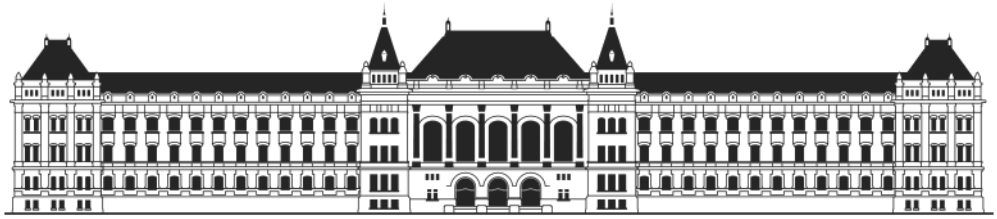


Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

Rejtett Markov-modell alapú gépi beszédkeltés

Doktori értekezés
Villamosmérnöki Tudományok Doktori Iskola

Tóth Bálint Pál
okl. villamosmérnök

Tudományos vezetők:
Németh Géza, Ph.D.
Olaszy Gábor, D.Sc.

Budapest, 2013

Tartalomjegyzék

Abstract

of the PhD Thesis of Bálint Pál Tóth,
„Hidden Markov-model based text-to-speech synthesis”

Hidden Markov model based text-to-speech (HMM-TTS) synthesis became one of the most intensively studied research area among speech experts recently. This Ph.D. thesis focuses on improvements of HMM-TTS including language independent and Hungarian language specific methods as well.

In the first chapter the basics of speech synthesis are introduced. In the second chapter, HMM-TTS is described in detail, including speech coding techniques, speech parameter modeling, context-dependent labels and decision trees, furthermore speaker dependent and speaker adaptive training, parameter and speech generation algorithms are also described.

In chapter 3 speech databases, the software environment, methodology of the research, and the parameters of the listening tests are discussed.

In chapter 4 the novel results of the research is summarized briefly in three thesis groups, which are described in detail below.

The first thesis group introduces a Hungarian hidden Markov model based speech synthesis system. The language specific modifications of speaker dependent HMM-TTS are described. This system produces synthetic voice in similar quality to other competitive Hungarian TTS alternative solutions. In the next step the quality of synthetic speech is increased by introducing distinctive features. Speaker adaptation applying distinctive features results in significantly better quality than the speaker dependent case. The manual correction of automatic transcription and phoneme boundaries has also been investigated. The results suggest that high quality phonetic transcription and phoneme labeling does not always cause significant improvement in the final speech quality.

The second thesis group examines unsupervised speaker adaptation in hidden Markov model based speech synthesis systems. An automatic, speech recognizer transcription based, unsupervised adaptation method is described. The effectiveness of the method is enhanced by setting the beam width of forced alignment iteratively. Even if the phoneme error rate is about 50%, the method can still produce synthetic speech with similar quality to the supervised speaker adaptation case.

The third thesis group focuses on porting HMM-TTS to low-resource devices, such as smartphones. The spectral representation of speech and the unvoiced excitation method are modified, the size of decision trees is limited and the parameter generation algorithm is tailored low-resource devices. As a result the response time of HMM-TTS is reduced to about 5% of the baseline version, while the quality of the synthetic speech does not change significantly. The measurements are carried out on three different modern smartphones.

The results have both theoretical and practical application benefits. In the last chapter of this Ph.D. thesis the author gives some possible application scenarios of his research and summarizes his results.

Kivonat

Tóth Bálint Pál

„Rejtett Markov-modell alapú gépi beszédkeltés”
című PhD értekezéséhez

A gépi beszédkeltés területén az egyik legnagyobb érdeklődést a rejtett Markov-modell alapú szövegfelolvasás váltotta ki az elmúlt években. Értekezésem ennek a tématerületnek nyelvfüggetlen és általánosságra törekvő, magyar nyelvű kutatásával foglalkozik.

A dolgozat a gépi beszédkeltés általános ismertetésével kezdődik, majd részletesen bemutatja a rejtett Markov-modell alapú szövegfelolvasás elméleti hátterét. Az értekezés az impulzus-zaj és kevert gerjesztésű beszédkódolók, a beszédkódolók paramétereinek folytonos és több-terű eloszlással való modellezése, a környezetfüggő címkék és a döntési fák témakörei mellett ismerteti a beszélőfüggő és beszélőadaptált tanítás, továbbá a paraméter- és hullámforma generálás részleteit.

A fentieket a kutatói munka során használt beszédkorpuszok, a felhasznált kísérleti konfigurációk, a beszédszintézis paraméter beállításai és a szubjektív gépi beszédminőség mérésére használt meghallgatásos tesztek bemutatása követi.

A kutatási célkitűzések megfogalmazása során a szerző tudományos eredményeit három fő téziscsoportba sorolja, mely csoportokon belül fogalmazza meg új tudományos állításait.

Az első téziscsoport a magyar nyelvű rejtett Markov-modell alapú szövegfelolvasó bevezetésével foglalkozik. A szerző módszert dolgozott ki a magyar nyelv sajátosságainak leírására beszélőfüggő rejtett Markov-modell alapú szövegfelolvasó rendszerben és megmutatta, hogy a megoldása versenyképes más magyar nyelvű szövegfelolvasókkal. Az így létrejött szövegfelolvasó rendszer minőségét megkülönböztető jegyek bevezetésével növelte. Eljárást dolgozott ki magyar nyelvű beszélőadaptációra, mellyel a beszélőfüggő esethez képest szignifikánsan jobb minőségű beszédet állított elő. Megvizsgálta a nagypontosságú kézi címkézés hatását a gépi beszéd minőségére.

A második téziscsoport témája a felügyelet nélküli beszélőadaptáció rejtett Markov-modell alapú szövegfelolvasó rendszerekben. A szerző eljárást dolgozott ki a beszédfelismerő kimenete alapján, emberi beavatkozás nélkül, új gépi hangkarakterű beszéd létrehozására. Az eljárás hatékonyságát a kényszerített illesztés keresési terének (beam szélességének) iteratív módon történő állításával növelte rossz minőségű felismerési kimenet esetén. Ennek eredményeként közel 50% fonémahiba-arány esetén is képes a felügyelt beszélőadaptációhoz hasonló minőségű gépi beszéd létrehozására.

A harmadik téziscsoport a rejtett Markov-modell alapú szövegfelolvasó korlátozott erőforrású eszközökre való illesztésével foglalkozik. A kutatás során a spektrális modellezés, a zöngétlen hangok gerjesztési módszere és a döntési fák mérete több lépésben módosításra kerültek. Továbbá a szerző a korlátozott erőforrású eszköz aktuális terhelését figyelembe vevő, párhuzamos működésű beszédelőállító eljárást dolgozott ki rejtett Markov-modell alapú szövegfelolvasókhoz. A fenti lépések sorozatával a rendszer válaszideje szignifikáns mértékben, a kiindulási rendszer közel egy huszadára csökkent, miközben a gépi beszéd minősége szignifikánsan nem változott.

A disszertációban bemutatásra kerülő eredmények a tématerület nemzetközi és hazai szintjén is új megközelítésre világítanak rá. Gyakorlati alkalmazásuknak lehetőségeit a szerző a dolgozat végén az összefoglalással együtt ismertet.

Előszó

A beszédkeltés bonyolult, komplex folyamat. Az agy nagy sebességgel, összehangolt módon működteti a hangképző szerveket (tüdő, légcső, gége, garat, száj- és orrüreg, ajkak) és a beszélő hallás útján kap visszacsatolást verbális kommunikációjáról. Éppen ezért a tökéletes gépi beszéd létrehozásához nemcsak a beszédkeltés mechanizmusát, hanem az agy működését is meg kell értenünk. Ameddig az agyban lejátszódó folyamatokat nem értjük meg teljes mértékben, addig csak közelítőlegesen van lehetőségünk modellezni az emberi beszédkeltést.

A modellezés **általános célja** a minél *természetesebb* és minél *érthetőbb* gépi beszéd létrehozása. Mindemellett **mérnöki szempontokat** is fontos figyelembe venni: *milyen erőforrás szükséges* a gépi beszéd előállításához, milyen rendszerekben, milyen eszközökkel valósítható meg?

Doktori értekezésemben mind az általános, mind pedig a mérnöki szempontok figyelembevételével ismertetem eredményeimet a gépi beszédkeltés területén. Dolgozatom címe „*Rejtett Markov-modell alapú gépi beszédkeltés*”, mely nagyméretű tématerületet jelöl. Tudomásom szerint magyar nyelven a témát részletesen összefoglaló dokumentum nincs, ezért fontosnak tartottam, hogy a tématerület alapelveit – a hazai és nemzetközi szinten is újnak számító tudományos eredményeim bemutatása mellett – magyar nyelven részletesen ismertessem.

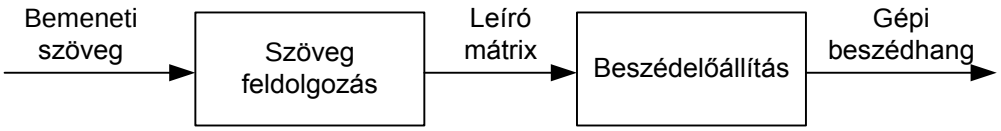
Dolgozatom **függelékében** a kutatási eredményeim bemutatása során használt fontosabb kulcsszavak, rövidítések jelentését foglaltam össze.

1. A gépi beszédkeltés

Az embereket régóta foglalkoztatja a beszédhang gépi előállítása. A gépi szövegfelolvasók fejlődése nemzetközi és hazai viszonylatban is több évtizedes múltra tekint vissza [1,2,3,4,5,6,7,8]. Általánosságban elmondható, hogy minden szövegfelolvasó két részből áll: a szövegfeldolgozó és a beszédelőállító alrendszerekből (1. ábra).

A szöveg feldolgozás során a bemeneti adatsort elemezve létrehozunk egy **leíró mátrixot**, melyben szerepelnek a bemeneti szöveget reprezentáló beszédhangok és a bemeneti szövegre jellemző hang- és frázisszintű információk. A szövegfeldolgozás magába foglalja a *fonetikus átírást*, továbbá tartalmazhat *érzelem kifejező* és *értelem módosító* információkat is (pl. hangsúlyozás, gyors-lassú, hangos-halk beszéd). A fonetikus átírást általánosságban véve a nyelvi sajátosságok alapján végezzük el, míg az érzelem kifejező és értelem módosító információk tökéletes modellezéséhez, mint említettem, az agy működését kellene ismernünk. Az utóbbi hiányában egyszerűbb, szöveg alapján készített empirikus modelleket alkalmazunk. A leíró mátrixban található információ halmazt továbbítjuk a gépi beszéd előállításáért felelős modulnak.

A gépi beszédelőállítás az a folyamat, amikor a bemeneti szövegből készített leíró mátrix alapján **emberi beszédre emlékeztető hangsorozatot hozunk létre**. Az elmúlt időszakban különböző megközelítésekkel modellezték a beszédelőállítás folyamatát.



1. ábra. A gépi szövegfelolvasók általános struktúrája.

A gépi beszédelőállítás 1791-ben Kempelen Farkas mechanikus *beszélőgépénél* kezdődött, mely beszédhangok és szavak megszólaltatására volt képes, hosszú mondatokat nem lehetett vele előállítani [9]. Működtetését ember végezte. A Bell laboratóriumban 1939-ben mutatták be a *VODER* (Voice Operation DEmonstratoR) nevű elektromechanikus gépet [10], mellyel már folyamatos, de nehezen érthető gépi beszédet lehetett előállítani. A szerkezet irányítása – Kempelen beszélőgépéhez hasonlóan – ez esetben is manuális volt. Mintegy 200 évvel Kempelen beszélőgépe után az igazi áttörést a gépi beszédállításban a számítógépek elterjedése jelentette. A számítógépek segítségével a manuális rendszereknél jóval gyorsabb és pontosabb vezérlést lehetett megvalósítani.

Az első számítógép alapú szövegfelolvasó eljárások közül az un. *formánsszintézis* volt a legjelentősebb technológia. Ezekben a rendszerekben egy egyszerűbb vagy bonyolultabb gerjesztő jel (impulzus-zaj generátor vagy a gége jelére emlékeztető hullámforma) hajtott meg egy lineáris, időváriáns szűrősort, a szűrők paramétereit pedig az emberi beszédre jellemző formánsok értékei szerint szabályalapon vezérelték [11,2,3,12]. Ez idő tájt az un. *artikulációs szintézis* volt a másik legígéretesebb megközelítés, mely a toldalékcső és az abban lejátszódó folyamatok kettő- vagy háromdimenziós fizikai modelljét készíti el [1]. A hangkeltés folyamatát a toldalékcső alakját befolyásoló számos paraméterrel, mint például a tüdő tágulása, nyelv pozíciója, ajkak formája, állkapocs helyzete vezérli. A modellek hatékonyak, hiszen hosszadalmas kézi elemzés után a formásszintézissel az emberi beszédre megtévesztésig hasonlító gépi beszédet tudtak már az 1960-70-es években előállítani. Az artikulációs szintézissel pedig jó minőségű magánhangzókat tudtak létrehozni. Azonban a mai gépi szövegfelolvasó rendszerekhez képest mind a formáns-, mind pedig az artikulációs automatikus szintézis hangminősége rossz. Ezt elsősorban azzal lehet magyarázni, hogy a

paraméterek nagyfelbontású és pontos automatikus előállítására szabály alapon egyelőre nincs megoldás. (A gépi tanulás alapú megközelítések pedig még kísérleti fázisban vannak [13].)

A *diád- és triád alapú beszédszintézis* az emberi hangból kivágott hullámforma összefűzésén alapul. Segítségével már sokkal inkább emberi hangra emlékeztető, ám még mindig gépies beszéd hozható létre [4,5]. Diádnak egy két hangból álló hangkapcsolat első hangjának a második felét, és második hangjának az első felét nevezzük. Triádnak egy három hangból álló hangkapcsolat első hangjának a második felét, a teljes középső hangját, és harmadik hangjának az első felét nevezzük. A szövegfelolvasó a diádokat vagy a triádokat a bemeneti szövegnek megfelelően fűzi össze, majd az így keletkezett hullámforma hangidőtartamait, alapfrekvencia menetét és intenzitását a leírómátrix alapján módosítja.

Napjaink egyik legfejlettebb technológiája a hullámforma *elemkiválasztáson alapuló beszédszintézis*. Ez a technológia jó minőségű, precízen felcímkézett emberi beszédből válogatja össze a felolvasandó szövegnek megfelelő hullámforma részeket, és ezeket fűzi össze [6,7]. Nagyméretű adatbázis esetén elemkiválasztáson alapuló *korpuszos beszédszintézisnek* is nevezzük. Minősége egy adott tématerületen kiváló (például időjárás jelentés), azonban más, vagy általános tématerületen nem minden esetben tud azonosan jó minőséget biztosítani. A technológia hátránya műszaki szempontból az, hogy futás időben akár több gigabyte tárkapacitás szükséges hozzá és igen számításigényes az eljárás. További hátrány, hogy új beszédhang létrehozásához a teljes beszédkorpuszt ismételten rögzíteni kell az adott célszemély hangján.

A fentieken kívül léteznek még más szövegfelolvasásban alkalmazott beszédkódoló technológiák is, amelyekre például a *szinusz hullám alapú* [14] és az *LPC alapú* beszédszintézis [15] épül.

Az elmúlt években a legnagyobb figyelmet a jelen értekezésben használt új módszer, a ***statisztikai parametrikus beszédszintézis, ezen belül a rejtett Markov-modell alapú gépi beszédkeltés*** kapta.

2. Gépi beszédkeltés rejtett Markov-modellekkel

Az elmúlt években a gépi beszédkeltés terén – számos előnyének köszönhetően – a ***statisztikai parametrikus beszédszintézis*** vált az egyik legaktívabb kutatási területté [16]. Igaz, hogy az 1. ábra felépítése alapján a statisztikai parametrikus beszédszintézis megnevezés az egész szövegfelolvasó struktúrát jelöli, mégis az eljárás általánosságban csak a *beszédelőállítás* folyamatára vonatkozik.

A statisztikai parametrikus beszédszintézis során először a jellemző paramétereket (pl. spektrális összetevők, hangmagasság, hangidőtartamok) kinyerjük a beszédadatbázisból (*un. beszédkorpuszból*), majd ezen paraméterek sokaságát generatív modellekkel helyettesítjük. Jelenleg általánosan elterjedt megoldást jelent a rejtett Markov-modell (Hidden Markov Model, HMM) alapú generatív modell. A modell paramétereinek becslésére ($\hat{\boldsymbol{\lambda}}$) a következő képlet szerint leginkább a *maximum likelihood* (vagy hozzá hasonló) becslést alkalmaznak:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \{ p(\mathbf{O} \mid W, \boldsymbol{\lambda}) \} \tag{1}$$

ahol $\boldsymbol{\lambda}$ a modell paramétereit, $\boldsymbol{O}$ a beszédkorpuszból származó jellemző paramétereket (tanítóadatok) és W az $\boldsymbol{O}$-hoz tartozó szósorozatot jelöli. A folyamat eddigi részét *tanításnak*, az ez utáni részét pedig *szintézisnek* nevezzük.

A gépi beszéd előállítása során a w szósorozathoz és $\hat{\boldsymbol{\lambda}}$ becsült modell paraméterekhez tartozó $\boldsymbol{o}$ paraméterek kimeneti valószínűségét az alábbiak szerint maximalizáljuk:

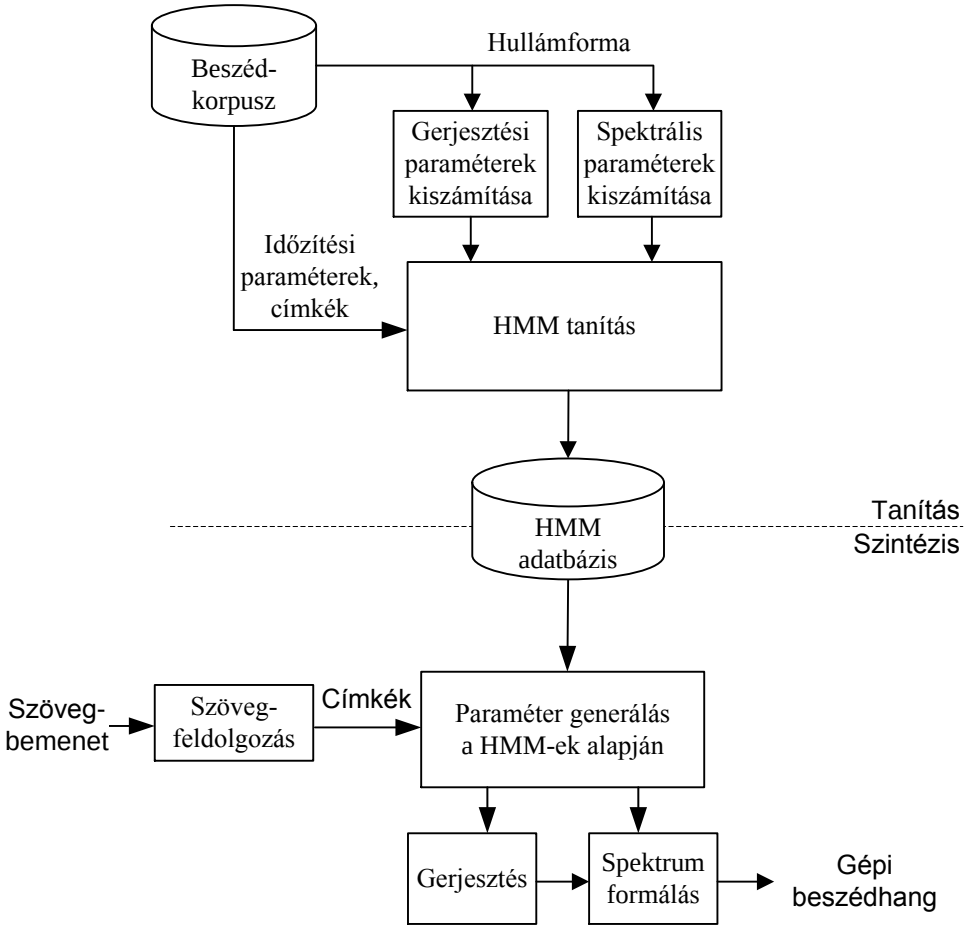
$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{ p(\mathbf{o} \mid w, \hat{\boldsymbol{\lambda}}) \} \tag{2}$$

Az így kialakult emberi beszédre jellemző paraméterhalmazból ezután készítjük el a gépi beszéd hullámformáját. A HMM generatív modell alapú statisztikai parametrikus beszédszintézist az irodalomban rejtett Markov-modell alapú szövegfelolvasásnak hívják, melyet az angol szakirodalomban "*Hidden Markov model based Text-To-Speech*" terminussal adnak meg, és *HMM-TTS*-ként hivatkoznak rá. A továbbiakban az egyszerűség kedvéért a nemzetközileg használt rövidítést (HMM-TTS) használom.

A korábbi technológiákhoz képest a HMM-TTS számos előnnyel rendelkezik. Kis futásidejű adatbázissal jó minőségű gépi beszéd előállítására képes [17], lehetőség van arra, hogy a rendszert adott célbeszélő hangkarakterisztikájához adaptáljuk [18,19], illetve több beszélő hangját interpoláljuk [20], és mindemellett érzelem kifejezésre is alkalmas [21].

Egy általános HMM-TTS blokkdiagramját a 2. ábra mutatja be. A beszédkorpusz hullámformáiból *kinyerjük a gerjesztési és spektrális paramétereket*, majd ezen paramétereket a nagyméretű beszédkorpuszhoz tartozó *címkékkel együtt átadjuk a HMM tanítási szakasznak*, mely az (1)-es egyenlet maximum likelihood becslését végzi. A tanítási szakasz eredményeképp előállnak a gépi beszéd előállításának (szintézisnek) alapját képező HMM generatív modellek, melyeket HMM adatbázisnak hívunk. A gépi beszéd előállítása során erre az adatbázisra támaszkodva a (2)-es egyenlet maximalizálását végezzük el: a bemeneti szósorozatot un. környezetfüggő címkékké alakítjuk, és a felolvasandó szöveg rejtett Markov-modelljeit a környezetfüggő címke sorozat szerint fűzzük össze. Ez a HMM-lánc generálja azt a gerjesztési és spektrális paramétersorozatot, ami alapján beszédkódoló eljárással előállítjuk a paraméter sorozatból a végső hullámformát, a gépi beszédet.

A HMM-TTS esetén alapvetően két fő típusát használják a beszédkódolóknak: az impulzus-zaj gerjesztésű és a kevert gerjesztésű beszédkódolót.



2. ábra. HMM-TTS általános blokkdiagramja beszélőfüggő tanítás esetén; [22] alapján, módosítva.

