Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel

Tóth Bálint, Németh Géza

Távközlési és Médiainformatikai Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem
1117 Budapest, Magyar Tudósok krt. 2.
{toth.b, nemeth}@tmit.bme.hu

Kivonat: Napjainkban számos automatikus szövegfelolvasási módszer létezik, de az elmúlt években a legnagyobb figyelmet a statisztikai parametrikus beszédkeltési módszer, ezen belül is a rejtett Markov-modell (Hidden Markov Model, HMM) alapú szövegfelolvasás kapta. A HMM-alapú szövegfelolvasás minősége megközelíti a manapság legjobbnak számító elemkiválasztásos szintézisét, és ezen túl számos előnnyel rendelkezik: adatbázisa kevés helyet foglal el, lehetséges új hangokat külön felvételek nélkül létrehozni, érzelmeket kifejezni vele, és már néhány mondatnyi felvétel esetén is lehetséges az adott beszélő hangkarakterét visszaadni. Jelen cikkben bemutatjuk a HMM-alapú beszédkeltés alapjait, a beszélőadaptációjának lehetőségeit, a magyar nyelvre elkészült beszélőfüggetlen HMM adatbázist és a beszélőadaptáció folyamatát félig spontán magyar beszéd esetén. Az eredmények kiértékelése céljából meghallgatásos tesztet végzünk négy különböző hang adaptációja esetén, melyeket szintén ismertetünk a cikkünkben.

1 Bevezetés

Napjainkban már számos lehetőség áll rendelkezésre gépi szövegfelolvasásra: a beszédkeltés mechanizmusát modellező formáns és artikulációs szintézistől kezdve a diádos és triádos hullámforma összefűzéses szintézisen át a hullámforma-elemkiválasztó (korpusz) szintézisig. A beszédszintetizátor által kiadott hangot érthetőség és természetesség szempontjából szokták minősíteni, a technológiai megoldást pedig olyan további műszaki paraméterekkel jellemzik, mint például számításigény, tárhely igény. Napjaink vezető technológiája a korpusz alapú hullámforma-elemkiválasztásos módszer, azonban adatbázisának a mérete igen nagy (gigabyte-os nagyságrendbe esik), az elemkiválasztás sok számítási kapacitást igényel és a beszélő hangkarakterét az adatbázis határozza meg. Így új beszédhangokhoz új, több gigabyte-os stúdióminőségű hangfelvételek vagy beszélő transzformációs eljárások szükségesek, melyek minőségromláshoz vezetnek.

A statisztikai parametrikus szintézis, ezen belül is a beszédfelismerő rendszerek technológiáját használó rejtett Markov-modell (Hidden Markov Model, HMM) alapú beszédszintézis [1] igen jelentős fejődésen ment keresztül az elmúlt években. Az általa generált beszéd minősége és természetessége megközelíti a korpuszos rendsze-

rek minőségét, de emellett számos előnnyel rendelkezik: a futáshoz szükséges adatbázis mérete kicsi (néhány megabyte) [2], könnyen lehet vele új beszédhangokat létrehozni [3], alkalmas érzelemkifejezésre [4] és beszélőadaptációra [5], [6]. A HMM-alapú beszédszintézis beszédépítési eljárása lényegesen különbözik az elemkiválasztásos 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 a tanítás során előállított rejtett Markov-modellek végzik.

A HMM-ek tanítására alapvetően két típusú eljárás létezik: a beszélőfüggő tanítás és beszélőadaptációs eljárás.

Az első esetben szükség van egy beszélőtől rögzített, minél hosszabb hanganyagra. A rendelkezésre álló hanganyagból kinyerjük a hullámformára jellemző spektrális, gerjesztési és a hangidőtartam paramétereket, majd ezekből egy – a hanganyagra jellemző – statisztikus modellt építünk.

