézi módszerekkel is ellenőrizni.

Amennyiben minden szükséges adat a rendelkezésünkre áll, elkezdődhet a HMM-ek tanítása. A tanítás célja, hogy az egyes hangokhoz és környezetükhöz rendelt paraméterek segítségével előállítsunk egy azokat minél pontosabban becslő függvényt (esetünkben Gauss-eloszlást használtunk). Ezen adato-

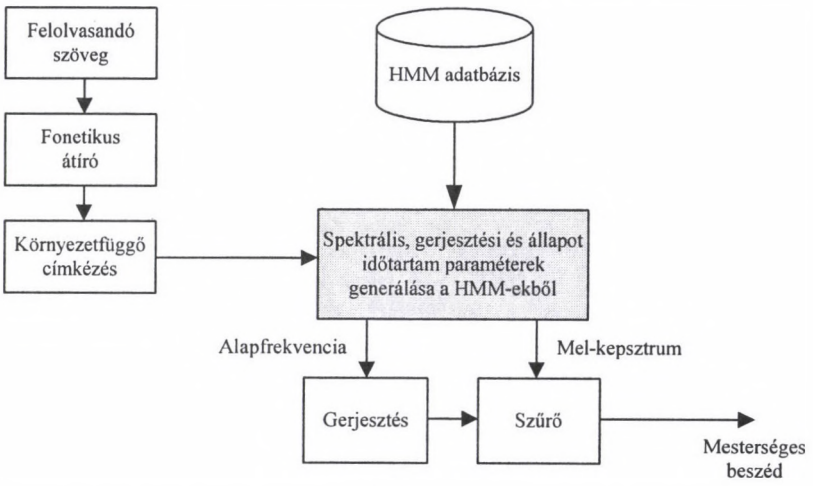
kat tároljuk el a HMM modelladatbázisban, melynek mérete (kb. 1-2 megabyte) nagyságrendekkel kisebb az eredeti beszédkorpusz méreténél (1-2 gigabyte).

A betanításhoz kétfajta módszert lehet használni. Betaníthatjuk a HMM-eket egy adott beszélőtől származó 2–4 órás beszédadatbázissal (ilyenkor erre a személyre jellemző hangszínezettel és stílussal beszél a szövegfelolvasó), illetve betaníthatjuk több beszélőtől gyűjtött adatbázisokkal (beszélőnként 1-1,5 óra hanganyag, minimum 3-4 különböző beszélő hangja). A tanítás után ekkor még szükséges a beszédhangot egy konkrét bemondónak a hangjára adaptálni. A személyre szabást egy 5–8 perces, tetszőleges személytől felvett beszédadatbázissal végezhetjük (Masuko 1997, Tamura 2001). Ezután ennek a személynek a hangjához hasonló hangszínnel és stílussal fog megszólalni a rendszer. Ezen túl még számos módszer létezik a beszédhang jellemzőinek a megváltoztatására (Yoshimura 1997, Tachibana 2005).

Beszéd előállítása a betanított HMM-ek segítségével

A beszéd előállítása során első lépésként elkészítjük a szöveg fonetikai átiratát környezetfüggő címkékkel (lásd alább), majd kinyerjük a várható hangidőtartamokat az állapot-időtartam valószínűség sűrűségfüggvényekből, ezután pedig a legvalószínűbb spektrális és gerjesztési paramétereket nyerjük ki a HMM modelladatbázisból. Ezen paraméterek alapján állítjuk elő a mesterséges beszédet beszédkódoló eljárás segítségével (Imai 1983, Maia 2007).

A HMM alapú beszédszintetizátor felépítését a 3. ábra mutatja.



3. ábra
A HMM-alapú szövegfelolvasó beszéd-előállítási mechanizmusa

Magyar nyelvű adaptáció

A kísérleteket a HTS keretrendszer segítségével végeztük el (Zen et al. 2007), amely egy szabadon felhasználható, programozható HMM-rendszer. A magyar nyelvű változat elkészítéséhez a következő fő elemekre és adatokra volt szükség: nagyméretű beszédadatbázis címkékkel ellátva, a beszédadatbázis mondatainak fonetikai átirata, egy hangkörnyezetfüggő címkéző eszköz és a magyar nyelvre jellemző döntési fák elkészítéséhez szükséges kérdések összeállítása. A következő pontokban áttekintjük a magyar változat létrehozásának ezen lépéseit.