2.1. Beszédkódolók

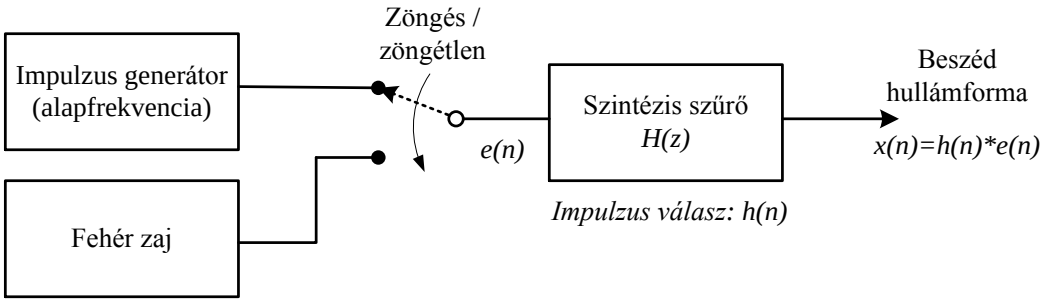
A HMM-TTS rendszerek forrás-szűrő modell alapú beszédkódolót használnak. A forrás készíti el a gerjesztési jelet, mely a tüdőből kiáramló levegő – a hangszalagok rezgése által megformált – hullámformájának felel meg. Ez a jel hajtja meg a szűrőt, mely a toldalékcső karakterisztikáját modellezi. A gerjesztési jel típusa nagymértékben meghatározza a HMM-TTS rendszerek minőségét. Doktori értekezésemben a HMM-TTS szempontjából a két legfontosabb beszédkódoló eljárást mutatom be: az impulzus-zaj és a kevert gerjesztésű beszédkódolókat. Ezen túl HMM-TTS rendszerekben más forrásmodellekkel is végeztek kutatást [23].

2.1.1. Impulzus-zaj gerjesztésű beszédkódoló

Impulzus-zaj alapú beszédkódoló esetén az *e(n)* gerjesztést periodikus és zajszerű jellel modellezzük [24]. A zöngés hangokat periodikus, a zöngétlen hangokat zajszerű jel írja le, ezért a gerjesztésük létrehozásáért rendre egy impulzus- és egy zajgenerátor felel. Az *e(n)* gerjesztéssel meghajtjuk a *H(z)* átviteli karakterisztikájú, *h(n)* impulzus válaszú szűrőt. A *H(z)* idő-variáns lineáris szűrő a toldalékcső karakterisztikáját írja le. A szűrő kimenetén a gerjesztési jel és a szűrő impulzusválaszának diszkrét konvolúciójaként megjelenik az *x(n)* beszédhang:

$$x(n) = e(n) * h(n) \tag{3}$$

A fentiek alapján impulzus-zaj gerjesztésű beszédkódoló esetén a beszédhang előállításához a következő paraméterekre van szükségünk: *gerjesztés típusa, alapfrekvencia (zöngés gerjesztés esetén), spektrális paraméterek, időzítési paraméterek*. Az eljárás vázlatos felépítését a 3. ábra mutatja be.



3. ábra. Impulzus-zaj gerjesztésű beszédkódoló felépítése; [24] alapján, módosítva.

2.1.2. Kevert gerjesztésű beszédkódoló

A folyamatos beszédben nem lehet szigorúan periodikus vagy zajszerű jelekként kezelni a zöngés és zöngétlen hangokat. A zöngés hangok esetén számolnunk kell zajszerű, zöngétlen esetben pedig periodikus komponensekkel. Mivel az impulzus-zaj gerjesztésű beszédkódolók ezt nem képesek modellezni, részben ezért gépies a hangzásuk. További problémát jelent az impulzus-zaj gerjesztésű beszédkódolóknál, hogy a beszéd irregularitását, az alapfrekvencia kisebb-nagyobb ingadozását (pl. az un. mikro-intonációt és a glottalizációt) nem modellezi. A fenti problémák kiküszöbölése érdekében a gerjesztés komplexebb modellezését valósítják meg a *kevert gerjesztésű beszédkódolók* (pl. *Mixed Excitation Linear Prediction, MELP*) [25].

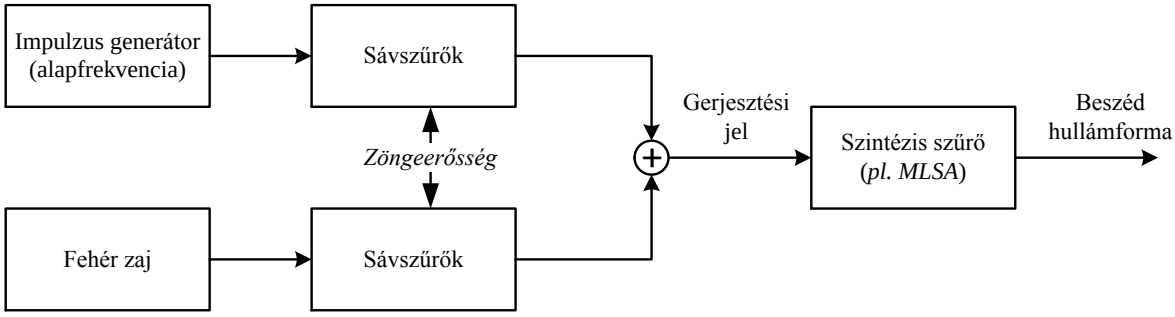
Kevert gerjesztés esetén a gerjesztés mind zajszerű, mind pedig periodikus komponenseket is tartalmaz, melyek arányát sávszűrők segítségével határozzuk meg. A beszédjelet sávokra osztjuk, és minden sávban kiszámoljuk a zöngeerősséget. Jelöljük a beszédjelet az *n*-edik minta esetén *s(n)*-el, az alapfrekvencia elemzés során használt ablakméretet *N*-el. Ekkor a *t*-edik minta esetén a *c(t)* zöngeerősséget a (4) egyenlet alapján számoljuk ki. HMM-TTS esetén az alábbi sávokra való felbontást használják elsődlegesen: 0-1000 Hz, 1000-2000 Hz, 2000-4000 Hz, 4000-6000 Hz, 6000-8000 Hz.

$$c(t) = \frac{\sum_{n=0}^{N-1} s(n)s(n+t)}{\sqrt{\sum_{n=0}^{N-1} s(n)s(n) \sum_{n=0}^{N-1} s(n+t)s(n+t)}} \tag{4}$$

Ezen túl meghatározzuk az alapfrekvenciát, továbbá inverz szűrő segítségével kiszámoljuk a Fourier magnitúdókat. A gerjesztési jel előállítása során a zöngeerősséggel meghatározzuk a periodikus és zajszerű részek arányát (4. ábra). A periodikus jelet az alapfrekvencia és a Fourier együtthatók segítségével állítjuk elő, majd a zöngeperiódus csúcsát a zöngeerősség függvényében változó mértékben eltoljuk (maximum a periódusidő 25%-ig). A zöngétlen hangok gerjesztésére fehérzajt használunk. Mindkét jelet külön-külön szűrjük, majd összegezzük őket. Így áll elő a végső gerjesztési jel. Ezzel a jellel hajtjuk meg a szintézis

szűrőt, mely a toldalékcső karakterisztikáját modellezi. A HMM-TTS esetén gyakran *MLSA* (*Mel Log Spectrum Approximation*) szűrő használatos [26].

A fentiek alapján egy kevert gerjesztésű beszédkódoló esetén a beszédhang előállítására a következő paraméterekre van szükség: *alapfrekvencia, Fourier magnitúdók, zöngeerősségek sávonként, spektrális paraméterek, időzítési paraméterek.*



4. ábra. Kevert gerjesztésű beszédkódoló felépítése; [25] alapján, módosítva.

2.2. A beszédkódoló paramétereinek modellezése

Klasszikus távközlési értelemben véve a beszédkódoló a paramétereket az átviteli csatorna egyik végén a beszédjel elemzése során számolja ki (*analízis*). Ezen paraméterek kerülnek továbbításra az átviteli csatornán, majd a túlsó végén a paraméterekből a beszédjel visszaállítása a cél (*szintézis*). A HMM-TTS rendszerekben az analízis a tanítás során történik. A szintézis során pedig a betanított generatív modellek feladata, hogy előállítsák a beszédkódoló számára a felolvasandó szövegre leginkább jellemző paraméterfolyamokat.

A rejtett Markov-modell alapú szövegfelolvasó rendszerekben rendszerint minden hangot ötállapotú, balról-jobbra tartó (*Left-To-Right*), visszacsatolás nélküli rejtett Markov-modellek írnak le. Az öt állapot a hang elejét, közepét és a végét, illetve az egyes részek közötti átmeneteket reprezentálja. Minden paraméterfolyamhoz külön-külön ötállapotú modellek tartoznak.

Mind az impulzus-zaj, mind a kevert gerjesztésű beszédkódolók esetén két típusú modellezés használatos: folytonos eloszlású rejtett Markov-modellek (*Continuous Density Hidden Markov Model, CD-HMM*) [27] és többterű eloszlású rejtett Markov-modellek (*Multi-Space probability Distribution Hidden Markov Model, MSD-HMM*) [28]. A folytonos eloszlású rejtett Markov-modellekkel a spektrumot leíró kepsztrális együtthatókat, a Fourier magnitúdókat és zöngeerősséget, az *MSD-HMM*-el pedig az alapfrekvenciát modellezzük.

Fontos megemlíteni, hogy a paraméterek modellezésénél a gépi beszéd dinamikájának javítása céljából a rejtett Markov-modellekkel az egyes paraméterfolyamok *delta* (*sebesség*) és *delta-delta* (*gyorsulás*) paramétereit is tanítják. Ezek számítását a gyakorlatban az (5) és a (6) egyenletek alapján végezzük.

$$\Delta x_t = -\frac{1}{2}x_{t-1} + \frac{1}{2}x_{t+1}, \tag{5}$$

$$\Delta^2 x_t = \frac{1}{4}x_{t-1} - \frac{1}{2}x_t + \frac{1}{4}x_{t+1} \tag{6}$$

A fentiek alapján a HMM-TTS rendszerek jellemzővektora impulzus-zaj gerjesztés esetén az 1. táblázatban, kevert gerjesztés esetén a 2. táblázatban megadott paraméterfolyamokat tartalmazza.

1. táblázat. HMM-TTS jellemzővektor impulzus-zaj gerjesztés esetén.

Paraméterfolyam	**Eloszlás**	**Paraméterek**
Spektrum	CD-HMM	mel-kepsztrális együtthatók Δmel-kepsztrális együtthatók Δ^2mel-kepsztrális együtthatók
Alapfrekvencia	MSD-HMM	alapfrekvencia Δalapfrekvencia Δ^2alapfrekvencia

2. táblázat. HMM-TTS jellemzővektor kevert gerjesztés esetén.

Paraméterfolyam	**Eloszlás**	**Paraméterek**
Spektrum	CD-HMM	mel-kepsztrális együtthatók Δmel-kepsztrális együtthatók Δ^2mel-kepsztrális együtthatók
Alapfrekvencia	MSD-HMM	alapfrekvencia Δalapfrekvencia Δ^2alapfrekvencia
Zöngeerősség	CD-HMM	zöngeerősség Δzöngeerősség Δ^2zöngeerősség
Fourier magnitúdó	CD-HMM	magnitúdó Δmagnitúdó Δ^2magnitúdó

2.2.1. Folytonos eloszlás: CD-HMM

A beszédfelismeréshez hasonlóan a folytonos eloszlású HMM-ek a toldalékcső modellezésére szolgálnak. A folytonos eloszlású HMM-ek véges állapotgépek, melyek minden lépésben egy állapotátmenetet hajtanak végre és a valószínűségi sűrűségfüggvényüknek megfelelő kimeneti vektort generálnak. Ezen kimeneti vektor alapján lehetséges a beszédjel spektrális formálása (lásd 5. ábra).

A spektrális formálás a következőképp történik: legyen a felolvasandó szöveget leíró paraméterfolyam D dimenziós o_t vektor. A felolvasandó szöveget osszuk N egyenlő szegmensre, az S_i-edik állapothoz tartozó szegmenst jelöljük d_i-vel. Az a_{ij} átmeneti valószínűség annak a valószínűségét adja meg, hogy az i-edik állapotból áttérünk a j-edik állapotba és eleget tesz az (7) egyenlőségnek. Ebben az esetben minden állapothoz M-összetevőből (mixture-ből) álló, Gauss eloszlású $b_j(o_t)$ kimeneti valószínűségi sűrűségfüggvény tartozik a (8) egyenlet szerint. A (8) egyenletben c_{jk} a mixture együtthatót, μ_{jk} a D-dimenziós várható érték vektort, Σ_{jk} a DxD méretű kovariancia mátrixot jelöli a j-edik állapotban a k-adik összetevő esetén. A Gauss eloszlások normalizálása céljából a valószínűségi c_{jk} együtthatók kielégítik a (9) és (10) egyenletet. A (8) egyenletben használt Gauss függvények összességének a segítségével lehetséges bármilyen véges, folytonos

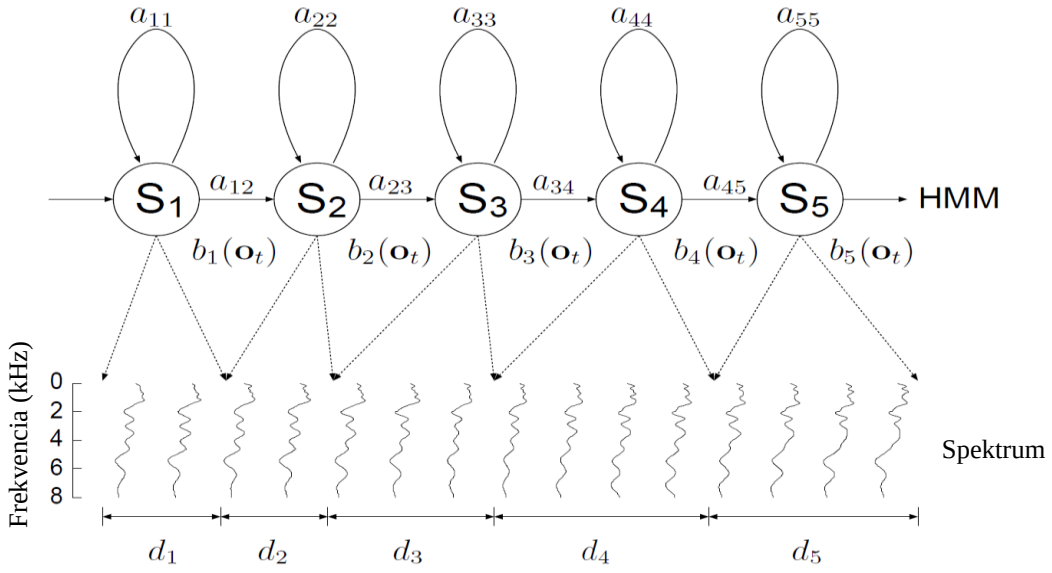
eloszlású függvényt becsülni. A beszéd spektrális összetevője is ilyen. A becslés pontos lépéseiről az irodalomban olvashatunk [29].

$$a_{ii} + a_{ij} = 1 \tag{7}$$

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk}) =$$
$$= \sum_{k=1}^{M} c_{jk} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_{jk}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(o_t - \mu_{jk})^T \Sigma_{jk}^{-1} (o_t - \mu_{jk})\right\} \tag{8}$$

$$\sum_{k=1}^{M} c_{jk} = 1; \qquad 1 \le j \le N \tag{9}$$

$$c_{jk} \ge 0; \qquad 1 \le j \le N, 1 \le k \le M \tag{10}$$



5. ábra. Ötállapotú rejtett Markov-modell sematikus ábrája a spektrum és az időzítési paraméterek modellezésének bemutatása céljából; [27] alapján.

2.2.2. Többterű eloszlás: MSD-HMM

Az alapfrekvencia menet a beszédjelben jelenlévő zöngés (periodikus) részeken folytonos, a zöngétlen (zajszerű) részeken pedig diszkrét értéket vesz fel. Emiatt az alapfrekvencia modellezése folytonos vagy diszkrét HMM-ekkel igen nehézkes. A HMM-TTS rendszerekben széles körben alkalmazott eljárás a többterű eloszlású rejtett Markov-modell, az MSD-HMM használata [28]. A többterű eloszlások magukba foglalják mind a diszkrét, mind pedig a folytonos eloszlásokat: a zöngés részben megfigyelt alapfrekvencia értékek egy dimenziós térből, zöngétlen esetben pedig nulla dimenziós térből veszik fel értéküket.

2.2.3. Időzítési paraméterek modellezése

A minél természetesebb hangzású gépi beszéd érdekében a hangszínezet és hangmagasság mellett igen fontos a beszéd ritmikájának modellezése. A beszéd ritmikáját HMM-TTS rendszerekben állapot átmeneti sűrűségfüggvényekkel modellezik: minden hang minden állapotátmenetéhez egy több dimenziós Gauss eloszlás tartozik [30].

2.3. Környezetfüggő címkék

A folytonos beszédben a beszédhangok alapfrekvenciáját, spektrális együtthatóit és az időtartamokat számos hang- és frázisszintű tulajdonság befolyásolja, és a hangok egymásra is hatással vannak. HMM-TTS rendszerekben *környezetfüggő címkék* segítségével írjuk le a beszéd e fajta akusztikai változatosságát. Egy adott beszédhang konkrét megvalósulását számos tényező befolyásolja. Ezen tényezők egy része nyelvfüggetlen, másik része pedig az adott nyelv szabályait követi. Például a koartikulációs hatások figyelembe vételére minden beszédhang esetén jelölnünk kell annak környezetét (*nyelvfüggetlen*), vagy például a hangsúlyozást szótag és frázis szinten is előnyös megadnunk (*nyelvfüggő*). A pontos modell létrehozása érdekében a tulajdonságokat hang-, szótagmag-, szó-, mondatrész- és mondatszinten kell meghatározni.

A környezetfüggő címkézést a generatív modellek tanítása és a szintézis során is automatikusan végezzük. Legyen például a környezetfüggő címkék halmaza a következő:

- az adott hang és környezete 2-2 hang távolságig (kvinfón),
- az adott szótag hangsúlyos-e vagy nem,
- hangok száma a szótagban,
- hang pozíciója a szótagban.[1]

Ezek alapján a környezetfüggő címkék a „*disszertáció*” szó első négy hangja esetén a következőképpen alakulnak:

„d” hang	hang kettővel az aktuális hang előtt = nem értelmezett; hang az aktuális hang előtt = nem értelmezett; aktuális hang = “d”; hang az aktuális hang után = “i”; hang kettővel az aktuális hang után = “ssz”; adott szótag hangsúlyos? = igen; hangok száma a szótagban = “3”; hang pozíciója a szótagban = “1”.
„i” hang	hang kettővel az aktuális hang előtt = nem értelmezett; hang az aktuális hang előtt = “d”; aktuális hang = “i”; hang az aktuális hang után = “ssz”; hang kettővel az aktuális hang után = “e”; adott szótag hangsúlyos? = igen; hangok száma a szótagban = “3”;

[1] A példában a szótaghatáron levő hosszú mássalhangzók nem tartoznak a nyelvtani szabályok szerint mindkét szótaghoz, hanem a jelölés alapján csak az elsőhöz, a következő szótag egyből a második hanggal kezdődik.

	hang pozíciója a szótagban = “2”.
„ssz” hang	hang kettővel az aktuális hang előtt = “d”; hang az aktuális hang előtt = “i”; aktuális hang = “ssz”; hang az aktuális hang után = “e”; hang kettővel az aktuális hang után = “r”; adott szótag hangsúlyos? = igen; hangok száma a szótagban = “3”; hang pozíciója a szótagban = “3”.
„e” hang	hang kettővel az aktuális hang előtt = “i”; hang az aktuális hang előtt = “ssz”; aktuális hang = “e”; hang az aktuális hang után = “r”; hang kettővel az aktuális hang után = “t”; adott szótag hangsúlyos? = nem; hangok száma a szótagban = “3”; hang pozíciója a szótagban = “2”.

A tömörebb tárolás és a könnyebb átláthatóság érdekében a környezetfüggő címkéket a gyakorlatban kódolni szokás, amit a tanítás és a szintézis során reguláris kifejezésekkel részekre lehet bontani. Egy tömörebb reprezentációra vizsgáljuk meg a példában szereplő környezetfüggő címkéket:

- az adott hang (p3) és környezete 2 hang távolságig (p1,p2,p4,p5): „p1^p2-p3+p4=p5”
- az adott szótag hangsúlyos-e: „@1_”, nem hangsúlyos: „@0_”
- hangok száma a szótagban: „_hangokszama:”
- hang pozíciója a szótagban: „:hangpozíció”

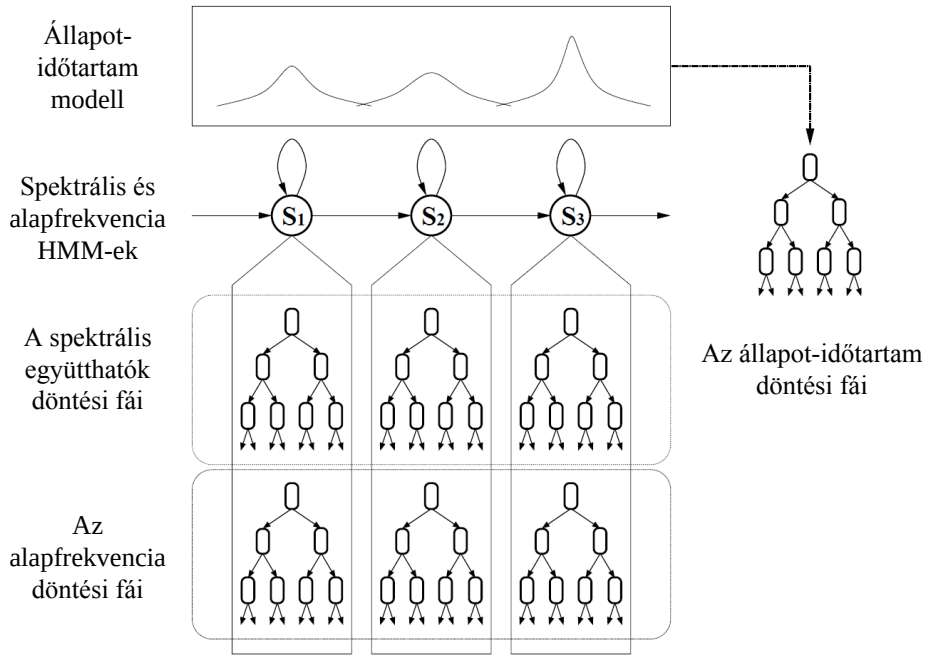
Ezek alapján a következőképp néz ki az első négy hangja a „disszertáció” szónak tömörebb reprezentációban:

x^x-d+i=ssz@1_3:1
x^d-i+ssz=e@1_3:2
d^i-ssz+e=r@1_3:3
i^ssz-e+r=t@0_3:2

2.4. Döntési fák

A környezetfüggő címkék lehetséges kombinációja túl nagy ahhoz, hogy megfelelően reprezentatív beszédkorpuszt tudjunk hozzá készíteni. A szemléltetés kedvéért végezzünk el egy gondolat kísérletet. Legyen egy adott nyelvben 40 beszédhang. Ekkor a kvinfón lehetséges változatainak száma $40^5 = 102.400.000$. Ehhez minden további környezetfüggő címkét hozzávéve az állapottér exponenciálisan növekszik. Már a HMM-TTS kutatások elején is a beszédhang és környezetének jelölésén túl 36 darab környezetfüggő címke volt a nemzetközi szinten kiindulásnak számító angol nyelvű mintarendszerben [22]. Ilyen méretű és felépítésű beszédkorpusz kialakítása lehetetlen, ezért a paraméterfolyamokat *döntési fák segítségével* osztályokba soroljuk [27].