A második esetben több beszélőtől kell minél hosszabb hanganyagokat gyűjtenünk, továbbá szükségünk van egy adott célbeszélőtől (akinek a hangkarakterisztikáját próbáljuk majd visszaadni a beszédelőállítás során) származó rövidebb felvételre. Az összegyűjtött szövegkorpuszból az első esethez hasonlóan kinyerjük a hullámformára jellemző spektrális, gerjesztési és fonéma hangidőtartam paramétereket, majd a több beszélőtől gyűjtött hosszabb felvételekből kinyert paraméterek segítségével megépítjük az ún. átlaghangra (*average voice*) jellemző statisztikus modellt, melyet az adott célbeszélő rövidebb felvételéből kinyert paraméterek segítségével a célbeszélő hangkarakterére adaptálunk.

Mindkét esetben az előállt modelleket adatbázisban tároljuk, majd a beszédelőállítása során az adatbázisban tárolt modellekből kinyert paramétereket használjuk fel.

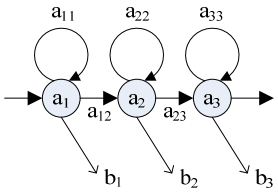
A beszédelőállításához beszédkódolási eljárást használunk, ahol a gerjesztési, szűrő és esetleges egyéb (pl. maradékjel) paramétereket HMM modellek generálják. A HMM-alapú magyar szövegfelolvasó beszélőfüggő tanításának lépéseiről korábban beszámoltunk [7], jelen cikkben röviden bemutatjuk a rejtett Markov-modellt, ismertetjük az átlaghang kialakítását, ennek adaptációs lehetőségeit és a betanított modellekből a beszédelőállításának folyamatát, továbbá bemutatjuk az általunk megvalósított szövegfelolvasó szubjektív méréséhez tervezett meghallgatásos teszt felépítését és eredményeit.

2 A rejtett Markov-modell

Gyakran használnak rejtett Markov-láncokat fizikai folyamatok modellezésére, ahol különböző megfigyelések alapján kell a folyamatot szimulálni. A beszédtechnológiában igen előnyösen lehet használni a rejtett Markov-modelleket, ekkor a beszédre jellemző, abból kinyert paramétereket kell tárolni, mely jelentősen hatékonyabb, mint a hangminta alapú rendszerek esetén a minták tárolása, hiszen a paraméterek jóval kevesebb helyet foglalnak el és jobban lehet belőlük általánosítani, mint az eredeti hullámformák esetén. A paraméterek (például spektrális jellemzők) kinyeréséhez úgynevezett akusztikus modelleket alkalmaznak. Régebben hangonkénti (ún.

monofón) akusztikus modellt használtak, manapság 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-modellek a beszédfelismerés alapjait képzik, szinte minden komoly rendszer erre a technológiára épül. 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 jelöli a beszédhang elejét, az a_2 a közepét és az a_3 pedig a végét. Az a_{12}, a_{23} élek a belső állapotok közötti átmeneti valószínűségeket jelentik, 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. A b_1, b_2, b_3 jelöli a megfigyelési valószínűségeket.



1. ábra. Három állapotú rejtett Markov-modell

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 rejtett Markov-modelleket [8] mutatja be részletesen.

A rejtett Markov-modell alkalmazása a beszédszintézis területén az elmúlt évtizedben merült fel és napjainkra egyre nagyobb figyelmet kapott. Az erre kidolgozott eljárás három 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 mintaválogatást hajtunk végre, 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, mely feladatokat szintén végezhetnek rejtett Markov-modellek. A harmadik fontos különbség pedig az, hogy trifón akusztikus modellek helyett sokkal összetettebb akusztikus modellt használunk, melyben az adott hanghoz közeli és távoli hangok szegmentális és szupraszegmentális szinten is beépülnek.