A magyar nyelvű adaptáció során a következő beszédhangokat definiáltuk (a hangokat itt helyesírási betűjelükkel adjuk meg): magánhangzók: *a, á, o, ó, u, ú, ü, ű i, í, e, é, ö, ő;* rövid és hosszú mássalhangzók: *b, d, gy, g, p, t, ty, k, m, n, ny, j, h, v, f, z, sz, c, dz, zs, s, cs, dzs, l, r.* A hanghosszúságot a mássalhangzóknál az időparaméterrel különböztetjük meg.

A beszédkorpusz előkészítése

A tanításhoz 600 időjárás-jelentést tartalmazó mondatot használtunk, melyeket professzionális női bemondó olvasott fel (2 óra hanganyag). A mondatokat 16 000 Hz-es mintavételezéssel, 16 bites felbontással tároltuk. A mondatok fonetikus átiratát elkészítettük, és a hang- és szóhatárokat bejelöltük először automatikus módszerekkel (Mihajlik et al. 2003), majd ezeken félautomatikus módszerrel finomítottunk (Olaszy–Bartalis 2008).

Környezetfüggő címkézés

Annak érdekében, hogy a HMM-ek a legmegfelelőbb elemeket válasszák majd ki a beszéd előállítás során, számos fonetikai jellemzőt adunk meg paraméterként. A jellemzőket minden egyes beszédhangra kiszámoljuk. A legfontosabb jellemzőket az 1. táblázat foglalja össze.

A szótagokat a szótagmagok alapján keressük, számoljuk és jelöljük, tehát nem a nyelvi szótagolási szabályokat vesszük figyelembe. A szótaghatárokat saját programmal állapítjuk meg. A programban a *A magyar helyesírás szabályai* (Akadémiai Kiadó 1985) című kiadvány *A szótagolás szerinti elválasztás* szabályait algoritmizáltuk.

A táblázatban megadott jellemzőket nevezzük címkéknek. Tehát a beszédadatbázisban található összes mondat összes hangjához minden paramétert ki kell számolni, amely az 1. táblázatban található. Egy hanghoz összesen 38 környezetfüggő címkét rendelünk, így például egy 100 hangból álló mondathoz 3800 címkére van szükségünk. A címkézést automatikusan végezzük, mely néhány esetben (pl. hangsúlyos szótagok meghatározása) hibás lehet. A Profivox címkézője 70%-os pontosságú, nyelvimodell-alapú kísérletek sem adnak jobb eredményt (Tamm–Olaszy 2005) megfelelő nyelvi modell hiányában. Ez azonban nem okoz jelentős problémát, hiszen a beszéd előállításakor is ugyanazt az algoritmust használjuk, így a HMM hibás címkézés esetén is következetesen fogja az adott hangoknak megfelelő paramétereket kiválasztani.

1. táblázat: A környezetfüggő címkézéshez használt jellemzők

Hangok	– Az aktuális hang, valamint a megelőző és követő két-két hang (kvinfón). A szüneteket is hangként jelöljük.
Szótag	– A szótaghangsúlyok jelölése (hangsúlyos/hangsúlytalan) az aktuális, az előző, és a követő szótagban. Erre a Profivox-rendszer hangsúly-meghatározó algoritmusát használtuk (Olaszy et al. 2000). – A beszédhangok száma az aktuális/előző/következő szótagban. – A szótagok száma az előző/következő hangsúlyos szótagtól/szótagig. – A szótag magánhangzója.
Szó	– A szótagok száma az aktuális/előző/következő szóban. – Az aktuális szó pozíciója a mondatrészben (elölről és hátulról is számítva). Mondatrésznek tekintünk két vessző, pontosvessző, gondolatjel közötti szövegrészt a mondatban.
Mondatrész (két írásjel közötti szakasz)	– A szótagok és szavak száma az aktuális/előző/következő mondatrészben. – Az aktuális mondatrész pozíciója a mondatban (elölről és hátulról is számítva).
Mondat	– A szótagok száma az adott mondatban. – A szavak száma az adott mondatban. – A mondatrészek száma az adott mondatban.