Más és más tényezők befolyásolhatják a gerjesztési, a spektrális és az időzítési paramétereket, ezért mindegyikhez külön döntési fát előnyös készíteni (lásd 6. ábra). A döntési fák a tanítás során jönnek létre. Szerepük, hogy az adott nyelvre, továbbá az emberi beszédmechanizmusra jellemző szabályok alapján a hasonló tulajdonságú környezetfüggő címkékhez tartozó paramétereket osztályokba sorolják. Ez az ún. *kérdések* alapján történik. A kérdések segítségével a címkék tetszőleges részhalmazát jelölhetjük ki. A kiválasztott kérdések lesznek a fák csomópontjaiban.



6. ábra. Döntési fák minden paraméterfolyamhoz, minden állapot esetén; [27] alapján.

A döntési fa építése a következőképp történik: a tanító minták környezetfüggő címkéihez tartozó paraméterek kezdetben egy halmazt alkotnak. Az algoritmus minden kérdést feltesz, és megvizsgálja, hogy a paramétereket melyik kérdés választja legegyenlőbb módon két halmazra. Ez a kérdés kerül a fa gyökerébe, ahonnan kettéágazik a fa, és a keletkezett két ágon megjelenik a kérdés alapján a két halmaz. A származtatott két elemet *gyereknek* nevezzük. Az algoritmust rekurzív módon folytatjuk a *gyerek* csomópontokra.

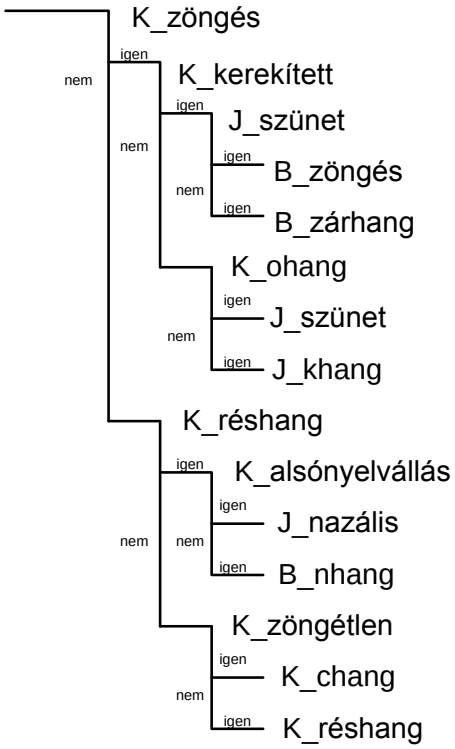
Minden egyes környezetfüggő címkéhez kérdéseket készítünk. Az előző pontban lévő példában szereplő környezetfüggő címkék esetén a kérdések egy lehetséges halmaza a következő:

- p1,p2,p3,p4,p5 mássalhangzó vagy magánhangzó?
- p1,p2,p3,p4,p5 zöngés-e vagy zöngétlen?
- p1,p2,p3,p4,p5 nazális hang?
- p1,p2,p3,p4,p5 alsó nyelvállású magánhangzó?
- p1,p2,p3,p4,p5 hátsó nyelvállású magánhangzó?
- p1,p2,p3,p4,p5 ajakkerekítéses magánhangzó?
- Hangsúlyos a szótag?
- Egy hang van a szótagban?
- Két hang van a szótagban?

- Három hang van a szótagban?
- Négy hang van a szótagban?
- Öt vagy annál több hang van a szótagban?
- Az aktuális hang az első hang a szótagban?
- Az aktuális hang az második hang a szótagban?
- Az aktuális hang az harmadik hang a szótagban?
- Az aktuális hang az negyedik hang a szótagban?
- Az aktuális hang az ötödik hang a szótagban?
- stb.

Amennyiben a kezdeti paraméterhalmazt túlságosan kis csoportokra bontjuk (nagyméretű fa), akkor az algoritmus *rátanul* a tanító adatbázisra, de rossz lesz az általanosító képessége (*túltanul*). Amennyiben túlságosan nagy csoportoknál állunk le (kisméretű fa), akkor az a tanító adatbázist, és így az emberi beszédet is rosszul fogja modellezni. Ezért fontos meghatároznunk, hogy mikor fejezzük be az algoritmust.

HMM-TTS esetén leálló feltételként jellemzően a *legkisebb leíró hossz* (*Minimum Description Length, MDL*) eljárást használják [31]. Az eredeti és a *gyerek* halmaz leíró hosszának a különbsége adja meg, hogy szétbontsuk-e új halmazra. Nagyszámú környezetfüggő címke, vagy kevés tanítóadat esetén a *MDL* használata sokszor ahhoz vezet, hogy a végső halmazok csak egy elemet fognak tartalmazni. Ezen problémát oldja meg a *minimális előfordulás* (*Minimal Occupancy*) használata, mely alapján ha egy halmazban kevesebb elem van mint a minimális előfordulás értéke, akkor azon az ágon, még akkor is leáll a döntési fa építése, ha a *leíró hossz* alapján az algoritmust még tovább kellene folytatni.



7. ábra. A spektrális paraméterekhez tartozó egy lehetséges döntési fa.

A 7. ábra egy adott beszédhanghoz tartozó általános spektrális paraméterekre vonatkozó döntési fára mutat példát. Az ábrán látható „K_” előtag azt jelenti, hogy épp a középső hangot vizsgáljuk a kvinfónból, 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). Ezután a fa következő szintjén lévő “K_kerekített” (középső hang labiális) és “K_réshang” (középső hang réshang) alapján való csoportokra bontás volt a legjobb. A fa további szintjeit ezzel az elvvel lehet tovább bontani.

A szintézis során a betanított döntési fákban a bemeneti szöveghez tartozó címkék alapján választjuk ki a bemenethez legjobban illeszkedő halmazt. A döntési fa csomópontjaiban lévő kérdésekre a válaszokat a címkék adják meg. A kiválasztott halmaz így a bemenethez legjobban illeszkedő környezetfüggő rejtett Markov-modelleket tartalmazza. Ezek a generatív modellek készítik el a szintetizálandó bemeneti szöveghez legjobban illő paraméterfolyamot.

2.5. A rejtett Markov-modellek tanítása

A HMM-TTS tanításához gondosan megtervezett, fonetikailag kiegyenlített tanító beszédadatbázis, un. beszédkorpusz szükséges. Ez az adatbázis legalább a következő elemeket tartalmazza: *a beszéd hullámformája, pontos időzítési paraméterek (hanghatárok), fonetikus átirat*. A tanítás megkezdése előtt a hanganyagból kinyerjük a gerjesztési és spektrális paramétereket, a fonetikus átiratból pedig előállítjuk a környezetfüggő címkéket. Így a rejtett Markov-modelleket végül a környezetfüggő címkékkel, továbbá a gerjesztési, spektrális és időzítési paraméterekkel tanítjuk. A HMM-TTS tanításának alapvetően két típusa van: a *beszélőfüggő* és a *beszélőadaptált* tanítás.

2.5.1. Beszélőfüggő tanítás

A beszélőfüggő tanítás esetén a tanító beszédkorpusz csupán egy beszélőtől származó hanganyagot tartalmaz. Az alapfrekvenciához, gerjesztési és időzítési paraméterekhez tartozó generatív, környezetfüggő modellek erre az egy beszélőre jellemző értékeket fogják megtanulni. Ezáltal a gépi beszéd hasonló hangkarakterrel és prozódiával fog megszólalni, mint az eredeti személy. Egy beszélő esetén – empirikus úton – minimum 1-2 óra hosszat javasolnak tanító adatbázis hosszának [32]. A beszélőfüggő tanítás sematikus felépítését a 2. ábra mutatja be.

2.5.2. Beszélőadaptált tanítás

A HMM alapú beszédszintézis egyik nagy előnye, hogy alkalmas beszélőadaptációra. A beszélőadaptáció annyit jelent, hogy a rendszer hangkarakterisztikáját egy adott célbeszélőhöz hasonlóra alakítjuk ki. Rejtett Markov-modellek esetén ehhez viszonylag rövid, 10-15 perces beszédkorpusz már elegendő az adott célbeszélőtől.

A beszélőadaptált tanítás folyamata hasonló a beszélőfüggő esethez, azonban itt a tanítást két fő részre oszthatjuk: a tanítást először egy *átlaghangra* kell elvégezni, melyet azután a második lépésben *célbeszélő hangkarakteréhez igazítunk* (8. ábra). Ebben az esetben így áll elő a szintézis alapját képező HMM adatbázis. Ezt követően a beszédhang-előállításának módszere megegyezik a beszélőfüggő esetben használt módszerrel. Az átlaghang előállításához legalább 4-5 beszélőtől minél hosszabb (személyenként legalább 1-2 óra) hangfelvételre, annak fonetikus átiratára és pontos hanghatár-jelöléseire van szükség. A HMM-eket az átlaghangra az összes beszélő adatbázisa, azaz minden beszélőre jellemző alapfrekvencia, hangidőtartam és spektrális paraméterek alapján tanítjuk be. Az átlaghang

tanítása a nagy mennyiségű adat miatt időigényes feladat: napjaink nagy teljesítményű számítógépein (pl. 2.6 GHz, 8 magos processzor, 16 GByte RAM) akár több hétig is eltarthat (folyamatos futással).

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 (10-15 perc) hanganyag áll rendelkezésre a célbeszélőtől, akkor előnyös *MLLR (Maximum Likelihood Linear Regression)* alapú adaptációt választani [18]. Akár már öt mondat is elegendő lehet ahhoz, hogy a gépi beszédhang a célszemély hangkarakterét és beszédstílusát visszaadja [18]. Az *MLLR* eljárás a lineáris transzformációk segítségével az átlaghang HMM modell paramétereit a célhang „irányába" módosítja. Az állapotkimenetek ekkor a következőképp alakulnak:

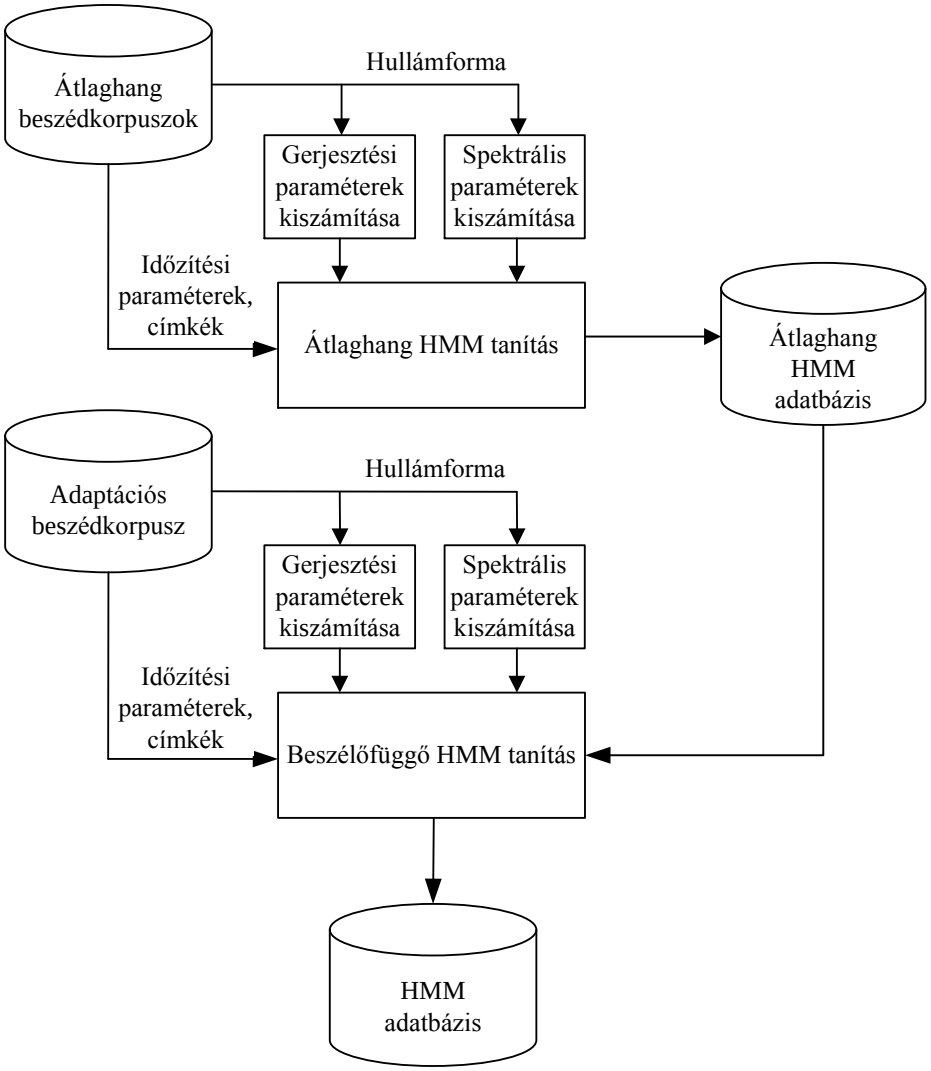
$$b_j(o_t) = N(o_t; \hat{\mu}_j; \hat{\Sigma}_j) \tag{11}$$

$$\hat{\mu}_j = A_{r(j)}\mu_j + b_{r(j)} \tag{12}$$

$$\hat{\Sigma}_j = H_{r(j)}^T \Sigma_j H_{r(j)} \tag{13}$$

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 szabályozható az adaptáció komplexitása és általánosítható képessége. Alapvetően az MLLR két fajtáját különböztetjük meg: azonos A és H lineáris-transzformációs mátrixok *esetén korlátozott MLLR*-ről (*Constrained MLLR, CMMLR*), egyébként pedig *korlátozás mentes MLLR*-ről (*Unconstrained MLLR*) beszélünk.

Amennyiben hosszabb (több mint 1 óra) adaptációs hanganyag is rendelkezésünkre áll, akkor a *Maximum A Posteriori* (*MAP*) technikával lehetséges az *MLLR*-hez képest minőségjavulást elérni, mely az előzőnél jobb minőségű adaptációt tesz lehetővé. [33].



8. ábra. HMM-TTS beszélőadaptált tanítás; [22] alapján, módosítva.

2.6. Paraméterfolyam és hullámforma generálás (szintézis)

Beszélőfüggő és beszélőadaptált esetben is a gépi beszédelőállítás a következő lépésekből áll:

1. Bemeneti szöveg fonetikus átírása.
2. Fonetikus átiratból a környezetfüggő címkék előállítása.
3. Állapotsorozat meghatározása és paraméterfolyam generálása a beszédkódoló számára.
4. A paraméterfolyamból beszédkódoló eljárással a hullámforma (gépi beszéd) előállítása.

A szöveg fonetikus átirata általánosságban a HMM-TTS rendszer működésétől függetlenül készül el. Ezután következik a környezetfüggő címkék előállítása, amelyek alapján meghatározzuk a legjobban illeszkedő állapotsorozatot és paraméterfolyamot. Ezek után a paraméterfolyamot a tanítás kezdeténél meghatározott beszédkódoló eljárással (lásd

2.1.) hullámformává alakítjuk, és így előáll a HMM-TTS alapú gépi beszéd. Vizsgáljuk meg részletesen a 3-dik lépést, a paraméterfolyam generálást.

2.6.1. Paraméterfolyam generálás a maximum likelihood kritérium alapján

A paramétergenerálást Tokuda és munkatársai több tanulmányban részletesen ismertetik [34,35]. A következőkben a paramétergenerálás értekezésem szempontjából is fontos részeit mutatom be. Legyen λ egy folytonos, többkomponensű (*mixture*) HMM az alábbiak szerint:

$$\lambda = (\mathrm{A}, \mathrm{B}, \pi) \tag{14}$$

ahol $A=\{a_{ij}\}$ jelöli az állapot átmeneti valószínűségeket, $B=\{b_j(o)\}$ az állapot kimeneteket a (9) alapján, és $\pi=\{\pi_i\}$ a kezdeti állapot eloszlását az *i*-edik állapotnak. Ezen HMM által generálandó

$$\mathrm{O} = [o_1^T,\ o_2^T, \dots, o_T^T] \tag{15}$$

paraméterfolyamot

$$\mathrm{P}(\mathrm{O} \mid \lambda) = \sum_{\forall Q} P(Q, O \mid \lambda) \tag{16}$$

maximalizáljuk O szerint, ahol

$$\mathrm{Q} = \{(q_1, i_1), (q_2, i_2), \dots, (q_T, i_T)\} \tag{17}$$

jelöli az állapotok és az összetevők (*mixture*-ök) sorrendjét (pl. *(q,i)* jelöli az *i*-edik összetevőjét a *q* állapotnak). Az o_t paraméter vektor mind statikus, mind pedig dinamikus jellemzőket tartalmaz:

$$o_t = [c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T \tag{18}$$

ahol

$$\mathrm{c}_t = [c_t(1), c_t(2), \dots, c_t(M)]^T \tag{19}$$

$$\Delta \mathrm{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} \omega^{(1)}(\tau)\mathrm{c}_{t+\tau} \tag{20}$$

$$\Delta^2 \mathrm{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} \omega^{(2)}(\tau)\mathrm{c}_{t+\tau} \tag{21}$$

A (16)-os egyenletet megoldása túl bonyolult, ezért a Viterbi algoritmushoz hasonlóan a megoldandó feladatot a

$$\mathrm{P}(\mathrm{O} \mid \lambda) = \max_Q P(Q, O \mid \lambda) \tag{22}$$

c szerinti maximalizálására vezetjük vissza. Feltételezve, hogy a paraméterfolyamok normális eloszlást követnek, $P(Q, O \mid \lambda)$ logaritmusa a (8) egyenlet alapján a következőképp írható fel:

$$\begin{aligned} \log \mathrm{P}(\mathrm{Q}, \mathrm{O} \mid \lambda) = \alpha \sum_{k=1}^{K} \log p_{q_k}\left(d_{q_k}\right) + \sum_{t=1}^{T} \log c_{q_t,i_t} - \\ -\frac{1}{2}(O-\mu)^T U^{-1}(O-\mu) - \frac{1}{2}\log|U| - \frac{3MT}{2}\log 2\pi \end{aligned} \tag{23}$$

ahol

$$\mu = \left[\mu_{q_1,i_1}^T, \mu_{q_2,i_2}^T, \dots, \mu_{q_T,i_T}^T\right] \tag{24}$$

$$U = diag\left[U_{q_1,i_1}, U_{q_2,i_2}, \dots, U_{q_T,i_T}\right] \tag{25}$$

és a q_t-edik állapot i_t-edik összetevőjéhez (*mixture)* tartozó súly c_{q_t,i_t}, a 3M × 1 méretű várhatóérték vektor μ_{q_t,i_t}, *a* 3M × 3M méretű kovarianca mátrix pedig U_{q_t,i_t}. A *T* keret során bejárt állapotok száma *K*, és $p_{q_k}\left(d_{q_k}\right)$ jelöli a d_{q_k} egymást követő megfigyelés valószínűségét q_k állapotban.

A (23) egyenlet alapján megállapítható, hogy $P(Q, O \mid \lambda)$ akkor maximális, ha $O = \mu$, tehát amikor az *U* kovariancia mátrixtól függetlenül a paraméterfolyam megegyezik a várhatóérték vektorral. Hogy elkerüljük ezt az esetet, $P(Q, O \mid \lambda)$ maximalizálását *Q* és *O* szerint a (20) és (21) egyenletek figyelembe vételével végezzük el. Jelen esetben a *Q* állapotsorozatot és az *O* paraméterfolyamot egyszerre kell meghatároznunk, ezért a Viterbi algoritmus esetén használt dinamikus programozás nem alkalmazható.

A probléma megoldásához először rögzített al-állapot sorozat mellett maximalizáljuk a $P(Q, O \mid \lambda)$ valószínűséget *c* szerint. A (20) és (21) egyenletek alapján a (23) egyenlet a következőre írható át:

$$\begin{aligned} \log \mathrm{P}(\mathrm{Q}, \mathrm{O} \mid \lambda) = \alpha \sum_{k=1}^{K} \log p_{q_k}\left(d_{q_k}\right) + \sum_{t=1}^{T} \log c_{q_t,i_t} - \\ -\frac{1}{2}\varepsilon(c) - \frac{1}{2}\log|U| - \frac{3MT}{2}\log 2\pi \end{aligned} \tag{26}$$

ahol

$$\varepsilon(c) = (Wc - \mu)^T U^{-1}(Wc - \mu) \tag{27}$$

és

$$W = [w_1, w_2, \dots, w_T]^T \tag{28}$$

$$w_t = \left[w_t^{(0)}, w_t^{(1)}, w_t^{(2)}\right] \tag{29}$$

$$w_t^{(n)} = \\ = \begin{bmatrix} 0_{M\times M_{első}}, \dots, 0_{M\times M}, \\ w^{(n)}\left(-L^{(n)}\right) I_{M\times M_{\left(t-L^{(n)}\right)-edik}}, \dots, w^{(n)}\left(-L^{(n)}\right) I_{M\times M_{t-edik}}, \\ \dots, w^{(n)}\left(-L^{(n)}\right) I_{M\times M_{\left(t+L^{(n)}\right)-edik}}, 0_{M\times M}, \dots, 0_{M\times M_{T-edik}} \end{bmatrix}^T \tag{30}$$

n=1,2,3 esetekben. $0_{M\times M}$ és $I_{M\times M}$ rendre az $M \times M$-es null és $M \times M$-es egységmátrixokat jelöli. Tegyük fel, hogy

$$c(t) = 0_M, t < 1, T < t \tag{31}$$

ahol 0_M az $M \times 1$ –es null vektort jelöli. Ekkor

$$\frac{\partial \log \mathrm{P}(\mathrm{Q}, \mathrm{O} \mid \lambda)}{\partial c} = 0_{TM} \tag{32}$$

alapján a következőt egyenleteket kapjuk:

$$Rc = r \tag{33}$$

ahol

$$R = W^T U^{-1} W \tag{34}$$

$$r = W^T U^{-1} \mu \tag{35}$$

A (33) egyenlet közvetlen megoldásához $O(T^3M^3)$ műveletre van szükségünk, mivel R egy $TM \times TM$ méretű mátrix. Ha $U_{q,i}$ diagonális mátrix, az azt jelenti, hogy a paraméterfolyam elemei függetlenek egymástól, így a megoldás komplexitása $O(T^3M)$-re csökken.[2]

[2] A (33) egyenlet hatékony megoldásával részletesen a 7.5. fejezetben foglalkozom.