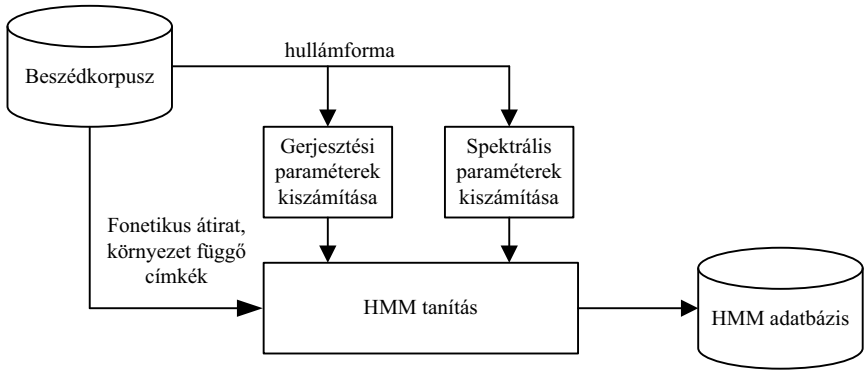
3 Rejtett Markov-modell alapú beszédszintézis

A HMM-alapú szövegfelolvasó két fő részből áll: a tanulási és a szintetizálási fázisból. A tanulás során a rejtett Markov-modelleket egy nagy, gondosan megtervezett és felcímkézett beszédadatbázis (és annak fonetikus átirata) segítségével tanítjuk be. A tanítási folyamat végére egy kisméretű HMM adatbázis áll elő, melyben a betanított beszédkorpuszra jellemző HMM paraméterek találhatóak. Ezekből válogatja majd ki a szintetizátor a beszéd előállítása során a szintetikus beszéd generálásához szükséges paramétereket. Ezen adatokból alakítja valamilyen beszédkódolási eljárással a paramétereket beszéddé.

A szintetizálási fázisban már csak a tanítás eredményét, egy néhány megabájtos adatbázist használunk. A bemeneti szöveg alapján meghatározzuk, hogy milyen hangsorozatot kell generálni és a HMM-adatbázisban tárolt paraméterekből kiválogatjuk azt a paramétersorozatot, amelyik legjobban reprezentálja az előállítani kívánt hangsorozatot. Ezekből állítjuk vissza a spektrális jellemzőket, a hangidőtartamokat, a szüneteket és az alapfrekvenciát, majd ezek alapján beszédkódoló eljárással elkészítjük a szintetizált beszéd hullámformáját.

A HMM modellek tanítására alapvetően kétfajta lehetőségünk van: beszélőfüggő modell tanítása vagy beszélőfüggetlen modell tanítása, majd az így előálló átlaghang adaptációja egy adott célszemély beszédhangjára.

Beszélőfüggő esetben a tanításhoz egy beszélő minél hosszabb hangfelvételére (legalább 1-1.5 óra), ennek fonetikus átiratára és pontos hanghatárjelölésekre van szükség. Fontos, hogy a hangfelvétel szövege fonetikusan kiegyenlített legyen. Hogy minél jobb minőségű hangot tudjunk előállítani, ügyelni kell arra, hogy a felvételek stúdió körülmények között legyenek rögzítve, továbbá hogy a fonetikus átirat és a címkézés precíz legyen. A hanghatárokat a gyakorlatban automatikus, úgynevezett kényszerített beszédfelismerési (forced alignment) módszerrel jelöljük meg. Ebből adódik bizonyos mértékű hiba. A beszélőfüggő tanítás lépéseit a 2. ábra mutatja be.