Döntési fák

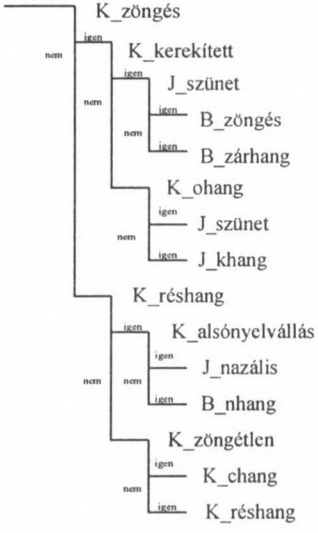
Az imént láthattuk, hogy számos környezetfüggő tulajdonság létezik, melyek összes lehetséges kombinációja óriási szám. Ha csupán a kvinfónok lehetséges változatait számoljuk meg, az is több mint 160 millió, de ezt a számot a többi környezetfüggő tulajdonság még exponenciálisan növeli. Ezért lehetetlen egy olyan, adott nyelvre jellemző, a tanításhoz szükséges beszédadatbázis-szöveget előállítani, melyben minden lehetséges kombináció szerepel. E probléma leküzdése érdekében be kellett vezetni a döntésifa-alapú klaszterezést (Odell 1995, Shinoda–Watanabe 2000), hogy a beszédszintézis során a HMM modelladatbázisban az adott környezetfüggő címkékhez legjobban hasonlító elemeket válassza ki a rendszer. Mivel a különböző tulajdonságok hatnak mind a spektrális, mind az alapfrekvencia-paraméterekre és az állapot-időtartamokra is, ezért ezeket külön-külön csoportokra bontottuk, így összesen ez a háromféle klaszterezés van jelen a paraméterek kiválasztásánál. A 2. táblázat mutatja, hogy milyen, a magyar nyelvre jellemző tulajdonságokat (Gósy 2004) használtunk fel a döntési fák építésekor.

A 4. ábrán egy példát láthatunk a spektrális paraméterekre vonatkozó döntési fára. A döntési fákat a program automatikusan készíti el. A döntési fa mindig a legelőnyösebb csoportra bontást próbálja megvalósítani. A K_ előtag azt je-

lenti, hogy a szabály az épp középső hangra igaz, a J_ előtag a középső hangot követőre (jobbra lévőre), a B_ előtag pedig a középső hangot megelőzőre (tőle balra lévőre) vonatkozik. A bemutatott példában azt láthatjuk, hogy a középső hang zöngés-zöngétlen csoportokra bontása volt a legelőnyösebb (a K_zöngés tulajdonság került a döntési fa legfelsőbb szintjére).

2. táblázat: A döntési fák építéséhez használt jellemzők

Beszédhangok	– Magánhangzó/mássalhangzó. – Zöngés/zöngétlen. – Rövid/hosszú. – A képzés helye a magánhangzóknál (hátul, középen, elöl). – Nyelvállás a magánhangzókra (felső, középső, alsó). – Ajakállás a magánhangzókra (kerekített, nem kerekített). – A képzés módja mássalhangzóknál (zárhang/réshang/zár-rés hang/pergőhang/nazális, közelítőhang).
Szótag	– Hangsúlyos/hangsúlytalan. – Az adott szótagra vonatkozó számszerű adatok (lásd 1. táblázat).
Szó	– Az adott szóra vonatkozó számszerű adatok (lásd 1. táblázat).
Mondatrész	– Az adott mondatrészre vonatkozó számszerű adatok (lásd 1. táblázat).
Mondat	– Az adott mondatra vonatkozó számszerű adatok (lásd 1. táblázat).



4. ábra
A spektrális paraméterekhez tartozó egy lehetséges döntési fa

Meghallgatásos teszt

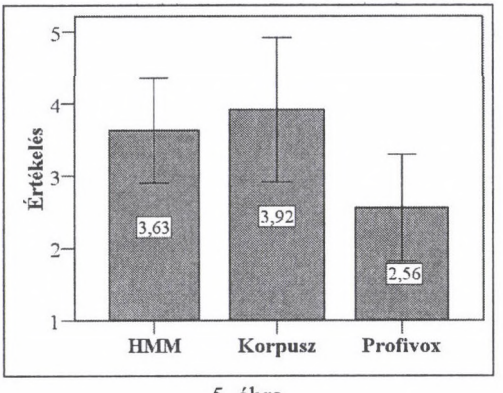
Annak érdekében, hogy objektíven tudjuk értékelni a magyar nyelvű HMM-alapú beszédszintézis minőségét egy MOS (Mean Opinion Score) meghallgatásos tesztet készítettünk el. A tesztben három beszédszintetizáló rendszer vett részt, egy triádokkal működő, egy korpuszos és a HMM-alapú szövegfelolvasó.