3. Kutatási célkitűzések

Az alapvető célkitűzésem a statisztikai parametrikus beszédszintézis, azon belül a rejtett Markov-modell alapú szövegfelolvasás kutatása volt. Munkám során több esetben magyar nyelvű beszédkorpuszokra támaszkodtam, azonban az értekezésben bemutatott megoldások nagy részében nyelv specifikus információt nem használtam fel. A konkrét célkitűzéseim, melyek a téziseimben is megjelennek, a statisztikai parametrikus beszédszintézis területén a következők voltak:

I. Rejtett Markov-modell alapú szövegfelolvasó kialakítása és továbbfejlesztése magyar nyelvre. (5. fejezet)

II. Beszédfelismerés kimenetén alapuló felügyelet nélküli, rejtett Markov-modell alapú szövegfelolvasó beszédhangjának adaptációja. (6. fejezet)

III. Rejtett Markov-modell alapú szövegfelolvasás illesztése korlátozott erőforrású eszközökre. (7. fejezet)

A témát újszerűsége és a benne rejlő számos megoldandó kutatási probléma miatt választottam. Egyben kihívást jelentett számomra, hogy tudomásom szerint Magyarországon elsőként foglalkozom a témával.

Doktori értekezésem soron következő fejezeteiben a rejtett Markov-modell alapú beszédszintézis kutatásával és korlátozott erőforrású környezetbe való illesztésével kapcsolatos célkitűzéseimet, az alkalmazott módszereket és eszközöket, továbbá a hazai és nemzetközi szinten is újnak számító tudományos eredményeimet mutatom be részletesen.

4. Anyag és módszer

Ebben a fejezetben a kutatásom során használt beszédkorpuszokat, eszközöket, illetve a kutatási munka eredményeként elkészült rendszerek kiértékelésének metodikáját ismertetem. Kutatásom bemutatása során minden fejezetben kitérek arra, hogy a kutatás adott fázisában milyen beszédkorpuszt használtam. Ismertetem továbbá azt is, amennyiben az elkészült rendszerek vizsgálata céljából a jelen fejezetben bemutatott módszertől eltérő eljárást használtam.

4.1. A kutatás során használt beszédkorpuszok

Beszédkorpuszon a hanganyagot, a felolvasott szöveg fonetikus átiratát és a szegmentálási címkék halmazát értem.

A tanításhoz felhasznált *hanganyag* felolvasott és félspontán[3] beszédet tartalmaz. A hanganyagot minden esetben 16 kHz, 16 bit, mono formátumra alakítottam át, a jellemző paramétereket ebből a formátumból nyertem ki.

A *fonetikus átirat* a hanganyagot reprezentáló fonémasort jelöli, valamint a nem beszédnek tekinthető hullámforma részeket (például szünet, levegővétel). Előállítása vagy teljesen automatikus módon, vagy pedig fél-automatikusan történt. Ez utóbbi esetben az automatikusan meghatározott egységek kézi ellenőrzéssel lettek javítva.

A *szegmentálási címkék* minden egyes fonéma és szünet esetén jelölik, hogy a hanganyagban hol kezdődik és hol ér véget. A szegmentálási címkék meghatározása is automatikus módszerrel, vagy automatikus módszer utáni kézi korrekcióval történt. *Kényszerített illesztésnek*[4] (*forced alignment*) nevezzük azt a gépi eljárást, mely során a beszéd hullámformája alapján a fonetikus átirat egyes elemeihez (hangjaihoz) automatikus módon időzítési paramétereket (hanghatárokat) rendelünk [36]. A kényszerített illesztésről a 6.3. fejezetben írok bővebben.

Kutatásom során eltérő beszédkorpuszokkal dolgoztam, illetve a kitűzött cél határozta meg, hogy mely paramétereit vizsgáltam a rendelkezésre álló adatoknak. Ezen adatbázisokat és paramétereiket mindig az adott fejezetben ismertetem.

4.2. A szintézis teszteléséhez készített mondatok

Célom egy általános, nem témaspecifikus megoldás létrehozása volt, így a kutatás eredményeként létrejövő kísérleti rendszerekkel általános tartalmú és szerkezetű, kijelentő mondatokat készítettem és a tesztek során ezeket használtam fel. Mivel kutatásom részben magyar, részben angol nyelven folyt, így a tesztmondatokat is vegyesen magyar és angol nyelven készítettem el.

[3] A félspontán beszéd fogalmát és főbb tulajdonságait a 6.2. fejezetben ismertetem.
[4] A magyar irodalomban kényszerített felismerésként is szoktak rá hivatkozni.

4.3. Kísérleti konfigurációk

Részben szabadon hozzáférhető eszközöket, részben pedig a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszékén (BME-TMIT) korábban készült megoldásokat használtam. Ezek a következők voltak:

- **HTS** (HMM-based Text-To-Speech System): A rejtett Markov-modellek tanítása és adaptációja. [37]
- **SPTK** (Speech Processing Toolkit): Paraméterfolyamok kinyerése és visszaállítása a beszéd hullámformából. [38]
- **getF0**: Alapfrekvencia kinyerése a hullámformából. [39]
- **STRAIGHT**: Kevert gerjesztés jellegű beszédkódoló modellezése. [40]
- **hts_engine**: Paraméter generálás a generatív HMM modellekből és hullámforma generálás impulzus-zaj alapú beszédkódoló segítségével. [37]
- **ProfiVox:** fonetikus átirat elkészítése, mondathangsúlyok megállapítása. [5]
- **HunPOS**: szófajok megállapítása. [41]
- Magyar nyelvű, automatikus beszédfelismerő. [42]
- Kényszerített illesztést végző modul (forced alignment). [42]
- **Praat**: hullámforma és címkézés vizsgálata. [43]

4.4. A beszédszintézis paraméter beállításai

A kutatási célnak alárendelve egységes, a nemzetközi irodalomban is leggyakrabban használt beállításokat követtem [16]. Ezeken csak abban az esetben módosítottam, amennyiben azt célkitűzésem megkívánta. Kutatásaim kezdetén a rendszer alapbeállítása a következő volt:

Jellemző paraméterek kinyerése: Az emberi beszéd tulajdonságait figyelembe véve 16 kHz-es, 16 bit-es hullámformából nyertem ki a spektrális és alapfrekvencia (F_0) paramétereket. Impulzus-zaj gerjesztésű beszédkódoló esetén az *SPTK* segítségével céltól függően 10-24-ed rendű *MGC* (*Mel Generalized Spectrum*) felbontást használtam [44], az F_0 kontúrt pedig a *getF0* programmal határoztam meg. Kevert gerjesztésű beszédkódoló esetén a STRAIGHT segítségével 39-ed rendű spektrális felbontást használtam és szintén a STRAIGHT segítségével megállapítottam az alapfrekvenciát, továbbá 5 különböző tartományban a periodikus és aperiodikus részek arányát, illetve a Fourier magnitúdókat. Mindkét beszédkódoló esetén a fenti paraméterek delta és delta-delta összetevőit is kiszámítottam +/- 1 időablakban, delta esetén –0.5, 0, 0.5, delta-delta esetén 1, -2, 1 súlyozással.

Rejtett Markov-modellek tanítása: A modelleket – beszédkódolási módszertől függően – az előző pontban ismertetett paraméterekkel tanítottam. A spektrális paramétereket Gauss eloszlással, a F_0-t pedig többterű valószínűségi-eloszlással (*MSD-HMM*) modelleztem. A jellemző paraméterekkel 5 állapotú, környezetfüggő, balról-jobbra tartó HMM-eket tanítottam. Döntési fák segítségével csoportokba soroltam a betanított, környezetfüggő rejtett Markov-modelleket az *MDL* (*Minimum Description Length*) kritérium alapján. A döntési fák méretét maximum 4000 csomópontra korlátoztam. Külön döntési fát építettem a különböző jellemző paraméterek számára. Bizonyos esetekben a környezetfüggő címkék közé és a döntési fa építésébe bevettem a szófajok jelölését, melyet a HunPOS segítségével határoztam meg.

Jellemző paraméterek generálása: A felolvasandó bemeneti szöveget a ProfiVox fonetikus átalakítójával hangsorozattá alakítottam [5], majd elkészítettem hozzá a környezetfüggő címkéket [16]. A Tokuda és munkatársai által kidolgozott paramétergenerálási eljárást [35] megvalósító hts_engine segítségével létrehoztam a jellemző paramétersorozatokat.

Hullámforma előállítása a jellemző paraméterekből: A generált jellemző paramétersorozatból beszédkódolási eljárástól függően a hts_engine, az SPTK vagy a STRAIGHT segítségével készítettem hullámformát. A hullámforma az emberi beszéd tulajdonságait követve 16 biten lineárisan kvantált, 16 kHz mintavételi frekvenciával készült el.

4.5. Meghallgatásos tesztek

A gépi szövegfelolvasás területén általánosan elterjedt az eredmények *MOS* (*Mean Opinion Score*) és *CMOS* (*Comparison Mean Opinion Score*) alapú értékelése [45]. A természetes és a gépi beszéd minőségének osztályozására én is ezen módszereket alkalmaztam. MOS alapú teszt esetén a tesztalanyok a hangmintákat 5 elemű skálán 1-től 5-ig értékelhetik (1: legrosszabb, 5: legjobb), CMOS esetén pedig két minta közül kell a tesztalanyoknak eldönteniük, szintén 5 elemű skálán, hogy melyik minta tesz jobban eleget a teszt osztályozási kritériumának (pl. minőség, természetesség, érthetőség). A tesztek során bizonyos esetekben a tesztalanyokra bíztam a *minőség* fogalmának értelmezését. Ekkor az osztályzás általános visszajelzést ad arról, hogy a tesztalanyok mennyire tartják jónak vagy rossznak a kísérletben résztvevő rendszer által készített szintetizált beszédet. Ez esetben a rendszer értékelésében számos paraméter, például *természetesség*, *érthetőség*, a hang által a tesztalanyban keltett *érzelem* játszik szerepet. Más esetekben külön felhívtam a tesztalanyok figyelmét arra, hogy például a bemondás természetességét osztályozzák. Minden szubjektív meghallgatásos teszt során volt egy kiindulási mintahalmaz, amelynek egy kisebb részét ál-véletlen módon választottam ki. A kiindulási mintahalmaz és a kiválasztott részhalmaz mérete a szubjektív meghallgatásos teszt felépítésétől és a várható tesztalanyok számától függött. A legtöbb teszt internet alapú volt, ahol minden esetben rögzítettem a tesztalany *nemét*, *korát*, *beszédtechnológiai ismereteit* (szakértő / nem szakértő) és a meghallgatáshoz *használt eszközt (hangszóró / fejhallgató)*. A hangmintákat adott meghallgatásos teszt során minden egyes tesztalanynak más-más sorrendben játszottam le, így *zárva ki* az esetleges *emlékezeti hatásokat* [46]. A szubjektív meghallgatásos tesztek pontos paramétereit és a tesztalanyokról a fontosabb információkat az adott fejezetek során részletesen ismertetem. A MOS és CMOS típusú meghallgatásos teszteken elért pontszámok átlagát és az átlagok körüli 95%-os konfidencia-intervallumot grafikonon, illetve oszlopdiagramon ábrázoltam. A *szignifikanciát* minden esetben vizsgáltam: amennyiben két eredményt hasonlítottam össze, MOS meghallgatásos teszt esetén a várható értékre vonatkozó *két mintás párosított t-próbával*, CMOS meghallgatásos teszt *esetén egymintás t-próbával*. Amennyiben kettőnél több adatot kellett összehasonlítanom, ott *ANOVA analízist* használtam a szignifikancia vizsgálatára. Ha az ANOVA alapján észlelhető volt szignifikáns különbség, *post hoc* összehasonlítás céljából a *Tukey-féle* eljárást használtam. A szignifikancia teszteknél minden esetben 95%-os konfidencia szinttel számoltam ($\alpha=0.05$).

Több esetben a meghallgatásos tesztekben viszonylag alacsony (3 körüli) MOS értékek jöttek ki, míg más meghallgatásos teszteknél, hasonló HMM-TTS rendszerek esetében magasabb (3.5-4 körüli) értékek mutatkoztak. Ez azzal magyarázható, hogy az előbbi esetben természetes bemondóktól származó minták is szerepeltek a tesztben, míg az utóbbi esetben csak gépi rendszerek vettek részt. A tesztalanyok a természetes bemondóktól származó minták miatt a gépi beszédet rosszabb minőségűnek érzékelhetik.

5. Rejtett Markov-modell alapú szövegfelolvasó kialakítása és továbbfejlesztése magyar nyelvre

Ebben a kutatási témakörben négy témával foglalkoztam. Létrehoztam egy magyar nyelvű HMM-TTS rendszert, melyet összehasonlítottam a korábban készült magyar nyelvű gépi szövegfelolvasókkal. Második lépésként a rendszer minőségét a megkülönböztető jegyek bevezetésével javítottam. Következő lépésként megmutattam, hogy beszélőadaptáció segítségével – a beszélőfüggő esethez képest – akár szignifikánsan jobb minőségű gépi beszédelőállítás is lehetséges. Végül negyedik lépéseként megvizsgáltam, hogy a beszélőfüggő és beszélőadaptált esetekben a beszédkorpusz címkézési pontossága milyen mértékben befolyásolja az előállított gépi beszéd minőségét. Eredményeimet objektív mérésekkel és szubjektív meghallgatásos tesztekkel igazoltam.

5.1. Előzmények

A rejtett Markov-modell alapú beszédszintézis kidolgozása elsődlegesen Keiichi Tokuda (Nagoya Institute of Technology, Japán) nevéhez fűződik. A témához kapcsolódó első cikkek a 90-es évek végén, illetve az ezredforduló első éveiben jelentek meg [47,28,25,22]. Az igazi áttörést az jelentette, amikor a 2005-ös és 2006-os gépi szövegfelolvasók nemzetközi versenyét, a *Blizzard Challange*-et az addig legjobb minőségűnek tekintett elemkiválasztáson alapuló szövegfelolvasóval szemben a rejtett Markov-modell alapú megoldás nyerte [48,49]. A sikeres verseny után a beszédtechnológiával foglalkozó szakemberek felfigyeltek a módszerre, és a megoldást széles körben kezdték el kutatni.

Korábban számos idegen nyelven készült HMM-TTS [22,50,51,52], azonban magyar nyelvű megoldás az általam ismert publikációk alapján nem állt rendelkezésre. Tudomásom szerint a mai napig nem született publikáció más magyar nyelvű HMM-TTS rendszerről.

5.2. Beszélőfüggő rejtett Markov-modell alapú szövegfelolvasó magyar nyelven

A rejtett Markov-modell alapú szövegfelolvasó magyar nyelvre való kidolgozása interdiszciplináris kutatási feladat. A magyar nyelv fonetikai és fonológiai tulajdonságait kell modellezni komplex, gépi tanulást megvalósító statisztikai parametrikus rendszerben. A magyar nyelvű HMM-TTS létrehozásával kapcsolatos kutatásaimat ebben a fejezetben ismertetem.

5.2.1. Beszédkorpusz

Korábban nem állt rendelkezésre célirányos, a HMM-TTS számára megfelelő kialakítású magyar nyelvű beszédkorpusz a tanításhoz, ezért ennek létrehozása kulcsfontosságú volt. A HMM-TTS beszédkorpuszának tervezése során fontos szempont volt, hogy a magyar nyelv hangjaira jellemző, fonetikailag kiegyenlített mondatokat tartalmazzon. Az MTBA beszédkorpusz 500 beszélőtől átlagosan 6-7 perc hosszú, telefonon keresztül rögzített, fonetikailag gazdag hanganyagot tartalmaz, elsősorban beszédfelismerési célokra [53]. Az MTBA beszédkorpusz mondatait megvizsgáltam, és alkalmasnak találtam HMM-TTS

számára, azonban beszédfelismeréssel szemben a beszédszintézis esetében minden beszélőtől legalább egy órányi, stúdió körülmények között rögzített hanganyagra volt szükség (minimum 44 kHz, 16 bit). Ezért az MTBA mondataira támaszkodva végeztük a BME-TMIT Beszédtechnológiai Laboratóriumának munkatársaival a szükséges hanganyagok felmondását, rögzítését és a beszédkorpuszok kialakítását. A beszédkorpuszok kialakításakor figyelembe vettük az MTBA feldolgozása során publikált tapasztalatokat [54,55].

Öt beszédadatbázis készült, 4 férfi bemondó (életkoruk 60, 50, 30, 29 év) és egy női (33 év) hangból. Minden bemondó egységesen ugyanazt a szöveget olvasta fel. A felhasznált beszédkorpuszokat az 3. táblázat foglalja össze. A *BF* jelöli azt, hogy beszélőfüggő tanítási folyamatot használtam, az *FF1, FF2, FF3, FF4* jelű a férfi bemondók hangját, az *NOI1* jelű pedig a női bemondóét. Az öt beszédadatbázishoz a kutatás ebben a fázisában az eredeti szövegek automatikus módszerrel készített fonetikus átiratát használtam. A fonetikus átiratot a 4.3. fejezetben bemutatott módon készítettem el.

3. táblázat. A beszélőfüggő (BF) rejtett Markov-modell magyar nyelvre való kidolgozása során használt tanító beszédkorpuszok.

Beszélő	**Mondatszám**	**Időtartam**	**Feldolgozás**
BF-FF1: 1. férfi beszélő, beszélőfüggő	1936	190 perc	automatikus
BF-FF2: 2. férfi beszélő, beszélőfüggő	1938	137 perc	automatikus
BF-FF3: 3. férfi beszélő, beszélőfüggő	1941	170 perc	automatikus
BF-FF4: 4. férfi beszélő, beszélőfüggő	1938	214 perc	automatikus
BF-NŐ1: 1. női beszélő, beszélőfüggő	1937	128 perc	automatikus

5.2.2. A magyar nyelv modellezése HMM-TTS rendszerben

A HMM-TTS magyar nyelvre való kialakításához a következőkre volt szükség: megfelelő beszédkorpuszokra a HMM modellek tanításához, a ProfiVox fonetikus átíró és hangsúlyjelölő integrálására, a szófajok meghatározására statisztikai feldolgozáshoz, továbbá a magyar nyelv általános modellezésére a környezetfüggő címkék és a döntési fák kialakítása során. A nyelvek szerkezetei közötti különbségek miatt a nemzetközi megoldásokat csak iránymutatóként tudtam használni, továbbá a magyar nyelv szerkezetének korábbi leírásai nem voltak alkalmasak a HMM-TTS rendszerekhez. Ezért ezt is szükséges volt kidolgoznom.

A HMM-TTS számára alkalmas és a tanításhoz felhasználható beszédkorpuszokat az 5.2.1. fejezetben ismertettem.

A kutatás kezdetén megvizsgáltam a ProfiVox fonetikus átírót az MTBA véletlenszerű, vegyes témájú, fonetikusan kiegyenlített 500 mondatára, és a referenciának tekintett kézi fonetikus átirathoz képest kevesebb, mint 0.2%-os fonémahiba-arányt tapasztaltam. Az észlelt fonémahibák jelentős részében a hasonulás jelölése maradt el. A tanítás és szintézis során következetesen használt kvinfón alapú beszédhang modellnek köszönhetően a nem jelölt hasonulásokat is adott hangkörnyezetben meg tudja tanulni a HMM-TTS rendszer. A kismértékű fonémahiba-arányt a kvinfón alapú beszédhang modellezés mellett megengedhetően ítéltem. Ezért magyar nyelvű HMM-TTS rendszerek esetén a ProfiVox fonetikus átírót használtam.[5]

[5] Fontos megjegyezni, hogy a ProfiVox fonetikus átírója diád és triád alapú szövegfelolvasókhoz lett kialakítva. Ezekben a rendszerekben a jobb érthetőség céljából bizonyos esetek nem lettek fonetikusan átírva. Továbbá a hasonulás több hangkapcsolat esetén is akusztikai szinten, a diád és triád adatbázisokban van modellezve – ezzel analóg megközelítés a HMM-TTS esetén a kvinfón model.

A nyelvi modellezés terén a következőket végeztem el. Beszédhangok definiálása: a magyar nyelvben összesen 64 fonéma, 14 magánhangzó és 50 mássalhangzó található [56] szerint. Ezen forrásra támaszkodva a magyar nyelvű HMM-TTS megalkotása során a következő beszédhangokat definiáltam (a beszédhangokat itt betűjelekkel adom meg):

magánhangzók: *a,á,e,é,i,í,o,ó,ö,ő,u,ú,ü,ű*
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*
(rövid és hosszú változatok egyaránt)

A szüneteket is beszédhangként jelöltem. A szünetek típusát nem vettem figyelembe, mindössze egy fajta szünettípust definiáltam. Ezt a környezetfüggő modellek miatt tehetem meg, hiszen a szünet milyenségét nem csak a szünet, hanem annak a környezete is meghatározza.

A hanghosszúságot mássalhangzóknál a fonetikus átirat alapján, eltérő kóddal különböztetem meg. A magánhangzók esetén a rövid és hosszú hangok már eleve külön kóddal vannak jelölve.

Ezek alapján a nemzetközi szakirodalmat követve határoztam meg a magyar nyelvre jellemző legfontosabb környezetfüggő címkéket, melyeket a 4. táblázat mutat be [16]. A környezetfüggő címkék meghatározása beszédhang, szótag, szó, mondatrész és mondat szinten a feldolgozás során automatikus úton történik. A szótag határokat a szótagolás szabályai alapján állapítottam meg [57].

A szóhatárok megállapítását a szöveges forma szóközeinek a fonetikus átiratra való átvetítésével a kényszerített illesztés kimenete alapján végeztem el. Mondatrésznek tekintettem két vessző, pontosvessző, gondolatjel írásjelek közötti szövegrészt a mondatban, a mondatokat pedig pont, felkiáltójel vagy kérdőjel választja el egymástól.

A szótaghangsúlyok jelölésére a ProfiVox rendszer hangsúly meghatározó algoritmusát használtam. A szófajok meghatározását a 4.3. fejezet alapján végeztem el.

4. táblázat. A magyar nyelvű HMM-TTS rendszerében használt környezetfüggő címkék.