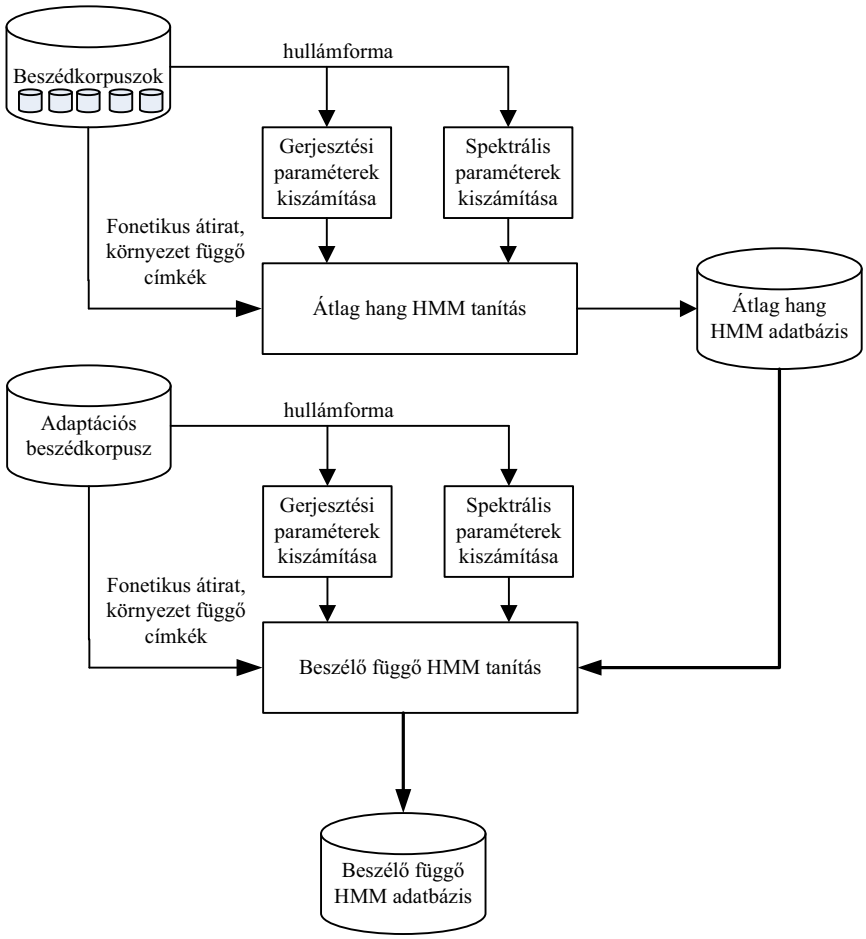
2. ábra. Beszélőfüggő HMM adatbázis tanítása.

A tanításhoz ezen túl szükségünk van az adott nyelvre jellemző környezetfüggő címkézésre és a döntési fák építéséhez egy nyelvspecifikus kérdésfájlra [9]. Ezek segítségével megkezdődhet a tanítás, mely a hosszú, több száz megabyte-ot elfoglaló

hanganyagból az adott beszélőre jellemző beszédhang paraméterek generálására alkalmas HMM adatbázist eredményez.

A HMM-alapú magyar beszédelőállításról korábban részletesen beszámoltunk [7], a továbbiakban a beszélőfüggetlen tanítás adaptációját ismertetjük.

Beszélőfüggetlen esetben először egy átlaghangot tanítunk, melyet utána egy célbeszélő hangkarakteréhez igazítunk. Ebben az esetben így áll elő a HMM adatbázis. Ezután a beszédhang előállításának módszere megegyezik a beszélőfüggő esetben használt módszerrel. A beszélőfüggetlen tanítás, majd adaptálás működési elvét a 3. ábra mutatja be.



3. ábra. Beszélőfüggetlen HMM adatbázisból kiinduló adaptált tanítás.

3.1 Beszélőfüggetlen átlaghang tanítása

A beszélőfüggetlen esetben először elő kell állítani egy ún. átlaghangot. Ennek előállításához több beszélőtől (legalább 4-5), minél hosszabb (személyenként legalább 1-1.5 óra) hangfelvételre, annak fonetikus átiratára és pontos hanghatárjelöléseire van szükség. A minél jobb minőség érdekében itt is érdemes figyelni arra, hogy a felvételek stúdió körülmények között legyenek rögzítve, illetve hogy a fonetikus átirat és a címkézés precíz legyen. Ezután automatikus módszerrel előállítjuk a beszédkorpuszhoz tartozó fonetikus átirat környezet függő címkéit, majd a HMM-eket az összes beszélő adatbázisa alapján tanítjuk be az átlaghangra, melyben jelen vannak minden egyes beszélőre az alapfrekvencia, hangidőtartam és spektrális paraméterek.