A teszt elején minden rendszertől 3-3 mondatot játszottunk le véletlenszerűen, amelyeket a tesztalanyoknak még nem kellett értékelni. Ez azt a célt szolgálta, hogy az alanyok hozzászokjanak a mesterséges hangokhoz, és hallják előre, hogy nagyjából milyen minőségre számíthatnak.

Ezután minden rendszer mintáiból 29 mondatot játszottunk le, minden tesztalany esetén más-más sorrendben, így zárva ki az esetleges „memóriahatásokat" (Santen 1993). A tesztmondatok tartalma időjárás-jelentés volt. A triádalapú rendszer kötetlen témakör szintézisére készült. A HMM-rendszer időjárás-jelentés tartalmú mondatokkal volt tanítva, illetve a korpuszos rendszer adatbázisa is időjárás-jelentéseket tartalmazott. Minden rendszerrel ugyanazt a 29 mondatot generáltuk, de egyik rendszer esetén sem szerepeltek ezek a mondatok az adatbázisban. A tesztalanyok a mondatokat egytől ötig értékelhették (egy volt a legrosszabb, öt a legjobb).

A meghallgatásos tesztet 12 tesztalany végezte el.

Az eredményt a 5. ábra mutatja.



5. ábra

A MOS meghallgatásos teszt eredménye az átlag és szórás értékekkel

A tudomány mai állása szerint a korpusz elvű beszédszintézis-módszer képviseli a legjobb hangminőséget kötött témakörre, ez tükröződik az eredményben is. A HMM-alapú rendszer megközelítette a korpuszost, tehát a beszédminősége szintén nagyon jónak mondható. A különbség a két rendszer között a műszaki paraméterekben viszont számottevő. A korpuszos rendszer

adatbázisa közel 11 órányi hanganyagot tartalmaz, míg a HMM-ek tanításához elegendő volt 1,5 órányi hanganyag, és tanítás után a HMM-szövegfelolvasó esetén az adatbázis mérete 2 megabyte alatt marad (szemben a korpuszos rendszer több mint egy gigabyte-os adatbázisával). A triádalapú rendszer általános témakörök lefedésére készült, semmilyen témakör-specifikus információ nem került bele. Ez is magyarázhatja az alacsonyabb értékelést. Az eredmények abszolút értéke kevésbé mérvadó, inkább az egymáshoz viszonyított arányok hordoznak érdemi információt.

Jövőbeli tervek

A jelen tanulmány a magyar nyelvű, HMM-alapú mesterséges beszédkeltés első kísérleti változatát ismertette. A jövőben számos továbbfejlesztési irányt tűztünk ki célul, melyek közül első lépésként az adaptív tanításhoz szeretnénk további beszédkorpuszokat rögzíteni, így érve el természetesebb hangzást, továbbá ezáltal lehetőségünk nyílik kis (5-8 perces) adatbázisok segítségével új beszédhangokat és érzelmeket betanítani (Krstulovic et al. 2007) a rendszerrel.

A kis adatbázisméret előnyei és a jó minőségű beszédhang miatt szeretnénk a rendszert mobil eszközökön is megvalósítani. Ennek érdekében, amennyiben a jelenlegi rendszer nem képes valós idejű működésre a limitált platformon, optimalizálni fogjuk az eljárást mobil környezetre, és mérésekkel meghatározzunk azt a minimális hardverigényt, amely szükséges a megfelelő működéshez.

Összefoglalás

A tanulmányban bemutattuk a rejtett Markov-modell alapú szintézis működésének az elvét, a magyar változat létrehozásának a lépéseit és az első magyar HMM-alapú beszédkeltéssel kapcsolatos kísérlet végén végzett meghallgatásos teszt eredményeit. A HMM-alapú szövegfelolvasó rendszerek igazi előnye, hogy kis adatbázis méretek mellett képesek jó minőségű beszédhangot előállítani. Az eljárás alkalmas arra is, hogy az eredeti hang karakterét megváltoztassuk más beszélő hangjára, illetve érzelmes beszédet is előállítsunk. Célunk, hogy ipari alkalmazásokban is használható magyar nyelven beszélő szövegfelolvasó rendszerré fejlesszük tovább a jelenlegi kísérleti rendszert.