Szint	**Címke**
Beszédhangok	Az aktuális beszédhang, valamint a megelőző és követő két-két beszédhang (kvinfón).
Szótag	Szótaghangsúlyok jelölése (hangsúlyos / hangsúlytalan) az aktuális, az előző, és a követő szótagban. A beszédhangok száma az aktuális / előző / következő szótagban. A szótagok száma az előző / következő hangsúlyos szótagtól / szótagig. A szótag magánhangzója.
Szó	Szótagok száma az aktuális / előző / következő szóban. Az aktuális szó pozíciója a mondatrészben (elölről és hátulról is számítva). Az aktuális szó szófaja.
Mondatrész	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.

A döntési fák építéséhez a szükséges kérdéseket a fent ismertetett környezetfüggő címkék és a magyar nyelv sajátosságainak figyelembe vételével készítettem el [56,12,8], melyek

közül a legfontosabbakat az 5. táblázat tartalmazza. A döntési fák építése során a kvinfónban szereplő minden beszédhang esetén feltettem az 5. táblázatban szereplő tulajdonságokra vonatkozó kérdéseket. A hatékony gépi tanulás érdekében diszjunkt beszédhang halmazok esetén is feltettem a halmazokat külön-külön kiválasztó kérdéseket. Például a kvinfón minden elemére nem csak azt a kérdést tettem fel, hogy egy beszédhang mássalhangzó-e, hanem azt a kérdést is definiáltam, hogy a beszédhang magánhangzó-e.

A környezetfüggő címkék igen részletes módon írják le az adott beszédhang, szótag, szó és mondatrész elhelyezkedését a mondatban. Ennek megfelelően minden ilyen jellegű numerikus adatra rá kell kérdezni. A minél pontosabb modellezés érdekében minden esetben a nyelvi előfordulás lehetséges határáig három típusú kérdést tettem fel az adott elem számosságára vagy pozíciójára vonatkozóan:

1. az adott elem számossága vagy pozíciója egyenlő egy adott értékkel,
2. az adott elem számossága vagy pozíciója kisebb-egyenlő egy adott értékkel,
3. az adott elem számossága vagy pozíciója nagyobb-egyenlő egy adott értékkel.

A szófajra vonatkozó kérdések az alábbi opciókat tartalmazták: főnév, melléknév, ige, határozószó, indulatszó, kötőszó, számnév, determináns, névutó, igekötő, hangutánzó, elöljárószó, névelő.

Kevert gerjesztésű beszédkódoló számára alkalmas paramétereket nyertem ki (lásd 2.1.2. fejezet) és tanítottam a rejtett Markov-modellekkel. Ennek megfelelően a szintézis során kevert gerjesztésű beszédkódolót használtam.

A szintézis esetén először elkészítettem a bemeneti szöveg fonetikus átiratát, majd ebből generáltam a fentebb ismertetett környezetfüggő címkéket. Ezen címkék segítségével a generatív modellek megadták a bemeneti fonémasorhoz legjobban illeszkedő paraméterhalmazt, melyből kevert gerjesztésű beszédkódoló eljárással előállítottam a hullámformát.

A beszédhangoknak az 5.3. fejezetben megkülönböztető jegyek alapján egy új strukturális elrendezését hoztam létre és alkalmaztam HMM-TTS rendszerben, mely minőségbeli javuláshoz vezetett a jelen fejezetben bemutatott reprezentációhoz képest.

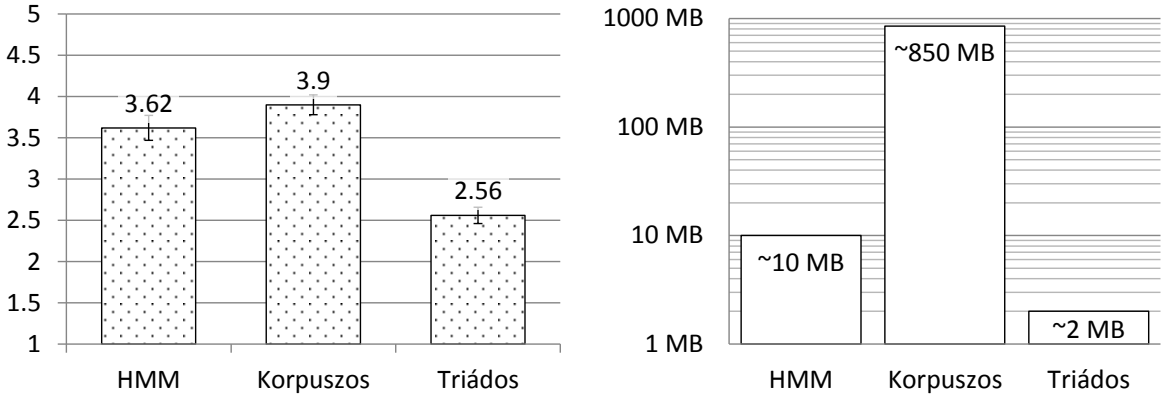
5. táblázat. A magyar nyelvű HMM-TTS döntési fáinak építéséhez használt tulajdonságok.

Szint	**Tulajdonság**
Beszédhangok	Magánhangzó / mássalhangzó. Zöngés / zöngétlen. Rövid / hosszú. Képzés helye a magánhangzóknál (hátul, középen, elöl). Nyelvállás a magánhangzókra (felső, középső, alsó). Ajakállás a magánhangzókra (kerekített, nem kerekített). Képzés módja mássalhangzóknál (zárhang / réshang / stb.).
Szótag	Hangsúlyos / hangsúlytalan. Az adott szótagra vonatkozó numerikus adatok (lásd 4. táblázat).
Szó	Az adott szó szófaja. Az adott szóra vonatkozó numerikus adatok (lásd 4. táblázat).
Mondatrész	Az adott mondatrészre vonatkozó numerikus adatok (lásd 4. táblázat).
Mondat	Az adott mondatra vonatkozó numerikus adatok (lásd 4. táblázat).

5.2.3. A szintetizált beszéd számszerű kiértékelése

Az elkészült rendszer minőségének szubjektív mérése céljából MOS típusú meghallgatásos tesztet készítettem. A BME-TMIT-en készült két korábbi rendszert (ProfiVox triádos szövegfelolvasó, ProfiVox beszédkorpusz alapú elem összefűzéses szövegfelolvasó) hasonlítottam össze az én megoldásommal. Tehát a kísérletben három magyar nyelvű TTS rendszer vett részt: egy triád alapú, egy korpuszos és a HMM alapú szövegfelolvasó. A teszt elején minden rendszertől 3-3 mondatot játszottam le véletlenszerűen, amelyeket a tesztalanyok még nem értékelhettek. Ez azt a célt szolgálta, hogy az alanyok hozzászokjanak a mesterséges hangokhoz, és tudják, hogy milyen minőségekre számíthatnak.

Ezután a három különböző rendszer által generált 29-29-29 mondatot játszottam le, minden tesztalany esetén más-más sorrendben. A tesztmondatok tartalma időjárás jelentés volt. (Fontos megjegyezni, hogy a korpuszos rendszert kifejezetten időjárás jelentések felolvasására tervezték, de másik két rendszert nem.) Minden rendszerrel ugyanazt a 29 mondatot generáltam, de ezen mondatok egyik rendszer esetén sem szerepeltek a tanító beszédkorpuszban. A meghallgatásos tesztet tizenegy fő, 8 férfi és 3 nő végezte el. A teszt internet alapú volt. A tesztalanyok átlag életkora 38 év volt, a legfiatalabb tesztalany 21, a legidősebb 64 éves volt. A tesztalanyok közül négyen beszédszakértők voltak. Ismert halláskárosodása egyik tesztalanynak sem volt. A tesztek eredményét a 9. ábra bal oldala mutatja be. Az eredmények alapján megállapítható, hogy a magyar nyelvű HMM-TTS minősége szignifikánsan nem eltérő a jelenleg legjobb minőséget jelentő korpusz alapú rendszerétől, de szignifikánsan jobb, mint a triádos rendszer. A 9. ábra jobb oldala a gépi beszéd előállításához szükséges futás idejű adatbázis méreteket szemlélteti. Ezek alapján a HMM-TTS futás idejű adatbázis mérete szignifikánsan kisebb a korpusz alapú rendszerénél, de a triád alapú szövegfelolvasóhoz képest nem mutat szignifikáns különbséget.



9. ábra. Gépi beszéd minőségének vizsgálata MOS meghallgatásos teszttel (bal oldal) és futás idejű adatbázis méretek (jobb oldal) a HMM-TTS, a korpuszos és a triád alapú szövegfelolvasó rendszerek esetén.

5.2.4. Konklúzió

A korpusz alapú, legjobb minőséget mutató szövegfelolvasó futás idejű adatbázisa mintegy 850 MByte (12 óra hanganyag), míg a HMM alapú rendszer futás idejű adatbázisa mintegy 10 MByte (2 óra hanganyag tanítása alapján). A korpusz alapú rendszer témaspecifikus területen működik csak megbízhatóan (időjárás jelentés). Ezzel szemben a HMM alapú rendszer általános témájú mondatokra is közel állandó minőséget produkál. (A beszédhangok ötállapotú parametrikus modellezéséből adódóan a HMM-TTS általános, témafüggetlen

szövegfelolvasó eljárásnak tekinthető. Az 5.3., 5.4., 5.5., 6.4., 6.5. és 6.6. fejezetek meghallgatásos tesztjei vegyes témájú mondatokat tartalmaztak (pl. hírek, időjárás jelentés, mese, árlista). Ezek a meghallgatásos tesztek is a témafüggetlenséget támasztják alá: a tématerületek között nem jelentkezett szignifikáns minőségbeli különbség a gépi beszédben.)

Jelen fejezet meghallgatásos tesztje során még témaspecifikus esetben sem volt szignifikáns minőségbeli különbség a korábban legjobb minőségűnek számító korpusz alapú szövegfelolvasó és a HMM-TTS között. Ezen okok indokolttá tették a HMM-TTS mélyebb vizsgálatát és lehetséges alkalmazását a korábbi szövegfelolvasó megoldásokkal szemben.

5.3. Megkülönböztető jegyek bevezetése a rejtett Markov-modell alapú szövegfelolvasó rendszerbe a minőségjavítás céljából

Az emberek nyelvtől függetlenül ugyanazokat a szerveiket használják a beszédhangok képzésére [56,12,8]. Míg a beszédhangképzés lehetősége univerzális, az egyes nyelvek hangzása különböző. A *megkülönböztető jegyek* (*Distinctive Features*) segítségével minden egyes beszédhangot nyelvfüggetlenül, a legtöbb esetben bináris értékek halmazával tudunk jellemezni. Mivel azonban sok esetben a beszédhangok nyelvenként különböznek, ezért minden nyelvhez külön meg kell határozni a megkülönböztető jegyek készletét és a beszédhangokat ennek megfelelően kell osztályokba sorolni. A legtöbb jegy megfeleltethető az általa definiált hang (fonéma) valamilyen fizikai (akusztikai) vagy fiziológiai (artikulációs) tulajdonságának [58,59,60]. A megkülönböztető jegyek fogalma és használata önmagában nem számít újnak a hazai és nemzetközi szakirodalomban, azonban HMM-TTS rendszerekben korábban – tudomásom szerint – még nemzetközi szinten sem alkalmazta őket senki. A megkülönböztető jegyek alapján a magyar beszédhangok osztályozását az általános nyelvészeti alapelveket és fogalmakat figyelembe véve, a HMM-TTS számára elsődlegesen mérnöki szempontok szerint határoztam meg. A megkülönböztető jegyeket a rejtett Markov-modell alapú szövegfelolvasó rendszerbe a *környezetfüggő címkék* és a *döntési fák* segítségével vezetem be. A megkülönböztető jegyek várhatóan általánosabb osztályokat hoznak létre, mint az előző fejezetben, az 5. táblázatban bemutatott jelölés. A megkülönböztető jegyekkel az 5.2. fejezetben bemutatott rendszert bővítettem ki.

5.3.1. Beszédkorpusz a megkülönböztető jegyek bevezetéséhez

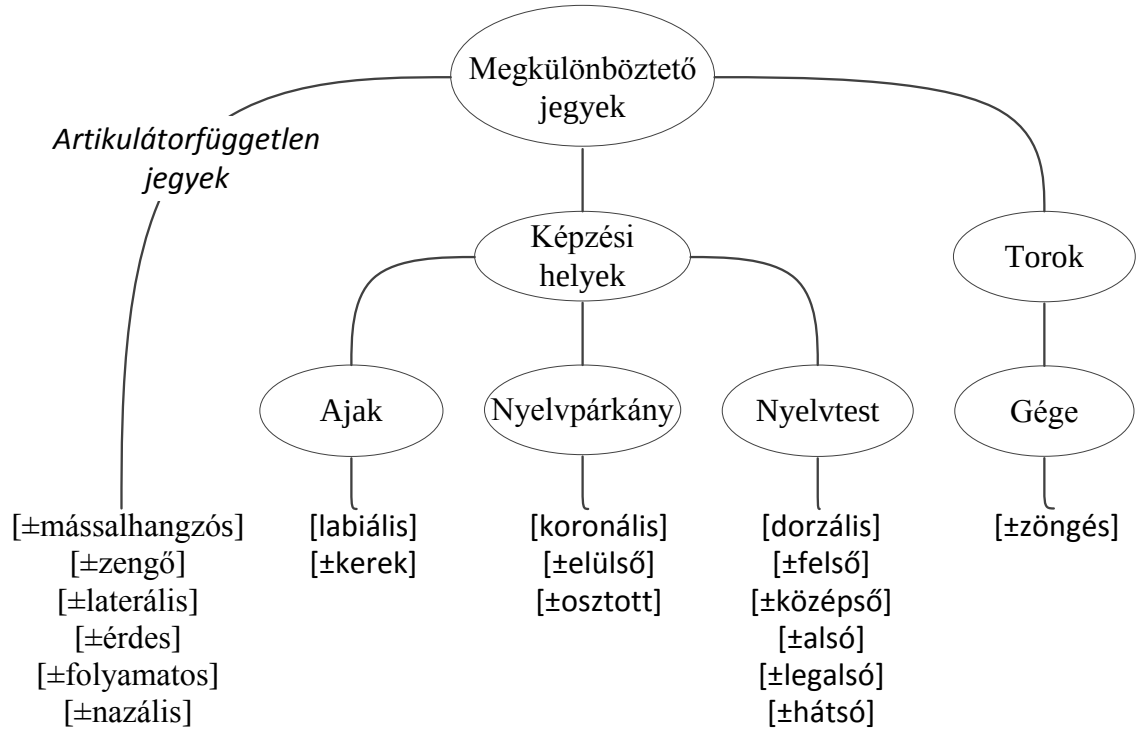
A megkülönböztető jegyek bevezetése során az 5.2.1. fejezetben ismertetett beszédkorpuszokkal dolgoztam. Ezáltal a megkülönböztető jegyek hatását az előző fejezetben bemutatott rendszerhez képest objektív és szubjektív módon tudtam mérni.

5.3.2. Megkülönböztető jegyek HMM-TTS rendszerben

A magyar beszédhangokat az 5.2.2. fejezetben bemutatott módon modelleztem a HMM-TTS rendszerben. Ezzel a módszerrel is már érthető, jó minőségű gépi beszédet állítottam elő. A megkülönböztető jegyek a beszédhangok egy általánosabb modellezését teszik lehetővé, segítségükkel pontosabban és nyelvfüggetlen(ebb) módon lehet leírni ugyanazokat a beszédhangokat. A jelen fejezetben bemutatásra kerülő rendszerezés során a korábbiaknak megfelelően a beszédhangokat betűkkel jelöltem, a rendszerezést mérnöki megközelítéssel, a HMM-TTS rendszerbe való bevezetés céljából az általános nyelvészeti fogalmak és szabályok

figyelembevételével végeztem. A megkülönböztető jegyek leírása során a beszédhangokat önmagukban ejtett fonémáknak veszem, a hangok egymásra hatását magában a leírásban nem veszem figyelemben [61,62]. A teljes rendszerben ez bizonyos mértékig a környezetfüggő címkék segítségével mégis modellezve van.

A *bináris* megkülönböztető jegyeket, amennyiben rendelkeznek az adott tulajdonsággal, a továbbiakban „+” jellel jelölöm, és „–” jellel, amennyiben nem rendelkeznek vele. Az *unáris* jelek jelenlétét „✓” jellel jelölöm. A „∅” pedig azt jelöli, hogy az aktuális *bináris* vagy *unáris* megkülönböztető jegy az adott hangra nincs értelmezve. A 10. ábra bemutatja a széles körben használt megkülönböztető jegyeket, amelyek alapján kidolgoztam a magyar HMM-TTS részére alkalmas leírást. A létrehozott leírást a következőkben mutatom be, illetve a 11. ábrán foglalom össze.



10. ábra. A szélesebb körben használt megkülönböztető jegyek; [61] alapján módosítva.

A 11. ábrán a könnyű átláthatóság érdekében a hosszú fonémákat nem jelöltem, ha létezik rövid, azonos megkülönböztető jegyekkel leírható párjuk. Ezekben az esetekben azért lehetnek ugyanazok a jegyek érvényesek a rövid és hosszú változatokra is, mivel a HMM-TTS számára bevezetésre kerülő hierarchiában nincsen a hangidőtartamra vonatkozó megkülönböztető jegy. A hanghosszúságokat ebben az esetben is az 5.2.2. fejezetben bemutatott módszerrel modelleztem.

A bevezetett megkülönböztető jegyek alapján kiterjesztettem a döntési fák építése során felhasznált kérdéseket. Minden egyes bináris megkülönböztető jegyhez két kérdést rendeltem. Az egyik kérdés a „+”, a másik a „–” értékkel jellemezhető csoportját jelöli ki az adott megkülönböztető jegynek. Az unáris jegyekhez egy-egy kérdés tartozik. Az aktuális beszédhanghoz tartozó kvinfón minden egyes eleméhez külön-külön, az adott megkülönböztető jegyre vonatkozó kérdéseket rendeltem.

A megkülönböztető jegyeket három fő osztályba sorolhatjuk: *artikulátorfüggetlen jellemzők, képzési helyek jellemzői és gégejellemző*.

5.3.2.1. Artikulátorfüggetlen jellemzők

Az a*rtikulátorfüggetlen jellemzők* azokat a megkülönböztető jegyeket jelölik ki, amelyek nem köthetők kizárólag egyetlen beszédszervhez [61]. Ebbe az osztályba a következő megkülönböztető jegyek tartoznak:

[±mássalhangzós]: mássalhangzós az a hang, melynek képzése során nagymértékű szűkület alakul ki a toldalékcsőben. A definíció alapján a „nagymértékű" nem jelöli meg egyértelműen a [+mássalhangzós] és [–mássalhangzós] közötti határt, ezért a nemzetközi irodalomra támaszkodva határoztam meg a két csoport elemeit. A [–mássalhangzós] csoportba soroltam a magánhangzókat és a siklóhangokat, azaz a félmagánhangzókat (magánhangzóra jellemző szerkezetű mássalhangzókat) és gégehangokat. A [+mássalhangzós] csoportba soroltam az összes megmaradt mássalhangzót (zárhangok, réshangok, zár-rés hangok, nazálisok, likvidák): *b, p, d, t, g, k, gy, ty, v, f, z, sz, zs, s, m, n, ny, l, r, dz, c, dzs, cs*. Ezek alapján a [–mássalhangzós] csoportban a magánhangzók (*a, á, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő*) mellett jelen van a *j* és *h* gégehang is [63].

[±zengő]: zengő az a hang, melynek képzése során a toldalékcsőben nyomásnövekedés nem lép fel. A [+zengő] csoportba tartoznak a magánhangzók, a nazálisok, a laterálisok, a pergőhangok, a félmagánhangzók és a gégehangok: *a, á, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő, j, h, m, n, ny, l, r*. A [–zengő] csoportba[6] tartoznak a zárhangok, a zár-rés hangok és a réshangok: *b, p, d, t, g, k, gy, ty, v, f, z, sz, zs, s, dz, c, dzs, cs*.

[±laterális]: laterális az a hang, melynek képzése során a toldalékcső középső részén a nyelv akadályt képez úgy, hogy a nyelv egy vagy mindkét oldalán levegőáramlás történik. A magyar nyelvben egyedül az *l* hang tartozik a [+laterális] a csoportba. A [–laterális] elemei az összes többi [+mássalhangzós]: *b, p, d, t, g, k, gy, ty, v, f, z, sz, zs, s, m, n, ny, r, dz, c, dzs, cs*. A többi hangra a [±laterális] jegy nem értelmezett.

[±érdes]: érdes az a hang, mely képzése során turbulens áramlás lép fel. Magyar nyelv esetén az [+érdes] csoportba a *v, f, z, sz, zs, s, dz, c, dzs, cs* hangok, az [–érdes] csoportba pedig a *b, p, d, t, gy, ty, g, k* zárhangok tartoznak. A többi hangra ez a jegy nem értelmezett.

[±folyamatos]: folyamatos az a hang, melynek képzése során a légáramlás nincsen teljes megszüntetve a szájüregben. Magyar nyelv esetén a [+folyamatos] csoportba az *l, r, v, f, z, sz, zs, s* hangok, a [–folyamatos] csoportba az *dz, c, dzs, cs, m, n, ny, b, p, d, t, g, k, gy, ty* tartoznak, hiszen az utóbbi esetben teljes levegőblokkolás van az orális csatornában. A további hangokra nem értelmezett ez a jegy.

[±nazális]: nazális az a hang, melynek képzése során a levegő az orrüregen keresztül áramlik. A nazális hangok létrehozása a lágy szájpad segítségével történik, ezért elméletben nem az artikulátorfüggetlen jegyekhez tartozik, a gyakorlatban mégis ezen jegyekhez sorolják, mert a nazálisoknak mindig van egy (aktív) artikulátoruk is (ajak, nyelv, stb.), amely a képzési helyet határozza meg. A magyar nyelv esetében az *m, n, ny* hangok a [+nazális], az összes többi hang pedig a [–nazális] csoportba tartozik (*b, p, d, t, g, k, gy, ty, v, f, z, sz, zs, s, l, r, dz, c, dzs, cs, j, h, a, á, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő)*.

5.3.2.2. Képzési helyek jellemzői

Képzési helyek jellemzői az artikulátorokkal (ajak, nyelvhegy, nyelvtest, nyelvtő, gége) kapcsolatos megkülönböztető jegyeket jelöli ki [61].

[labiális]: unáris jegy, labiális az a hang, melynek kialakítása során az aktív artikulátor az alsó ajak (a passzív lehet a felső ajak, vagy a felső fogsor). Ebbe a csoportba tartoznak a labiális zárhangok, zár-rés hangok, réshangok, nazálisok és siklóhangok. A magyar nyelvben

[6] Ezt a csoportot a nemzetközi szakirodalom obstruensnek, a magyar szakirodalom zörejhangnak nevezi.

a *b, p, v, f, m* beszédhangok és a kerekített magánhangzók tartoznak ide (*a, u, ú, ü, ű, o, ó, ö, ő)*. A többi hang esetében a jegy nincs értelmezve.