Érdekes kérdés, hogy az átlaghang tanításához férfi, női, vagy kevert hangokat használjunk. Amennyiben nagy mennyiségű férfi és női hanganyag áll rendelkezésre, a leghatékonyabb megoldást a nemfüggő átlaghang használata jelenti. A gyakorlatban azonban általában az egyik, vagy mindkét nemtől csak korlátozott mennyiségű hanganyagunk van, ezért a kevert nemű átlaghang előállítást célszerű választanunk, majd ebből adaptálni mind férfi, mind női hangra. Meg lehet csinálni, hogy ellentétes nemű átlaghangból adaptálunk női / férfihangra, azonban [10] beszámol arról, hogy ez jelentős minőség- és természetességcsökkenést okoz a végső hangnál a nemfüggő átlaghanghoz képest. [11] egy olyan eljárásról számol be, mely segítségével kevert nemű átlaghangból a nemfüggő átlaghanghoz képest minimális minőség- és természetességromlás mellett lehet női és férfihangra adaptálni.

3.2 Beszélőadaptáció

Miután elkészültek az átlaghang HMM modelljei, a célbeszélőtől származó hangfelvételekkel tudjuk a modellt az adott személy hangkarakteréhez és beszédstílusához igazítani, adaptálni. A beszélőadaptációjára alapvetően kétfajta lehetőségünk van.

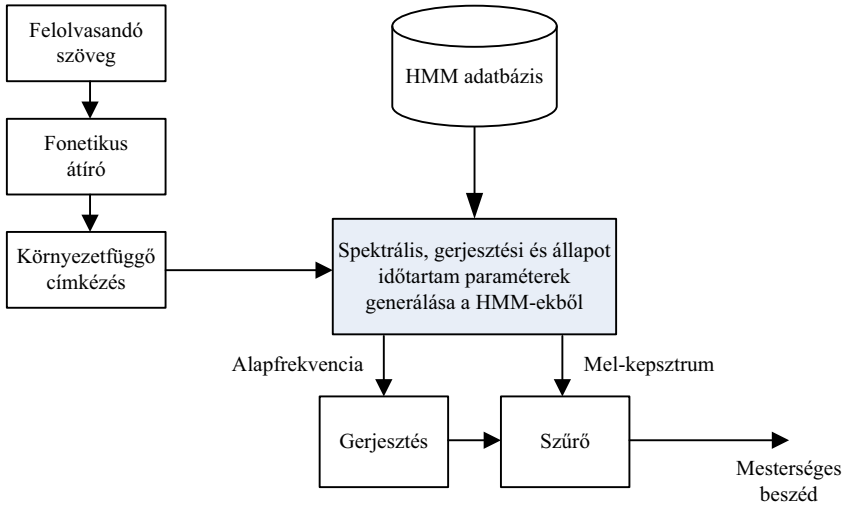
Amennyiben kevés hanganyag áll rendelkezésre a célbeszélőtől, akkor előnyös Maximum Likelihood Linear Regression (MLLR) alapú adaptációt választani [5]. [12] irodalomban ismertetett kísérlet alapján akár már öt mondat is elegendő lehet ahhoz, hogy a célszemély hangkarakterét és beszédstílusát többé-kevésbé visszaadja a mesterségesen előállított hang.

Amennyiben hosszabb adaptációs hanganyag is elérhető, akkor a Maximum A Posteriori (MAP) technikát érdemes használni [6], mely az előzőnél jobb minőségű mesterségesen generált hangot eredményez. Ennek a technológiának az új változatai, mint például a CSMAPLR (Constrained Structural Maximum A Posteriori Linear Regression) közel azonos minőséget és természetességet képviselnek, mint a beszélőfüggő tanítás esetén előállított mesterséges beszéd [13].

Természetesen mindegyik adaptációs technológia esetén szükséges az adaptációs hanganyag fonetikus átirata és a pontos hanghatárjelölések.

3.3 Beszéd előállítása

A beszéd előállítása megegyezik a beszélőfüggő esetben használt eljárással. A beszéd előállítása során a HMM által generált alapfrekvencia, hangidőtartam és spektrális paramétereket használjuk fel. A HMM-ek tanításától függően a beszéd előállítását végezheti egészen egyszerű beszédkódoló is (pl. LPC-10). A jobb minőség érdekében használhatunk ennél bonyolultabb technológiákat, mint például a MELP (Mixed Excitation Linear Prediction) kódoló. Természetesen ebben az esetben a beszédkorpuszból számolt maradék jeleket is be kell tanítanunk a HMM-ekkel.