Néhány HMM alapú szintézissel készített mondat meghallgatható a következő Internet címen: http://speechlab.tmit.bme.hu/hmm/

Irodalom

A magyar helyesírás szabályai 1985. 11. kiadás. Akadémiai Kiadó, Budapest. http://mek.oszk.hu/01500/01547/index.phtml (A letöltés ideje: 2008. július 3.)

Gósy Mária 2004. *Fonetika, a beszéd tudománya.* Osiris Kiadó, Budapest.

Imai, Satoshi 1983. Cepstral analysis synthesis on the mel frequency scale. *Proceedings of ICASSP 1983.* 93–96.

Krstulovic, Sacha – Hunecke, Anna – Schröder, Marc 2007. An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In: *Proceedings of Interspeech 2007.* 1897–1900.

Maia, Ranniery –Toda, Tomoki – Zen, Heiga – Nankaku, Yoshihiko – Tokuda, Keiichi 2007. A trainable excitation model for HMM-based speech synthesis. In: *Proceedings of Interspeech 2007.* 1909–1912.

Masuko, Takashi – Tokuda, Keiichi – Kobayashi, Takao – Imai, Satoshi 1997. Voice characteristics conversion for HMM-based speech synthesis system. In: *Proceedings of ICASSP 1997.* 1611–1614.

Mihajlik, Péter – Révész, Tibor – Tatai, Péter 2003. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica* 49/3–4. 407–425.

Mihajlik Péter – Fegyó Tibor – Tatai Péter 2006. Új eljárás a gépi beszédfelismerés környezetfüggő beszédhangmodelljeinek kialakítására. *Beszédkutatás 2006.* 218–230.

Odell, Julian James 1995. *The use of context in large vocabulary speech recognition.* PhD thesis. Cambridge University, Cambridge.

Olaszy Gábor – Bartalis Mátyás 2008. Jelfeldolgozási algoritmusok kombinációja a gépi hanghatárjelölés javítására. *Beszédkutatás 2008.* 208–220.

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.

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE.* 257–286.

van Santen, Jan P. H. 1993. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language* 7. 49–100.

Shinoda, Koichi – Watanabe, Takao 2000. MDL-based context-dependent subword modeling for speech recognition. *Journal of Acoustical Society of Japan (E)* 21/2. 79–86.

Tachibana, Makato – Yamagishi, Junichi – Takashi, Masuko – Kobayashi, Takao 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions Information and Systems* 88/11. 2484–2491.

Tamm Anne – Olaszy Gábor 2005. Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. In Alexin Zoltán – Csendes Dóra (szerk.): *III. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Szeged, 383–393.

Tamura, Masatsune – Masuko, Takashi – Tokuda, Keiichi – Kobayashi, Takao 2001. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In: *Proceedings of ICASSP 2001.* 805–808.

Yoshimura, Takayoshi – Tokuda, Keiichi – Masuko, Takashi – Kobayashi, Takao – Kitamura, Tadashi 1997. Speaker interpolation in HMM-based speech synthesis system. In: *Proceedings of Eurospeech 1997.* 2523–2526.

Yoshimura, Takayoshi 2002. *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems.* PhD thesis. Nagoya Institute of Technology, Nagoya.

Zen, Heiga – Nose, Takashi – Yamagishi, Junichi – Sako, Shinji – Masuko, Takashi – Black, Alan W. – Tokuda, Keiichi 2007. The HMM-based speech synthesis system version 2.0. In: *Proceedings of ISCA SSW6.* Bonn. (CD)

Ezúton szeretnénk kiemelten köszönetet mondani Olaszy Gábornak szakmai tanácsaiért. Köszönjük a szubjektív kiértékelésben részt vevő tesztelőknek aktivitásukat. Külön köszönet illeti Bartalis Mátyást a web-es tesztfelület elkészítéséért és Mihajlik Pétert a magyar nyelvű beszédfelismerő eszközök használatához nyújtott segítségéért. A kutatást az NKTH a NAP projekt keretében (OMFB-00736/2005) részben támogatta.