[±kerek]: kerek az a hang, melynek kialakításában a nyitott ajkak vesznek részt. Ez a megkülönböztető jegy a magánhangzókra vonatkozik, a [+kerek] csoportba az *a, u, ú, ü, ű, o, ó, ö, ő* hangok, a [–kerek] csoportba pedig az *á, e, é, i, í* hangok tartoznak. A többi hangra nem értelmezett.

[koronális]: unáris jegy, koronális hang esetén a nyelvpárkány (*apex*) elemelkedik a semleges pozícióból. Ide tartoznak a dentális, alveoláris, palato-alveoláris és palatális hangok: *d, t, gy, ty, z, sz, zs, s, j, n, ny, l, r, dz, c, dzs, cs*. A [koronális] jegy nincs a többi hangra értelmezve.

[±elülső]: elülső az a hang, melynek képzése során a szűkület a szájüreg palato-alveoláris területe előtt keletkezik. Csak a [koronális] hangokra értelmezhető a HMM-TTS számára kialakított jelen hierarchiában. Az [+elülső] csoportot a dentális és alveoláris hangok (*d, t, z, sz, n, l, r, dz, c*), a [–elülső] csoportot a palato-alveoláris és palatális hangok (*gy, ty, zs, s, j, ny, dzs, cs*) alkotják. A többi hangra nincs értelmezve az [±elülső] jegy.

[±osztott]: osztott az a hang, melynek képzése során a légáramlás irányában jelentős a szűkület kiterjedése. Az előző csoporthoz hasonlóan csak a [koronális] hangokra értelmezett a jelen hierarchiában. Eredetileg azért vezették be ezt a tulajdonságot, hogy a koronális hangokat két csoportra bontsák: laminális és apikális hangokra. Az [+osztott] csoporthoz tartoznak a palato-alveoláris és palatális hangok (*gy, ty, zs, s, j, ny, dzs, cs*). Az [–osztott] csoport elemei az alveoláris hangok (*d, t, z, sz, n, l, r, dz, c*). A jegy a többi hangra nincs értelmezve.

[dorzális]: unáris jegy, dorzális az a hang, melynek elsődleges képzőszerve a nyelvhát (dorsum). A magyarban a [–mássalhangzós] hangok között a magánhangzók[7] és a hátsó félmagánhangzók artikulációja (*a, á, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő*), a [+mássalhangzós] hangok esetén a velárisok (*g, k*) és uvulárisok artikulációja történik így. A többi hangra nincs értelmezve.

[±felső], [±középső], [±alsó], [±legalsó]: a nyelv függőleges mozgását írja le. A nyelv függőleges mozgását 3 megkülönböztető jeggyel szokás leírni, ekkor azonban nem elég egy megkülönböztető jegy a nyelvállás pontos meghatározásához. A HMM-TTS rendszerben való megvalósítás érdekében ezért megtartottam a négy szintet. Ez igaz, hogy redundanciát jelent, azonban a tanítás során a részletesebb leírás jobb általánosító képességet eredményezhet. A jelen hierarchiában a hangok besorolása a négy kategóriába a következőképp történt (a fel nem sorolt hangokra az adott jegy nincs értelmezve):

[+felső]: *i, í, u, ú, ü, ű*	[–felső]: *a, á, e, é, o, ó, ö, ő*
[+középső]: *é, o, ó, ö, ő*	[–középső]: *a, á, e, i, í, u, ú, ü, ű*
[+alsó]: *a, e*	[–alsó]: *á, é, i, í, u, ú, ü, ű, o, ó, ö, ő*
[+legalsó]: *á*	[–legalsó]: *a, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő*

[±hátsó]: a nyelv vízszintes mozgását írja le. A [+hátsó] a hátul (*a, á, u, ú, o, ó*), a [–hátsó] az elöl (*e, é, i, í, ü, ű, ö, ő*) képzett magánhangzókat írja le. A többi hangra a jegy nincs értelmezve.

5.3.2.3. Gégejellemző

A *gégejellemző* csoport azt a megkülönböztető jegyet tartalmazza, amely a gégeműködéssel van kapcsolatban [61].

[7] Ez vitatott kérdés. Vannak szerzők, akik szerint az elöl képzett magánhangzók koronálisok, és csak a hátul képzettek dorzálisok. Esetünkben a [±hátsó] jegy szolgálja azt a célt, hogy a dorzális képzési helyen belül a magánhangzóknak ezt a két osztályát elkülönítse.

[±zöngés]: zöngés az a hang, melynek képzése során a gerjesztés zöngés tulajdonsággal rendelkezik. A magyar nyelv esetében megkülönböztetünk zöngés, zörejes (turbulens) és kevert gerjesztésű beszédhangokat, így a [+zöngés] csoportba a zöngés és kevert gerjesztésű hangok (*b, d, g, gy, v, z, zs, j, m, n, ny, l, r, dz, dzs, a, á, e, é, i, í, u, ú, ü, ű, o, ó, ö, ő*), a [–zöngés] csoportba pedig a zörejes gerjesztésű hangok tartoznak (*p, t, k, ty, f, sz, s, h, c, cs*).

5.3.3. Számszerű kiértékelés

A megkülönböztető jegyek hatását objektív módon a döntési fák vizsgálatával, szubjektív módon meghallgatásos tesztekkel mértem. Az alábbiakban ismertetem a mérések eredményeit.

5.3.3.1. Döntési fák vizsgálata

Objektív módon a döntési fák elemzésével vizsgáltam a megkülönböztető jegyek hatását. Az egyes jellemző paraméterfolyamokhoz külön-külön döntési fák tartoznak (pl. spektrális paraméterek, alapfrekvencia, stb.). Amennyiben adott környezetfüggő címkére vonatkozó kérdés a döntési fában szerepel a döntési fa építésének leálló kritériumai mellett, az annyit jelent, hogy az adott paraméterfolyam esetén ez a kérdés a döntési fa adott szintjén a legjobban írja le a tanító beszédkorpuszt. A döntési fák mérete és felépítése jelentős hatással van a gépi beszéd minőségére. Továbbá a tanító beszédkorpuszt a döntési fa legfelsőbb szintjein jelenlévő kérdések jellemzik a legjobban, ezért fontos ezen szintek vizsgálata is.

A HMM-TTS rendszer tanítása után először elemeztem a kialakult döntési fákat. Ezáltal objektív módon lehet vizsgálni, hogy a gépi tanulás a megkülönböztető jegyeket mennyire találja „fontosnak”. A megkülönböztető jegyek előfordulási gyakoriságát az 5.2. fejezetben ismertetett kevert gerjesztésű magyar nyelvű HMM-TTS rendszerben a 6. és 7. táblázat mutatja be. A könnyebb átláthatóság érdekében a paraméterfolyamok öt állapotához tartozó döntési fákat a táblázatokban összegezve jelenítem meg. Mindkét táblázat esetén az eredményeket az 5.2.1. fejezetben ismertetett beszédkorpuszokkal végzett tanítások átlagaként számoltam ki. Tehát a táblázatokban minden érték mögött valóságban 5 állapot × 5 beszélő = 25 döntési fa áll. Mindkét táblázat fejlécében a kevert gerjesztés jellemző paraméterfolyamait láthatjuk.

A 6. táblázatot megvizsgálva látható, hogy az egyes paraméterfolyamokat milyen mértékben befolyásolták a megkülönböztető jegyek. A megkülönböztető jegyek a legnagyobb hatással a spektrális paraméterekre voltak, de hatásuk a többi paraméterfolyam esetén is jelentős. A 7. táblázat a döntési fákban előforduló tíz leggyakoribb megkülönböztető jegyet mutatja be az adott paraméterfolyamokra a felhasznált beszédkorpuszok esetén. A táblázat alapján az artikulátorfüggetlen jegyek több mint 50%-os arányt képviseltek.

	Artikulátorfüggetlen jegyek						Képzési helyek jegyei											Gége
	mássalhangzós	zengő	laterális	érdes	folyamatos	nazális	labiális	kerek	koronális	elülső	osztott	dorzális	felső	középső	alsó	legalsó	hátsó	zöngés
b	+	-	-	-	-	-	✓	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	+
p	+	-	-	-	-	-	✓	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	-
d	+	-	-	-	-	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
t	+	-	-	-	-	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	-
g	+	-	-	-	-	-	∅	∅	∅	∅	∅	✓	∅	∅	∅	∅	∅	+
k	+	-	-	-	-	-	∅	∅	∅	∅	∅	✓	∅	∅	∅	∅	∅	-
gy	+	-	-	-	-	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	+
ty	+	-	-	-	-	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	-
v	+	-	-	+	+	-	✓	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	+
f	+	-	-	+	+	-	✓	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	-
z	+	-	-	+	+	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
sz	+	-	-	+	+	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	-
zs	+	-	-	+	+	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	+
s	+	-	-	+	+	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	-
j	-	+	∅	∅	∅	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	+
h	-	+	∅	∅	∅	-	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	-
m	+	+	-	∅	-	+	✓	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	+
n	+	+	-	∅	-	+	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
ny	+	+	-	∅	-	+	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	+
l	+	+	+	∅	+	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
r	+	+	-	∅	+	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
dz	+	-	-	+	-	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	+
c	+	-	-	+	-	-	∅	∅	✓	+	-	∅	∅	∅	∅	∅	∅	-
dzs	+	-	-	+	-	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	+
cs	+	-	-	+	-	-	∅	∅	✓	-	+	∅	∅	∅	∅	∅	∅	-
a	-	+	∅	∅	∅	-	✓	+	∅	∅	∅	✓	-	-	+	-	+	+
á	-	+	∅	∅	∅	-	∅	-	∅	∅	∅	✓	-	-	-	+	+	+
e	-	+	∅	∅	∅	-	∅	-	∅	∅	∅	✓	-	-	+	-	-	+
é	-	+	∅	∅	∅	-	∅	-	∅	∅	∅	✓	-	+	-	-	-	+
i	-	+	∅	∅	∅	-	∅	-	∅	∅	∅	✓	+	-	-	-	-	+
u	-	+	∅	∅	∅	-	✓	+	∅	∅	∅	✓	+	-	-	-	+	+
ü	-	+	∅	∅	∅	-	✓	+	∅	∅	∅	✓	+	-	-	-	-	+
o	-	+	∅	∅	∅	-	✓	+	∅	∅	∅	✓	-	+	-	-	+	+
ö	-	+	∅	∅	∅	-	✓	+	∅	∅	∅	✓	-	+	-	-	-	+

11. ábra. Megkülönböztető jegyek egy lehetséges kialakítása a magyar nyelv beszédhangjaira HMM-TTS számára. A hosszú fonémák rövid párjukkal azonos megkülönböztető jegyekkel írhatóak le. Jelölés: „+”: rendelkezik a tulajdonsággal; „–”: nem rendelkezik a tulajdonsággal; „✓”: unáris jegy jelen van; „∅”: nem értelmezett.

6. táblázat. A megkülönböztető jegyek aránya a kevert gerjesztésű, magyar nyelvű HMM-TTS döntési fáiban.

	Alap-frekvencia	**Spektrális paraméterek**	**Hang-időtartam**	**Zönge-erősség**	**Σ**
Csomópontok száma	13821	3272	1153	4486	22732
Megkülönböztető jegyek	2664	1411	314	1018	5407
Megk. jegyek előfordulása	19.3%	43.1%	27.2%	22.7%	23.8%

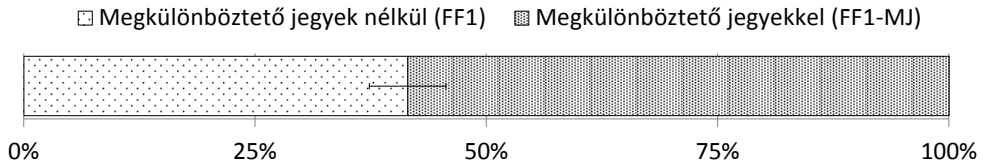
7. táblázat. A döntési fákban előforduló tíz leggyakoribb megkülönböztető jegy (az artikulátorfüggetlen jegyek dőlten, félkövéren vannak kiemelve).

	Alapfrekvencia	**Spektrális paraméterek**	**Hangidőtartamok**	**Zöngeerősségek**
1.	***zengő***	hátsó	***laterális***	***zengő***
2.	alsó	***zengő***	***zengő***	***folyamatos***
3.	***folyamatos***	kerek	***folyamatos***	***nazális***
4.	***laterális***	***nazális***	kerek	felső
5.	***nazális***	koronális	zöngés	kerek
6.	kerek	alsó	***nazális***	***mássalhangzós***
7.	zöngés	magas	alacsony	***laterális***
8.	magas	***laterális***	***érdes***	zöngés
9.	***érdes***	***folyamatos***	***mássalhangzós***	***érdes***
10.	hátsó	labiális	alsó	alsó

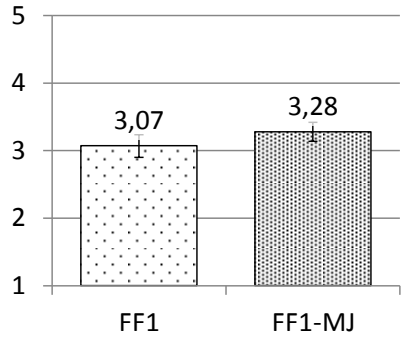
5.3.3.2. Meghallgatásos teszt

Szubjektív módon CMOS és MOS meghallgatásos tesztekkel mértem a megkülönböztető jegyek hatását az 5.2. fejezetben bemutatott rendszerhez képest. A meghallgatásos tesztben az *FF1* beszélő alapján végzett eredeti tanítás (jelölés: *FF1*), illetve a megkülönböztető jegyek bevezetése után elvégzett tanítás alapján (jelölés: *FF1-MJ*) létrehozott rendszerek vettek részt. A teszt első része egy MOS típusú, második része egy CMOS típusú pár-összehasonlításos teszt volt. Az *FF1* és *FF1-MJ* rendszerrel is ugyanazt a 40 mondatot generáltam le, és ebből a 40 mondatból véletlenszerűen, egyenletes eloszlással választottam ki 20-20-at a meghallgatásos teszt két részéhez.

A pár összehasonlítás során az emlékezeti hatások elkerülése céljából véletlenszerű volt, hogy az *FF1* vagy az *FF1-MJ* mintát játszottam-e le előbb. A tesztalanynak az *FF1* és *FF1-MJ* minta közül kellett eldöntenie, hogy melyik hangzása természetesebb. Az első részben a tesztalanyoknak 10 mintapárt kellett értékelniük. A meghallgatásos teszt második részében összesen 20 mintának a természetességét kellett értékelniük. A meghallgatásos tesztet tizenhárman végezték el, 12 férfi és 1 nő. Internet alapú volt a teszt, az átlag életkor 30 év volt, a legfiatalabb tesztalany 22, a legidősebb 60 éves volt. 7 tesztalany beszédszakértő volt, és a tesztalanyoknak nem volt ismert halláskárosodása. A tesztek eredményeit a 12. ábra és a 13. ábra mutatja. A 12. ábra alapján az *FF1-MJ*-t az *FF1*-el szemben többen választották, ami minőségjavulásra utal. A 13. ábra alapján is jobban teljesített az *FF1-MJ*.



12. ábra. A megkülönböztető jegyek hatásának vizsgálata CMOS meghallgatásos teszttel.



13. ábra. A megkülönböztető jegyek hatásának vizsgálata MOS meghallgatásos teszttel.

5.3.4. Konklúzió

A megkülönböztető jegyek bevezetésével a gépi beszéd minősége javul és a megkülönböztető jegyek jelentős mértékben megjelentek a döntési fák felépítésében is. Ennek gyakorlati hasznosságán – a jobb beszédminőségen – túl az elvi jelentősége az, hogy a jegyek segítségével a HMM-TTS működését közelebb hoztam az emberi beszédkeltés fiziológiájához.

5.4. Beszélőadaptált rejtett Markov-modell alapú szövegfelolvasó létrehozása magyar nyelven

A rejtett Markov-modell alapú gépi szövegfelolvasó egyik legfontosabb előnye a többi technológiával szemben az, hogy lehetőséget ad a beszélőadaptációra. A beszélőadaptáció azt jelenti, hogy egy célszemély hangkarakterisztikájához hasonlóra szabom a gépi beszéd hangszínezetét, alapfrekvencia menetét és a hangidőtartamokat. A HMM-TTS parametrikus tulajdonságából adódóan lehetséges a paraméterhalmazokat egy adott beszélő értékeinek az irányába „eltolni”. Fontos kérdés, hogy milyen lesz a minősége a beszélőadaptált rendszernek a beszélőfüggő tanításhoz képest? Elvárásom az volt, hogy legalább olyan minőségű rendszert kell tudni beszélőadaptáció segítségével létrehozni, mintha beszélőfüggő tanítást végeznék. Számos nemzetközi megoldás létezik HMM-TTS beszélőadaptációra [22,50,51], azonban az 5.2. és 5.3. fejezetekben ismertetett kutatások eredményeire támaszkodva új megoldást hoztam létre.

5.4.1. Beszédkorpusz

Az átlaghang létrehozásához a 8. táblázatban szereplő adatbázisokat használtam alapként. Ezután a beszélőadaptációt további két, az átlaghangban nem szereplő egy férfi és egy női beszélő hangjával végeztem el. Az adaptációs beszédkorpuszok tulajdonságait a 9.

táblázatban foglalom össze. A *BA-FF1* az *FF1* férfi beszélőre adaptált esetet, a *BA-NŐ1* a *NŐ1* női beszélő beszélőadaptált tanításához felhasznált beszédkorpuszát jelöli. A beszélőfüggő esettel való összehasonlítás érdekében az 5.3. fejezetben létrejött szövegfelolvasó rendszerek közül kiválasztottam ugyanezen *FF1* és *NŐ1* hangokhoz tartozó HMM-TTS-eket. Ezeket a korábbiaknak megfelelően *BF-FF1* (1. férfi beszélő, beszélőfüggő eset) és *BF-NŐ1* (1. női beszélő, beszélőfüggő eset) módon jelölöm.

8. táblázat. A beszélőadaptált rejtett Markov-modell átlaghangjához használt beszédkorpuszok.

Beszélő	**Mondatszám**	**Időtartam**	**Feldolgozás**
FF2: 2. férfi beszélő	1938	137 perc	automatikus
FF3: 3. férfi beszélő	1941	170 perc	automatikus
FF4: 4. férfi beszélő	1938	214 perc	automatikus
FF5: 5. férfi beszélő	1992	198 perc	automatikus
NŐ2: 2. női beszélő	1992	193 perc	automatikus
Összesen	9801	912 perc	automatikus

9. táblázat. Az adaptációhoz használt beszédkorpuszok.

Beszélő	**Mondatszám**	**Időtartam**	**Feldolgozás**
BA-FF1: 1. férfi beszélő, beszélőadaptált	104	10 perc	automatikus
BA-NŐ1: 1. női beszélő, beszélőadaptált	164	11 perc	automatikus

5.4.2. Beszélőadaptáció magyar nyelven

A beszélőadaptációhoz kevert gerjesztésű beszédkódolót használtam. Ennek megfelelően a 2.1.2. fejezetben ismertetett paraméterfolyamokat nyerem ki mind az átlaghang tanításához szükséges, mind pedig a beszélőadaptációhoz használt beszédkorpuszokból (8. és 9. táblázat). Az átlaghang tanítása során a korábbiakhoz hasonlóan automatikus módszerrel készített hanghatár-jelölést és fonetikus átiratot használtam. A környezetfüggő címkéknek és a rájuk vonatkozó kérdéseknek az 5.3. fejezet alapján megkülönböztető jegyekkel kibővített halmazát használtam. Érdekes kérdés a nemek közötti jelentős frekvenciakülönbségek miatt, hogy az átlaghang tanításához férfi, női, vagy kevert hangokat érdemes-e használni. Amennyiben nagy mennyiségű férfi és női hanganyag áll rendelkezésre, a leghatékonyabb megoldást a nemtől függő 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 a rendszert mind férfi, mind női hangra. Lehetséges ellentétes nemű átlaghangból a másik nem hangjára adaptálni, azonban Isogai és munkatársai beszámolnak arról, hogy ez jelentős minőség és természetesség csökkenést okoz a nemtől függő átlaghanghoz képest [64]. Yamagishi egy olyan módszert mutatott be, melynek segítségével kevert nemű átlaghangot a nem függő átlaghanghoz képest minimális minőség és természetesség romlás mellett lehet női- és férfihangra adaptálni [33]. Ezen utóbbi irodalom alapján végeztem el az eloszlások normalizálását az átlaghang létrehozása során.

Miután elkészültek az átlaghang HMM modelljei, az adaptációs beszédkorpuszokból kinyert jellemző paraméterek irányába CMLLR alapú eljárással transzformáltam az átlaghang modelljeit [18]. Az adaptáció során a paraméterfolyamokat különböző súlyozással módosítottam. Az alapfrekvencia, zöngeerősség és időzítési paraméterekhez tartozó paraméterfolyamok esetén „közelebb” maradtam az átlaghang értékeihez (kisebb súly). A

spektrális együtthatókhoz tartozó paraméterfolyam esetén pedig a célbeszélő paraméterfolyamának az értékeihez kerültem „közelebb” (nagyobb súly). A súlyozási faktorok a nemzetközi szakirodalmat követve a következőek voltak [33]:

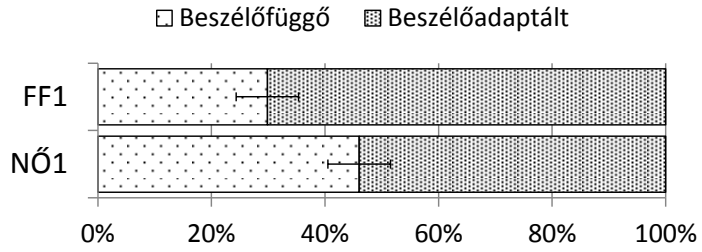
- alapfrekvencia: *0.2*
- zöngeerősség: *0.2*
- időzítési paraméterek: *0.2*
- spektrális együtthatók: *1*

Erre a súlyozásra azért volt szükség, mert a kisebb súlyú paraméterfolyamoknál az adaptációs beszédkorpuszban a teljes eltolás esetén a túltanulás veszélye állna fel a viszonylag rövid hosszúságú adaptációs korpusz miatt. Ez annyit jelentene, hogy az adaptációs beszédkorpuszban előforduló környezetfüggő címkék esetén jól teljesítene a rendszer, az általánosító képessége azonban rossz volna. Kompromisszumos megoldás, hogy az átlaghanghoz „közelebb” maradva megmarad a rendszer általánosító képessége, de a CMLLR miatt bizonyos mértékben mégis visszaadja a célbeszélőre jellemző hangkarakterisztikát. Nagyobb méretű adaptációs beszédkorpusz használatával lehetséges növelni a súlyozási faktorokat, ami a célbeszélő paraméterfolyam értékeihez közelebb eső értékeket eredményezne. A spektrális együtthatók esetén a paraméterfolyam felépítéséből adódóan a viszonylag rövid adaptációs beszédkorpuszból is nagyszámú paramétert tudtam kinyerni, így nem áll fenn a túltanulás veszélye.