4. ábra. A beszédhang előállítása a HMM adatbázisból.

4 Magyar nyelvű tanítás és adaptáció

A magyar nyelvű HMM-alapú beszélőadaptált szövegfelolvasó elkészítésének bizonyos lépései hasonlóak a beszélőfüggő változathoz. A döntési fák építéséhez és a környezet függő címkézéshez a korábban bemutatott eljárást használtuk [7]. Jelen cikkünkben az adaptációhoz használt adatbázisokat és az alkalmazott adaptációs technológiát ismertetjük.

4.1 A felhasznált beszédkorpuszok

Az átlaghang építéséhez négy férfi és egy női beszélőtől rögzített adatbázist használtunk. Az adatbázisokat stúdió körülmények között vettük fel, az adatbázisok szövege gondosan megtervezett, fonetikusan kiegyenlített mondatokat tartalmaz. Az átlaghang készítéséhez felhasznált adatbázisok további jellemzőit az 1. táblázat mutatja.

1. táblázat: Az átlaghang létrehozásához használt beszédkorpuszok (formátum: 44 kHz, 16 bit, mono).

Beszélő	Mondatszám	Időtartam	Méret
1. férfi beszélő	1941	170 perc	857 MB
2. férfi beszélő	1938	137 perc	694 MB
3. férfi beszélő	1944	191 perc	966 MB
4. férfi beszélő	1938	214 perc	1082 MB
1. női beszélő	1940	129 perc	652 MB

Miután készen lett az átlaghang HMM adatbázis, négy különböző beszélőtől rögzített, félig spontán hanganyagot használtunk fel közepesen zajos környezetből az adaptációhoz, melyek tulajdonságait a 2. táblázat mutatja. Mind a négy esetben publikusan elérhető parlamenti felvételeket használtunk, melyek előre megtervezettek, de spontán módon előadottak.

2. táblázat: Az adaptációhoz használt beszédkorpuszok (formátum: 44 kHz, 16 bit, mono).

Beszélő	Mondatszám	Időtartam	Méret
1. férfi beszélő	87	19 perc	94 MB
2. férfi beszélő	48	17 perc	89 MB
3. férfi beszélő	30	11 perc	58 MB
4. férfi beszélő	26	9 perc	44 MB

4.2 Az alkalmazott adaptációs technológia

A beszélőadaptáció során MLLR eljárást használtunk. Az MLLR lineáris transzformációk segítségével az átlaghang HMM modell paramétereit a cél hang 'irányába' módosítja. Az állapotkimenetek ekkor a következőképp alakulnak:

$$b_j(o_t) = N(o_t; \hat{\mu}_j; \hat{\Sigma}_j) \quad (\mathbf{1})$$

$$\hat{\mu}_j = A_{r(j)}\mu_j + b_{r(j)} \quad (\mathbf{2})$$

$$\hat{\Sigma}_j = H_{r(j)}^T \Sigma_j H_{r(j)} \quad (\mathbf{3})$$

ahol $\hat{\mu}_j$ és $\hat{\Sigma}_j$ a j-edik állapotra jellemző kimeneti sűrűségfüggvényhez tartozó várható érték vektor ill. kovariancia mátrix a lineáris transzformáció után. $A_{r(j)}$, $b_{r(j)}$ és $H_{r(j)}$ a várható érték lineáris-transzformációs mátrixa, a hozzá tartozó eltolás vektor és a kovariancia lineáris-transzformációs mátrixa az r(j)-edik regressziós osztályban. Az adott állapotokra jellemző kimeneti sűrűségfüggvényeket regressziós-fa

segítségével osztályokba soroljuk, egy adott osztályban azonos lineáris-transzformációs mátrixokat és az eltolás vektort használunk. A regressziós fa méretének az adaptációs anyag mennyiségéhez való igazításával tudjuk szabályozni az adaptáció komplexitását és általánosítható képességét. Alapvetően az MLLR két fajtáját különböztetjük meg: azonos A és H lineáris-transzformációs mátrixok esetén erőltetett MLLR-ről (Constrained MLLR, CMMLR), egyébként pedig szabad MLLR-ről (Unconstrained MLLR) beszélünk. A jelen cikkben ismertetett rendszer esetén CMLLR-t használtunk.