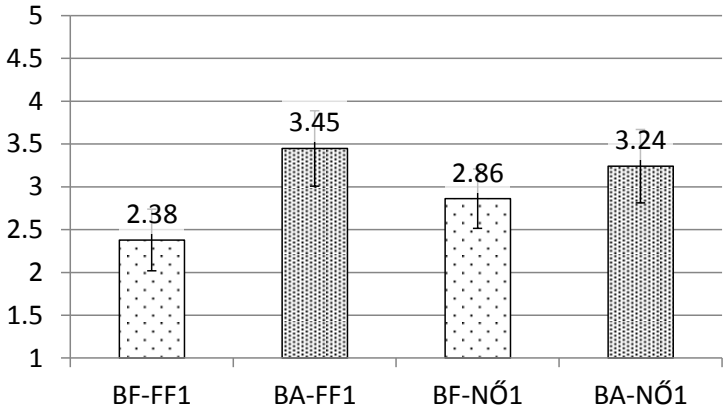
5.4.3. Számszerű kiértékelés

A beszélőadaptáció során létrejött *BA-FF1* és *BA-NŐ1* rendszerek minőségét MOS és CMOS típusú meghallgatásos tesztekkel határoztam meg. Annak érdekében, hogy a korábbi beszélőfüggő esettel összevessem a beszélőadaptált megoldást, a meghallgatásos tesztekben a *BF-FF1* és *BF-NŐ1* rendszerek hangját is használtam. A MOS teszt során ezek természetességét egymástól függetlenül mértem, a CMOS teszt során pedig a *BA-FF1* ↔ *BF-FF1* és *BA-NŐ1* ↔ *BF-NŐ1* rendszereket párban hasonlítottam össze. A meghallgatásos teszthez mind a négy rendszerrel száz-száz hangmintát készítettem, majd ebből egyenletes eloszlás szerint, ál-véletlen módon válogattam ki a teszt során felhasznált hangminták halmazait.

A meghallgatásos teszt első részében mindegyik rendszerből 10-10 hangmintát választottam ki. Egy tesztelő a 10 mintából 2 mintapárt hallgatott meg a beszélőfüggő férfi és női, illetve beszélőadaptált férfi és női HMM rendszerek által generált hangminták közül. Páronként a minták szövege azonos volt. A tesztalanyok ötelemű skálán osztályozták, hogy a mintapár mintái ugyanolyan természetesek-e, illetve hogy valamelyik mintát kicsit vagy sokkal természetesebbnek érzik-e a másik mintánál. A meghallgatásos teszt második felében az adaptációs célszemély eredeti, természetes bemondásához hasonlították a tesztalanyok a célszemély hangjára adaptált gépi beszédet annak értékelésére, hogy mennyire adja vissza annak hangzását. Ebben az esetben a négy rendszerből egy tesztalany esetén összesen 10 hangmintát használtam fel. Ezeket 1-től 5-ig terjedő skálán kellett osztályoznia. Az 1-es osztályzat itt azt jelentette, hogy egyáltalán nem adja vissza az eredeti beszélő hangkarakterét, az 5-ös pedig, hogy a szintetizált hangminta összetéveszthető az eredeti beszélővel. A mintákat a négy rendszerből egyenletes eloszlás szerint válogattam ki. A tesztet összesen 29-en végezték el, 14 férfi és 15 nő. Az átlagéletkor 31 év volt, a legfiatalabb tesztalany 20, a legidősebb 65 éves volt. 6 tesztalany beszédszakértő volt. A teszt internet alapú volt, böngészőből lehetett kitölteni, a hangminták MP3 kódolással voltak tárolva 128 kbps, 16 bit minőségben.



14. ábra. Beszélőfüggő és beszélőadaptált HMM-TTS hangok vizsgálata szubjektív CMOS meghallgatásos teszttel.



15. ábra. Beszélőfüggő és beszélőadaptált HMM-TTS hangok természetességének vizsgálata szubjektív MOS meghallgatásos teszttel.

A meghallgatásos tesztek eredményeit a 14. ábra és a 15. ábra mutatja be. A 14. ábra a pár összehasonlítást értékeli, melyen megfigyelhető, hogy az *FF1* és a *NŐ1* esetén is a tesztalanyok a beszélőadaptált rendszert részesítették előnyben. A 15. ábra a rendszerek természetességét vizsgáló MOS tesztek eredményeit ismerteti. A beszélőadaptált rendszerek (*BA-FF1, BA-NŐ1*) ebben az esetben is jobban teljesítettek a beszélőfüggő rendszerekkel (*BF-FF1, BF-NŐ1*) szemben. Mindkét esetben szignifikáns különbség mutatkozott a beszélőadaptált eset javára. A férfi beszélő esetén a MOS és CMOS tesztek is nagyobb minőségbeli javulást mutattak, mint a női esetben. Ez az eltérés elsősorban abból adódhat, hogy az átlaghang négy férfi és egy női beszédkorpusz alapján került kialakításra. Várhatóan az átlaghangban a női beszélők számának növelésével nagyobb mértékű minőségbeli javulást lehetne elérni női hanggal történő beszélőadaptáció esetén is.

5.4.4. Konklúzió

Az eredmény elsőre meglepőnek tűnhet, hiszen látszólag kisebb beszédkorpusszal jobb minőségű gépi beszéd érhető el, mint nagyobbal. Beszélőadaptált esetben mintegy 10-15 percnyi hangfelvétel elég ahhoz, hogy a HMM-TTS az eredeti beszélő hangkarakterisztikájához hasonló hangon szólaljon meg, azonban az átlaghang előállításához az 5.4.1. fejezetben bemutatott beszédkorpuszokat is felhasználtam. Ezáltal az átlaghang elkészítéséhez használt „tudás” mérete nagyságrenddel több volt a beszélőfüggő tanításhoz használt beszédkorpuszhoz képest, ennek köszönhetően a nyelvi és prozódiai sajátosságokat a HMM-ek jobban be tudják tanulni. Amennyiben a beszélőfüggő tanítás jelentősen hosszabb (több 10 órányi) felvételekkel történne, vélhetőleg az így létrejött HMM-TTS elérné (esetleg

túl is lépné) a beszélőadaptált esetek minőségét. Azonban ekkor nem használnánk ki a HMM-TTS egyik legfontosabb előnyét, hogy új beszédhangok viszonylag kis munkával hozhatóak létre.

Ismereteim szerint nemzetközi szinten is új megoldás a megkülönböztető jegyek használata beszélőadaptáció során. Az eredménynek nagy előnye van alkalmazás szempontjából, mert ez alapján a jövőben az új HMM-TTS hangkarakterisztikák létrehozásához rövidebb hangfelvételek készítése elegendő.

5.5. A címkézési pontosság növelésének hatása a rejtett Markov-modell alapú szövegfelolvasó beszédminőségére

A beszélőfüggő és beszélőadaptált magyar nyelvű HMM-TTS-ek létrehozása után a beszédkorpusz címkézési pontossága és a HMM-TTS beszédminősége közötti összefüggést vizsgáltam. Arra a kérdésre szerettem volna választ adni, hogy vajon javul-e a gépi beszéd minősége, ha a beszédkorpusz automatikusan végzett címkézését utólagosan kézi ellenőrzésnek vetem alá és az előforduló címkehibák számát így gyakorlatilag nullára csökkentem.

Elemkiválasztáson alapú rendszerek esetén már vizsgálták a beszédkorpuszban jelenlévő hibák és a szubjektív minőség közötti összefüggést [65,66]. Az eredmény, hogy ezen szövegfelolvasó rendszerekben az adatbázisban jelentkező szegmentálási és fonémajelölési hibák jelentős hatással vannak a gépi beszéd minőségére. HMM-TTS-ek esetén azt elemezték, hogy milyen hatással van az adatbázis mérete a szubjektív beszédminőségre beszélőfüggő [67] és beszélőadaptált esetekben [18]. A nemzetközi kutatások során általánosan bevett szokás, hogy a hanghatárokat automatikus úton állapítják meg, a fonetikus átirat kézi ellenőrzésére a gyakorlatban a jelentős többletmunka miatt általában nem térnek ki [32,52]. Arról, hogy a tanító és adaptációs adatbázisban jelenlévő fonéma- és szegmentálási hibák hogyan hatnak a HMM-TTS rendszerek által generált beszéd minőségére tudomásom szerint korábban nem született publikáció.

5.5.1. Beszédkorpusz

A kutatás során mind a beszélőfüggő, mind a beszélőadaptált eseteket vizsgáltam. Az előző fejezetek alapján ezeket *BF* (beszélőfüggő) és *BA* (beszélőadaptált) módon jelölöm. A *kézi* a kézi javításnak alávetett, az *auto* pedig az eredeti, automatikus eljárással készült beszédkorpuszokra utal.

A beszélőfüggő tanításhoz az 5.2. fejezetben bemutatott *BF-FF1* és *BF-NŐ1* beszédkorpuszokat, és azoknak kézileg ellenőrzött változatait használtam. Beszélőadaptált esetben az 5.4. fejezetben ismertetett korpuszok segítségével készítettem el az átlaghangot. Az adaptációs beszédkorpuszt az *FF1* és *NŐ1* korpuszból származtattam.

Az átlaghang tanítása során a szövegek fonetikus átírása az eredeti szövegből kizárólag automatikus úton történt. (Az ideális eset az volna, ha az átlaghang adatbázisai mind kézzel ellenőrzöttek lennének. Kutatásom ezen szakaszában azonban még nem állt rendelkezésre javított hanghatárokkal az összes beszédkorpusz kézi átirata, ezért a konzekvens adatbázis-építés elvét alkalmazva mindegyik beszédkorpuszból az automatikus átiratot használtam.)

5.5.2. A beszédkorpusz kézi ellenőrzése HMM-TTS rendszerben

A kézi ellenőrzés a beszédkorpusz két részére terjedt ki: a fonetikus átiratra és a szegmentálásra. A fonetikus átiratban a fonéma tévesztések javítása, a szegmentálás esetén pedig a hanghatárok pontos jelölése volt a cél. A kézi javítás során a BME-TMIT Beszédtechnológiai Laboratórium munkatársai voltak a segítségemre. Ahogyan korábban írtam, a tanító beszédkorpuszt az MTBA-ra támaszkodva alakítottuk ki, ezért a kézi javítás során is figyelembe vettük az MTBA feldolgozásáról publikált tapasztalatokat [54,55].

5.5.2.1. Fonémahiba-arány

Módszert dolgoztam ki a *fonémahiba-arány* (Phone Error Rate, PER) megállapítására. A következő hibafajtákat kezeltem:

- A bemondó mást olvas fel, mint ami a szövegben van. 1. példa: szöveg: „és”, fonetikai átirat: „és”, kimondva: „s”. 2. példa: szöveg: „lehtetlen”, átirat: „lehtetlen”, kimondva: „lehetetlen”.
- A fonetikai átíró mást jelöl, mint az elhangzott elem. Példa: szöveg: „900”, fonetikai átirat: „kilencszáz”, kimondva: „kilencáz”.

A hibák javításának módszere: a gépileg felcímkézett adatállományokat kézi ellenőrzésnek vetem alá. Ennek során a fonetikus átiratot és az elhangzott hanganyagot egybevetettem, és a hibákat javítottam a megfelelő helyen. Minden esetben a kézzel javított fonetikus átiratot tekintem referenciának. A kézzel javított átiratról feltételezem, hogy hangról hangra azt tartalmazza, ami a beszédkorpusz hanganyagában szerepel. A fonémahiba-arány meghatározása során a referenciához képest vizsgálom az automatikusan készített fonetikus átiratot.

A fonémahiba-arány számítása közben a fontos jellemzők:

- *Fonémák száma*: összes fonéma száma a beszédkorpuszban.
- *Helyes fonémák száma*: a kézi átirathoz képest mennyi fonéma azonos.
- *Törlések*: az automatikus átiratban a kézi átirathoz képest fonéma törlések száma.
- *Helyettesítések*: az automatikus átiratban a kézi átirathoz képest a fonéma helyettesítések száma.
- *Beszúrások*: az automatikus átiratban a kézi átirathoz képest az új fonémák száma.

A javítások száma a törlések, helyettesítések és beszúrások számának összege. A fonémahiba-arányt a következőképp számolom [68]:

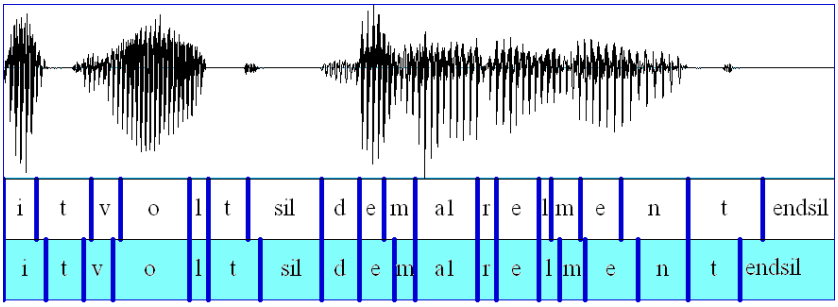
$$PER = \frac{Fonémák\ száma - Helyes\ fonémák\ száma}{Fonémák\ száma} \times 100\% \tag{36}$$

5.5.2.2. Szegmentálás

Szegmentáláson azt a folyamatot értem, melynek során a hangfelvételek hullámformáján minden hang és szünet kezdetét (hanghatár) címkével jelöljük. Szegmentálási szempontból kétféle címkét különböztetünk meg, a hang elejét jelző jelölést és a szünetek elején elhelyezett címkét (hangsor belseji szünet és hangsor végi). Ezeket a hanghatárokat a korábbiakkal azonos módon kényszerített illesztéssel határoztam meg. Fontos figyelembe venni, hogy hibátlan fonetikus átirat esetén is lehetnek hibás hanghatár címkék. Ezért a kényszerített

illesztés szegmentálási eredményét kézi ellenőrzéssel szükséges volt javítani (16. ábra). Mivel a kézi ellenőrzés nagy emberi erőforrást igényel, ezért a kutatás ezen fázisában ezt a felhasznált beszédkorpuszoknak csak egy részében tudtuk elvégezni (*FF1* és *NŐ1*).

A hanghatár hibákat osztályokba soroltam attól függően, hogy hány *ms* eltérés van a referencia és az automatikus hanghatár meghatározás között a következőképp: *10–19 ms*, *20–29 ms*, *30–39 ms*, *40–49 ms*, *50–59 ms*, *60–69 ms*, *70–79 ms*, *80–89 ms*, *90 ms*-nál több. A szegmentálás végeztével előáll(nak) a HMM tanításhoz, illetve adaptációhoz szükséges beszédkorpusz(ok).



16. ábra. A hanghatárok automatikus, gépi elhelyezése (fent) és a kézi javítás eredménye (lent) az „*Itt volt, de már elment.”* mondatban.

5.5.3. Számszerű kiértékelés

Az eredmények ismertetését három részre bontom:

- megvizsgálom a fonémahiba-arányt (PER) a beszédkorpuszokban,
- ismertetem az automatikus és a kézi szegmentálás különbségeit, illetve
- meghallgatásos tesztekkel vizsgálom a pontos címkézés hatását a gépi beszéd minőségére HMM-TTS rendszerekben.

5.5.3.1. Fonémahiba-arány vizsgálata

A kísérletek során az *FF1* és *NŐ1* hangfelvételekből kiindulva összesen 8 féle beszédkorpuszt vizsgáltam: 4 korpuszt a beszélőfüggő tanításhoz és 4 korpuszt a beszélőadaptált változathoz (ezek a korábbi korpuszokat, illetve azok részhalmazait tartalmazták). Az adaptációs beszédkorpuszt az *FF1* és *NŐ1* korpuszból származtattam és elemeit úgy választottam ki, hogy benne az adott *FF1* és *NŐ1*-ből az összes törlés, helyettesítés és beszúrás szerepeljen. Ezáltal az automatikus fonetikus átiratból a legtöbb hibát tartalmazó részeket tartottam meg, továbbá véletlen módon választottam ki hozzá annyi hanganyagot, hogy 10 perc körüli időtartam álljon rendelkezésre. Az eredményeket a 10. táblázat mutatja be. A táblázat alapján megállapíthatjuk, hogy beszélőfüggő esetekben a teljes korpusz nagyságához képest a hibák száma elenyésző (0.83%, 0.52%), illetve, hogy az adaptációs beszédkorpusz esetén már nagyobb mértékű hibaaránnyal kell számolnunk (15.5%, 6%).

10. táblázat. A beszélőfüggő és beszélőadaptált tanításhoz felhasznált korpuszok tulajdonságai kézi és automatikus címkézés esetén a fonémákra vonatkozva.

	BF-FF1-kézi	**BF-FF1-auto**	**BF-NŐ1-kézi**	**BF-NŐ1-auto**	**BA-FF1-kézi**	**BA-FF1-auto**	**BA-NŐ1-kézi**	**BA-NŐ1-auto**
Mondatszám	1936	1936	1937	1937	104	104	164	164
Időtartam [perc]	190	190	128	128	10	10	11	11
Fonémák száma	80964	81053	80893	81058	4281	4370	6934	7099
Helyes fonémák száma	80964	80380	80893	80663	4281	3697	6934	6674
Törlések	-	32	-	51	-	32	-	51
Helyettesítések	-	57	-	114	-	57	-	114
Beszúrások	-	584	-	260	-	584	-	260
Javítások száma	-	673	-	425	-	673	-	425
PER	0%	0.83%	0%	0.52%	0%	15.5%	0%	6%

5.5.3.2. A szegmentálás vizsgálata

A hanghatár hibák összehasonlításából származó eredményeket beszélőfüggő tanítás és beszélőadaptáció során használt beszédkorpusz esetén a 11. táblázat mutatja be. A táblázat értékei azt adják meg, hogy hány darab beszédhang esetén kellett az oszlopok fejlécében megadott mértékű hanghatár javítást elvégezni. A táblázatban szereplő fonémák számát a 10. táblázatban feltüntetett értékekkel összevetve megállapítható, hogy a BF-FF1, BF-NŐ1, BA-FF1 és BA-NŐ1 beszédkorpuszok hanghatárai rendre 31%, 21%, 29%, 17.5% százalékban lettek javítva.

11. táblázat. Az automatikus hanghatár meghatározás pontossága.

	10-19ms	**20-29ms**	**30-39ms**	**40-49ms**	**50-59ms**	**>60ms**
BF-FF1-automatikus	17238	5355	1664	555	188	169
BF-NŐ1-automatikus	13854	2317	656	227	91	92
BA-FF1-automatikus	884	264	86	25	8	6
BA-NŐ1-automatikus	1037	148	36	15	7	4

5.5.3.3. Meghallgatásos teszt

Az eredmények kiértékelése céljából meghallgatásos tesztet állítottam össze. Ez esetben a kutatás célja a kialakult szövegfelolvasó rendszerek hangjának összehasonlítása volt annak érdekében, hogy megállapíthassam, hogy a kézi címkézés okoz-e minőségbeli javulást beszélőfüggő és beszélőadaptált esetekben.

A kísérlet során összesen nyolc féle HMM-TTS rendszert vizsgáltam: négy beszélőfüggő tanítás alapján előállt rendszert (*BF-FF1-kézi, BF-FF1-auto, BF-NŐ1-kézi, BF-NŐ1-auto*), valamint négy beszélőadaptált HMM-TTS rendszert (*BA-FF1-kézi, BA-FF1-auto, BA-NŐ1-kézi, BA-NŐ1-auto*). A meghallgatásos teszt két részből állt. Az első részben a tesztalanyok a hangminták hangzásának természetességét osztályozták a következő rendszerek esetén (CMOS): *BF-FF1-kézi ↔ BF-FF1-auto, BF-NŐ1-kézi ↔ BF-NŐ1-auto, BA-FF1-kézi ↔*

BA-FF1-auto, BA-NŐ1-kézi ↔ BA-NŐ1-auto. Minden rendszerből 10 hangmintát használtam fel, páronként a minták szövege azonos volt. Egy tesztelő beszélőfüggő férfi és női, illetve beszélőadaptált férfi és női HMM rendszerek által generált 10 hangminta közül 2 mintapárt hallgatott meg. A tesztalanyok ötelemű skálán osztályozták, hogy a mintapár mintái ugyanolyan természetesek-e, illetve hogy valamelyik mintát kicsit vagy sokkal természetesebbnek érzik-e a másik mintánál. A meghallgatásos teszt második felében a tesztalanyok az eredeti beszélő természetes bemondásához hasonlították a gépi beszédet. Ebben az esetben mind a nyolc rendszer részt vett a tesztben. A meghallgatásos teszt második felében is összesen 10 hangmintát használtam fel, egy tesztalany egy rendszerből egy hangmintát hallgatott meg, és 1-től 5-ig kellett osztályozni, hogy a minták mennyire adják vissza az eredeti beszélő hangkarakterét.

A tesztet összesen 29-en végezték el, 14 férfi és 15 nő, közülük 6 tesztalany beszédszakértő volt. Az átlagéletkor 31 év volt, a legfiatalabb tesztalany 20, a legidősebb 65 éves volt. A teszt internet alapú volt, a hangminták MP3 kódolással voltak tárolva 128 kbps, 16 bit minőségben.

A meghallgatásos teszt eredményeit a 17. (első rész) és a 18. ábra (második rész) mutatja. A 17. ábra alapján a kézi és az automatikus címkézés között három esetben nem volt szignifikáns különbség. Egyedül a férfi beszélővel tanított beszélőfüggő rendszer esetén volt tapasztalható szignifikáns eltérés a kézi ellenőrzés javára (legfelső sor). A 18. ábrán az eredményeket páronként megfigyelve (első-második, harmadik-negyedik, ötödik-hatodik, hetedik-nyolcadik oszlop) láthatjuk, hogy a természetes bemondáshoz képest sem volt szignifikáns eltérés a kézi és automatikus módszerek között. A beszélőadaptált női hang esetén a kézi ellenőrzés minimálisan rosszabb pontokat kapott, mint az automatikus módszer. Ez a nem várt eredmény a mérést terhelő zajból származhat, nagyobb számú tesztalany esetén ez a különbség vélhetőleg megszűnik.