5 Eredmények

A rejtett Markov-modell alapú szövegfelolvasó beszélőadaptációjának megvalósításához a HTS rendszer [9] módosított, magyar nyelvű változatát vettük alapul [7]. A tanításhoz és adaptációhoz a 4.1 szakaszban ismertetett beszédkorpuszokat használtuk fel. Az összeállított rendszer minőségének szubjektív mérése céljából egy meghallgatásos tesztet állítottunk össze.

5.1 A teszt felépítése

A tesztben a korábban ismertetett adaptációs anyagok alapján négy különböző férfihangra adaptált rendszer vett részt. A teszt két részből áll. A teszt első felében a tesztalanyoknak 16 mintát (rendszerenként négyet) kellett 1-től 5-ig tartó skálán értékelniük természetesség szempontjából. Az 1 azt jelentette, hogy a hangminta zavaróan gépies hangzású, az 5 pedig azt, hogy teljesen természetes. A teszt második felében a beszélők eredeti hangkarakteréhez viszonyítva kellett a tesztalanyoknak szintén 1-től 5-ig tartó skálán megmondaniuk, hogy mennyire adja vissza a szintetizált hang az eredeti beszélő hangkarakterét. Az 1-es itt azt jelentette, hogy egyáltalán nem adja vissza, az 5-ös pedig hogy a szintetizált hangminta összetéveszthető az eredeti beszélővel. A teszt második felében minden rendszerből 5 mintát, így összesen 20 mintát hallgattak meg.

Mindkét részben a minták pszeudovéletlenszerűen lettek kiválogatva egy 40 darabos halmazból, ügyelve arra, hogy a minták előfordulási gyakorisága egyenletes eloszlást kövessen. A különböző rendszerekből kiválogatott mintákat ezután véletlen sorrendbe rendeztük. Ezen lépésekre azért volt szükség, hogy elkerüljük a memóriahatást a teszt során, tehát hogy a tesztalanyok által adott értékeket nem csak a minták tartalma, hanem a minták sorrendje is befolyásolja (pl. egy rosszabb minta után következő jobb minta sok esetben jobb pontszámot kaphat, mintha előtte is egy hasonló minőségű minta állna).

A tesztet összesen 25-en végezték el, 19 férfi és 6 nő. Internet alapú volt a teszt, az átlag életkor 35 év volt, a legfiatalabb tesztalany 21, a legidősebb 67 éves volt. 10 tesztalany beszédszakértő volt.

5.2 Az eredmények értékelése

A teszt eredményeit a 3. táblázat mutatja.

3. táblázat: A meghallgatásos teszt eredményei. Mindkét oszlopban az első érték az átlag, a második az átlagos szórás, a harmadik, zárójelben lévő érték pedig a konfidenciát jelöli =0.05 mellett.

Adaptációs korpusz	A hangminta természetessége	Hasonlóság az eredeti beszélő hangjához
1. férfi beszélő	3.2 ± 1.09 (0.2)	2.9 ± 1.08 (0.2)
2. férfi beszélő	3.1 ± 1.08 (0.2)	2.9 ± 1.05 (0.2)
3. férfi beszélő	3 ± 1.17 (0.2)	2.7 ± 1.05 (0.2)
4. férfi beszélő	3 ± 1.11 (0.2)	2.6 ± 1.06 (0.2)

Az értékekből kitűnik, hogy a hangminta természetessége a különböző beszélők esetén közel azonos, a hosszabb adaptációs anyag nem okozott szignifikáns különbséget a rövidebbhez képest. Ez azzal magyarázható, hogy mindegyik hang az átlaghangból lett adaptálva, mely már önmagában is elég információt hordoz természetes hang létrehozásához.

A teszt második felében, a hangminta összehasonlítás során azonban már meg lehet figyelni, hogy rövidebb adaptációs anyag (ld. 2. táblázat) esetén az eredeti beszélőhöz való hasonlóság csökken.

6 Jövőbeli tervek

A jövőben kísérleteket fogunk végezni azzal kapcsolatban, hogy a félig spontán, közepes minőségű adaptációs anyagokat stúdió minőségű, tervezett beszédűre cserélve hogyan változik a generált hang minősége és természetessége. A hanghatárok automatikus jelölését ellenőrizni fogjuk félautomatikus és kézi módszerekkel. Ezen túl más típusú adaptációs technológiákat is kipróbálunk (például MAP vagy CSMAPLR). Méréseket végzünk ezek minőségével kapcsolatban.

Kiemelt fontosságúnak tartjuk a beszédelőállításának folyamatát mobil platformokra optimalizálni. A fentebb ismertetett megoldás futás idejű tárhely igénye (1-2 MB) elméletileg lehetővé teszi kevés erőforrással rendelkező eszközökre való átvitelét, azonban számítás-igénye jelentős optimalizációra szorul.

7 Összefoglaló

Cikkünkben röviden áttekintettük a rejtett Markov-modell alapjait, kapcsolatát a beszédtechnológiával, és különösen a beszédszintézissel. Röviden összefoglaltuk a magyar nyelvű, beszélőfüggő HMM-alapú mesterséges beszédelőállítás elemeit, majd részletesen ismertettük a beszélőadaptációhoz szükséges lépéseket. Ezután ismertet-

tük a megvalósított rendszer szubjektív méréséhez tervezett meghallgatásos teszt felépítését és annak eredményeit. Végezetül jövőbeli terveinkre tértünk ki.

A beszédszintézis területén jelenleg az egyik leggyorsabban fejlődő terület a rejtett Markov-modell alapú beszédelőállítás. Szeretnénk a világgal lépést tartva magyar nyelven is megvalósítani a legújabb technológiákat, illetve új eredményekkel hozzájárulni a terület gyorsabb fejlődéséhez.

Hivatkozások

1. Black, A., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In Proc. ICASSP (2007), 1229-1232
2. Kim, S.-J., Kim, J.-J., Hahn, M.-S: HMM-based Korean speech synthesis system for hand-held devices. IEEE Trans. Consumer Electronics 52 (4) (2006) 1384–1390
3. N. Iwahashi, Y. Sagisaka: Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks", Speech Communications, Vol. 16, no. 2 (1995) 139–151
4. Tachibana, M., J. Yamagishi, Masuko, T., Kobayashi, T.: Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing. IEICE Trans. Inf. Syst., Vol. E88-D, no.11 (2005) 2484-2491
5. Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Adaptation of Pitch and Spectrum for HMM-Based Speech Synthesis Using MLLR. In Proc. ICASSP 2001, (1998) 805-808
6. Ogata, K., Tachibana, M., Yamagishi, J., Kobayashi, T.: Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis. In Proc. ICSLP 2006, (2006) 1328–1331
7. Tóth, B., Németh, G.: Hidden Markov model based speech synthesis system in Hungarian, Infocomm., Vol. 63, no. 7 (2008) 30–34
8. Rabiner, Lawrence R: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. (1989) 257–286
9. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K.: The HMM-based speech synthesis system version 2.0, in Proc. ISCA SSW6. (2007) 294–299
10. Isogai, J., Yamagishi, J., Kobayashi T.: Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis. In Proc. EUROSPEECH 2005 (2005), 2597–2600
11. Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T., Tokuda, K.: Improved Average-Voice-based Speech Synthesis using Gender-Mixed Modeling and A Parameter Generation Algorithm considering GV, Proc. ISCA SSW6, Aug. (2007)
12. Tamura, M., Masuko, T., Tokuda, K., Kobayashi T.: Speaker adaptation for HMM-based speech synthesis system using MLLR, Proc. ESCA/COCOSDA Workshop on Speech Synthesis (1998) 273-276
13. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J. Analysis of Speaker Adaptation Algorihms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm.' IEEE Audio, Speech, & Language Processing Vol.17 issue 1 (2009) 66-83 2009