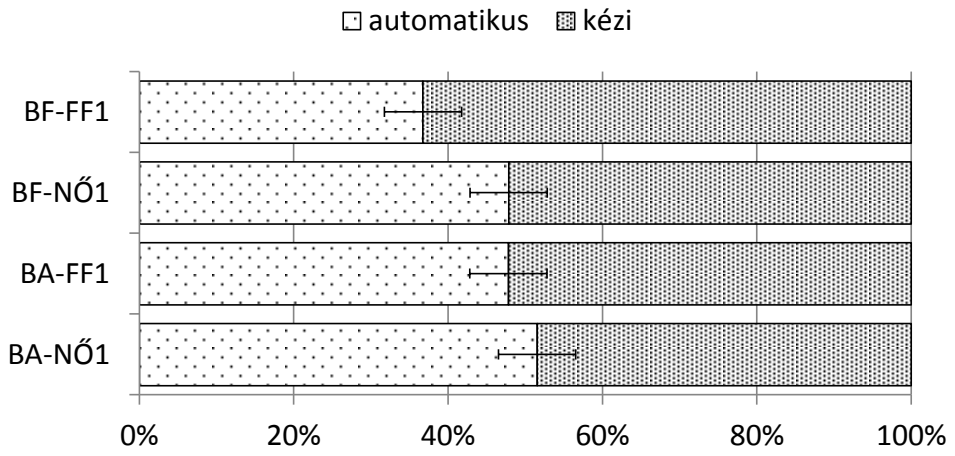
5.5.4. Konklúzió

Az eredmények alapján léteznek olyan esetek, amikor a rengeteg fáradságos, nagy szakértelmet és precizitást kívánó kézi címkézést meg lehet spórolni, hiszen a manuálisan végzett munka nem okoz szükségszerűen szignifikáns minőségjavulást HMM-TTS rendszerekben.

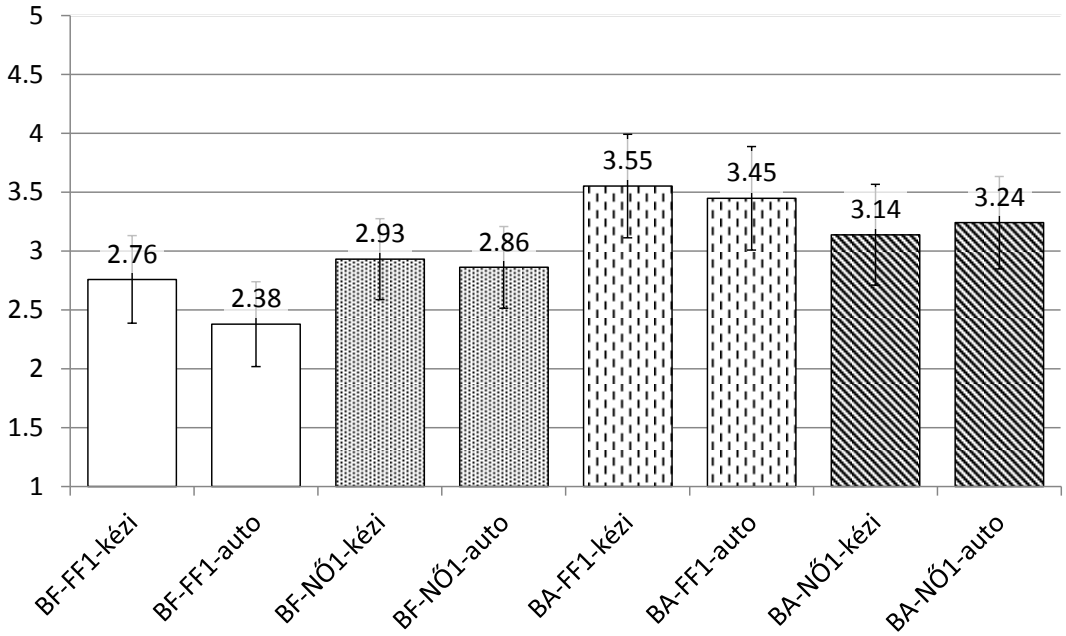
A kézi ellenőrzés egyedül egy esetben okozott szignifikáns minőségjavulást az automatikussal szemben (*BF-FF1*), azonban a meghallgatásos teszt második része során ugyanebben az esetben már nem volt szignifikáns minőségbeli különbség észlelhető. Ez azzal magyarázható, hogy a teszt első részében a két rendszer között összehasonlítás történt (relatív minősítés), ami a kis eltéréseket is jelentősen fel tudja nagyítani. A teszt második felében a rendszereket külön-külön vizsgáltam meg (abszolút minősítés), és az így kapott eredményeket hasonlítottam össze, és ekkor már a *BF-FF1* esetén sem mutatkozott szignifikáns különbség. Az összes többi esetben (*BA-FF1, BF-NŐ1, BA-NŐ1*) a jelenlegi magyar nyelven működő rejtett Markov-modell alapú szövegfelolvasó rendszerben a kézi hanghatár-jelölés és kézi fonetikus átirat nem okozott szignifikáns javulást. Ezt azzal lehet magyarázni, hogy a HMM-TTS generatív modelleket épít a tanító beszédkorpuszból. A beszédkorpuszban szereplő fonémák nagyszámú előfordulásának köszönhetően az automatikus eljárások által a rendszerben jelenlévő hibák a statisztikai módszerek hatására kiátlagolódnak, és így nem okoznak észlelhető minőségromlást az elkészített szintetizált beszédben.

Beszélőadaptált esetben CMOS és MOS tesztek sem mutattak szignifikáns különbséget az automatikus és kézi módszerek között. Ez elsősorban a több mint 15 órányi hanganyaggal tanított átlaghang modellel magyarázható: a beszédhangoknak és a magyar nyelvnek ez a modell egy olyan reprezentációját tartalmazta, melyet az adaptációs hanganyagban szereplő

fonéma és hanghatár hibák nem rontottak szignifikáns mértékben, miközben a HMM-TTS rátanult a célszemély hangkarakterére és beszédritmusára. Az eredmények alapján fontos vizsgálni, hogy mi az a hibahatár, ami már jelentős minőségromlást okoz beszélőadaptált esetben. Abban az esetben, ha a generatív modellek nagyobb hibák esetén is még megfelelő minőséget képesek produkálni, akkor automatikus beszédfelismerő és kényszerített illesztés alapján lehetőség nyílik új beszédhangok létrehozására felügyelet nélküli beszélőadaptációval. Ezzel a kérdéssel a következő fejezetben foglalkozom.



17. ábra. Páros összehasonlítás az automatikusan címkézett és a kézileg utólagosan javított adatbázisok felhasználásával készített gépi szövegfelolvasó rendszerek között.



18. ábra. Az automatikusan címkézett és a kézzel utólagosan javított adatbázisok felhasználásával készült gépi szövegfelolvasó rendszerek MOS teszttel való mérése.

5.6. Összegzés

Jelen fejezet keretében a rejtett Markov-modell alapú szövegfelolvasó magyar nyelvre történt bevezetésével és vizsgálatával kapcsolatos kutatásaimat, eredményeimet mutattam be. A fejezetben megfogalmazott új kutatási eredményeimet téziscsoportba foglalt tézisek formájában a 8. fejezetben mutatom be.

A kutatás első lépéseként a magyar nyelvnek a mesterséges tanuló algoritmusok számára alkalmas leírását dolgoztam ki és ennek felhasználásával létrehoztam egy magyar nyelvű HMM-TTS rendszert. Meghallgatásos tesztekkel igazoltam, hogy a korábbi, legjobb minőségű gépi beszédet előállító korpusz alapú szövegfelolvasótól szignifikánsan nem eltérő minőségű gépi beszéd állítható elő kevesebb erőforrással (I.1. tézis: 5.2. fejezet). Következő lépésként a HMM-TTS rendszerek számára modellt dolgoztam ki megkülönböztető jegyek alapú hangosztályozásra, és megmutattam, hogy segítségükkel lehet a gépi beszéd minőségén javítani. Rámutattam, hogy a megkülönböztető jegyek a fentiek mellett közelebb hozzák a HMM-TTS rendszert az emberi beszédkeltés fiziológiájához (I.2. tézis: 5.3. fejezet). A létrehozott modellezési eljárást beszélőadaptációra használtam és megmutattam, hogy ezáltal a beszélőfüggő esethez képest lehetséges szignifikánsan jobb minőségű gépi beszédet előállítani (I.3. tézis: 5.4. fejezet). A fejezet utolsó pontjában pedig megvizsgáltam, hogy az automatikus címkézés kézi javítása növeli-e a gépi beszéd minőségét (I.4. tézis: 5.5. fejezet). Az eredmény meglepő: a szaktudást és sok emberi munkaórát kívánó kézi javítás nem okozott szükségszerűen szignifikáns minőségjavulást.

6. Rejtett Markov-modell alapú szövegfelolvasó felügyelet nélküli adaptációja

A HMM-TTS egyik nagy előnye, hogy a hangszínezet és a prozódia a paramétereken keresztül módosítható. Ahogyan korábban bemutattam, beszélőadaptáció esetén egy célbeszélő hangjához és stílusához lehet szabni a gépi beszédet [33]. A beszélőinterpoláció során már meglévő gépi beszédhangokból új hangkarakterisztikákat tudunk létrehozni [20]. A paraméterek módosításával akár érzelmes beszéd kifejezésére is lehetőség van [69].

Rejtett Markov-modell alapú szövegfelolvasók esetén a beszélőadaptáció a 2.5.2. fejezetben bemutatott módon történik. Röviden összefoglalva: szükség van a célbeszélőtől származó mintegy 10 perc hosszúságú megfelelő minőségű hangfelvételre[8], a hangfelvétel időzítési paramétereire (hang-, szó- és mondathatárokat beleértve), továbbá a hangfelvétel fonetikus átiratára.

Vajon lehetséges-e emberi beavatkozás nélkül új HMM-TTS beszédhangok létrehozása? Ha igen, akkor a beszélőadaptációhoz elég, ha csupán a hanganyag áll rendelkezésünkre az adott célbeszélőtől és ennek segítségével automatikusan elkészíthető a célbeszélő hangkarakterisztikáját visszaadó gépi beszédhang. Az 5.5. fejezetben bemutatott kutatásom eredményei alapján megállapítottam, hogy a HMM-TTS adaptációs beszédkorpuszában jelenlévő hibák nem feltétlen okoznak szignifikáns minőségromlást a hibátlannak tekintett beszédkorpusszal szemben. Ez felveti azt a kérdést, hogy vajon mi az a fonémahiba-arány, amely már szignifikáns minőségromlást okoz? Lehetséges-e pontatlan fonetikus átirat esetén is a HMM-TTS beszélőadaptációja.

Jelen fejezetben megvizsgálom a felügyelet nélküli beszélőadaptáció lehetőségét rejtett Markov-modell alapú szövegfelolvasó rendszerekben. A korábbi nemzetközi eredmények feldolgozása alapján új, mások által még tudomásom szerint nem alkalmazott megközelítést dolgoztam ki. Ennek lényege, hogy a beszédfelismerő kimenetét használtam fel az adaptációs beszédkorpusz fonetikus átiratának alapjául. A fonéma határokat ebből kényszerített illesztéssel, automatikus módon kontrollált beam-el állapítottam meg. Először kidolgoztam a felügyelet nélküli adaptációs eljárást, majd megvizsgáltam a különböző, a HMM-TTS beszélőadaptációja szempontjából fontos aspektusokat, illetve kiértékeltem a megoldás hatékonyságát. Az eredmények igazolását CMOS és MOS alapú szubjektív meghallgatásos teszttel végeztem. Kutatásom tartalmaz nyelv specifikus elemeket, azonban az alkalmazott módszertan nem nyelvfüggő.

6.1. Áttekintés

A felismerési pontosság növelése céljából a HMM alapú beszédfelismerőknél vegyesen használják a felügyelt és felügyelet nélküli adaptációt. A beszédfelismerők felügyelet nélküli adaptációja kevesebb kézi munkát igényel, de nagyobb adaptációs adatbázisra van szükség a felügyelt esethez képest [42].

A rejtett Markov-modell alapú szövegfelolvasók területén már korábban is születtek felügyelet nélküli adaptációra megoldások. A King és munkatársai által kidolgozott eljárás szerint lehetséges kézi feldolgozás nélkül új HMM beszédhang karakterek létrehozása oly

[8] A HMM-TTS szempontjából megfelelő minőségűnek lehet azt a hangfelvételt tartani, amely legalább 16000 Hz / 8 bit-en lett rögzítve, a beszéd folyamatos a tartalomnak megfelelő hosszúságú szünetekkel, továbbá a beszéd tiszta, jól érthető és nincs háttérzaj.

módon, hogy a beszélőadaptációhoz használt transzformációs mátrixot pusztán a trifón modellek alapján, azaz csak a fonémák szerint állítják elő [70]. Ezzel a módszerrel a környezetfüggő címkék által hordozott információt elhagyjuk az adaptációs folyamatból. A tanulmányban bemutatott eredmények alapján a rendszer alacsony minőségű gépi beszédet képes csak előállítani (MOS: 1.9-2.5). Feltételezhető, hogy ez elsősorban a környezetfüggő címkék elhagyásából adódott, és nem a felügyelet nélküli adaptációból.

Gibson két lépcsős döntési fa építéssel valósítja meg a felügyelet nélküli beszélőadaptációt [71]. Ez a módszer a környezetfüggő HMM-ekhez tartozó döntési fákat először szegmentális szempontok alapján, majd szupraszegmentális szinten alakítja ki. Az eredmények alapján nem tapasztalható minőségbeli különbség a felügyelt és a felügyelet nélküli eset között. Ebben a kutatásban beszédkorpuszként a beszédfelismerő számára gyűjtött adatbázist használtak. Ez nem előnyös a HMM-TTS számára, mert sok beszélőtől rövid hangfelvételeket tartalmaz. Ezzel magyarázható a bemutatott rendszer igen rossz minősége (MOS: 1.9-2.0). Ilyen alacsony MOS érték mellett nem mérhető megbízhatóan a felügyelet nélküli adaptáció és a felügyelt megoldás közötti minőségbeli különbség.

Később Gibson és munkatársai a két lépcsős döntési fa elvét terjesztették ki nyelvközi felügyelet nélküli beszélőadaptációra [72]. A nyelvközi beszélőadaptáció azt jelenti, hogy egy adott nyelven beszélő gépi hangot (pl. magyar) más nyelven beszélő ember (pl. angol) hangkarakterisztikájához hasonlóvá alakítják [73]. Ezt a kiinduló és a célbeszélő rendszerekben lévő fonémák vagy állapotok összerendelésével érik el. A fenti tanulmány ezt az eljárást terjeszti ki úgy, hogy kézi beavatkozást ne igényeljen. Az eredmények azt mutatták meg, hogy nyelvközi adaptáció esetén a két lépcsős döntési fa építéssel hasonló minőséget lehet elérni, mint azonos nyelven történő beszélőadaptáció során. A gépi beszéd minősége ebben az esetben is igen alacsony volt (MOS: 1-3). Oura és munkatársai a nyelvközi adaptáció más megközelítését használja [74]. Ebben az esetben a kiinduló és a célbeszélő közötti kapcsolatot Kullback-Leibler divergencia alapján kiválasztott állapot párok segítségével teremtették meg. A célbeszélő korpuszának fonetikus átiratát ebben az esetben beszédfelismerő készíti el, azonban a kutatásnak ez csak egy eszköze és nem a célja. A cikk nem vizsgálja a beszédfelismerő által okozott hibák hatásait az adaptáció minőségére. Ez a rendszer nagyméretű (2000 mondat), témaspecifikus adaptációs beszédkorpusz esetén közepes minőségű gépi beszédet képes előállítani (MOS: 3). A döntési fa kiszorításos (*decision tree marginalization*) módszer alkalmazásával szintén nyelvközi, felügyelet nélküli adaptáció lehetséges [75]. A forrás azonos nyelven történő felügyelet nélküli adaptációról is beszámol. Ebben az esetben is használták a beszédfelismerő kimenetét, azonban ebből csak triádos modelleket építettek, a környezetfüggő címkék halmazát kvinfónokkal nem készítették el. Mindazonáltal az eredményük ígéretes, átlagosan 75%-os fonémahiba-arány mellett is volt olyan eset, hogy a felügyelt és a felügyelet nélküli adaptáció minősége hasonló volt (MOS: 2.5-3).

Több nemzetközi kutatás is foglalkozik felügyelet nélküli beszélőadaptációval, ezek közül már van, amelyik bizonyos mértékben használja a beszédfelismerő kimenetét. Azonban tudomásom szerint a beszédfelismerő kimenetén fellépő fonéma tévesztések hatásával teljes környezetfüggő leírás mellett a felügyelet nélküli beszélőadaptációban elsőként foglalkozom. Kutatásom fő kérdése, hogy lehetséges-e pusztán a beszédfelismerő kimenetére támaszkodva a felügyelt esethez hasonló minőséget elérni?

Célom az volt, hogy amennyiben létezik megoldás, akkor a felügyelet nélküli beszélőadaptáció teljesen szabadon történhessen. Ezért a kutatást félspontán beszéddel végeztem.

6.2. A félspontán beszéd

A HMM-TTS beszédkorpuszaiban leggyakrabban használt tervezett beszéd igen szigorú megkötést jelentene a felügyelet nélküli adaptáció esetén, és így erősen korlátozná a megoldás hatékonyságát. Továbbá a tervezett beszéd esetén feltételezhető, hogy megvan a szöveg fonetikus átirata, így értelmét vesztené a felügyelet nélküli adaptáció. Ezeket figyelembe véve a kutatási munkámat a spontán beszéd jellemzőit is magában hordozó félspontán beszéddel végeztem.

Félspontánnak, vagy fél-reproduktívnak nevezzük azt a beszédtevékenységet, mely az élőszó igényével lép fel, de rendszerint az előadó által egy korábban megfogalmazott, elmondásra szánt írott szövegen alapszik [76]. A beszélő nem ragaszkodik mereven az előzetes tervekhez, hanem beszéd közben egyes részeit szó szerint, más részeit pedig átalakítva mondja el. Részeket elhagy, más részeket kiegészít, vagy teljesen új részekkel pótol. Tehát a szövegalkotás (gondolkodás) és szövegmondás (beszéd) néha a spontán beszédhez hasonlóan, szimultán módon zajlik le, máshol pedig a szövegalkotás megelőzi a szövegmondást. A parlamenti beszédeket, vagy annak a vázlatát a politikusok rendszerint előre leírják. Előadás közben erre a leiratra támaszkodnak, de attól sokszor eltérő, kibővített formában mondják el a beszédet. Ezért a HMM-TTS szempontjából vizsgálva a politikai beszédeket a félspontán kategóriába soroltam.

A félspontán beszéd legfontosabb jellemzői a beszélőadaptáció szempontjából a következőek:

1. A *hangszínezet* (spektrális összetevők) hol a spontán, hol pedig a tervezett beszédre jellemző. Ezek a részek nem különíthetőek el egyértelműen, mert az átmenet sokszor folytonos. Spontán beszéd során az egyéni hangszínezethez viszonyítva nagy eltérések lehetnek az érzelmek, attitűdök és szituációk váltakozásának megfelelően. Tervezett beszéd esetén rendszerint átlagos hangszínezet valósul meg, melyet azonban a műfaji sajátosságok jelentős mértékben befolyásolhatnak (pl. mesemondás).
2. A *hangerő* igen változatos képet mutat. A hangsúlyozás néhol váratlan, „illogikus”, máshol pedig követi a nyelvtani szabályokat.
3. A *hangmagasság* ingadozása folyamatosan változik a spontán és a tervezett részek között. A spontán részeknél a hangmagasság az érzelmek és szituációk alapján igen változatos. A tervezett részeknél nagyjából azonos hangmagasságban mozog a beszéd, de egy szekunddal vagy terccel magasabban, mint a spontán részeknél.
4. A *beszédtempó* és *szünetek* néhol nagyon változatos formát mutatnak, míg máshol szabályosak, olykor szinte metronómszerűek.
5. Változatos légzésmód, hangadás és hangindítás.
6. A spontán beszédnél igényesebb, de a tervezett beszédhez képest néhol elnagyolt *artikuláció*.
7. A *hangkapcsolódást* a nyelvi rétegnek, a beszélő nyelvi szintjének és az adott szituációnak megfelelően módosult nyelvi szabályok határozzák meg.

6.3. Kényszerített illesztés a felügyelet nélküli hanghatár-jelöléshez

Az automatikus hanghatár-jelölés fontos része a felügyelet nélküli beszélőadaptációnak. Ahogyan korábban szó volt róla, *kényszerített illesztéssel* a beszéd hullámformája alapján a

fonetikus átirat egyes elemeihez (hangjaihoz) időzítési paramétereket automatikus módon lehet rendelni [36].

A hanghatár-jelölés pontossága jelentősen függ a hangfelvétel minőségétől, a fonetikus átirat pontosságától és a kényszerített illesztés eljárásától. A kényszerített illesztést a beszédszintézis területén a beszédkorpusz címkézése során használják. Felépítésükből adódóan ez elsősorban az elemkiválasztásos és a HMM-TTS rendszerekben fontos.

A hanghatárok meghatározása spontán, félspontán, de még tervezett beszéd esetén is szubjektív feladat, legtöbb esetben nem lehet *ms* pontossággal manuális úton sem megállapítani, hogy hol ér véget egy beszédhang, és hol kezdődik a következő. Számos forrás adatai alapján maximum *20 ms*-os eltérést mutat a beszédhang határok mintegy 90%-a különböző beszédkorpuszokon, különböző beszédtechnológiai szakemberek által végzett címkézés esetében [77,78,79].

Több elvi megközelítés is létezik a kényszerített illesztésre, napjainkban a leginkább a rejtett Markov-modell alapú rendszerek terjedtek el. A HMM alapú beszédfelismerőt ebben az esetben illesztésre használjuk, hiszen tudjuk, hogy mi hangzott el. A betanított rejtett Markov-modellek maximum likelihood becslés alapján megadják a legpontosabbnak vélt beszédhang határokat.

Az eljárás alapját képző Viterbi algoritmus egy adott (nem végső) időpillanatához számos részleges felismerési hipotézis tartozik. A felismerési hipotézisek azért részlegesek, mert az algoritmus nem ért a jellemzővektor sorozat végéhez. Ezen részleges felismerési hipotézisekből igen számításigényes (és általában értelmetlen) volna mindet megtartani. Ezért vagy az adott időpillanatban a legjobbnak ítélt felismerési hipotézishez képest adott mértékben lemaradó felismerési hipotéziseket tartjuk meg a likelihood érték alapján, vagy pedig adott számú legjobb hipotézist. A gyakorlatban a kettő kombinációját is szokták használni. Keresési mélységnek, vagy beam szélességnek nevezzük a megtartott felismerési hipotézisek számát/mértékét.

A HMM alapú megoldáson túl kényszerített illesztés céljából használható több frekvenciasávos eljárás, dupla diád kereséses eljárás [80], szabály alapú eljárás [78] és neurális hálózat alapú hanghatár jelölés is [81,82].

Kutatómunkám során alkalmazott beszédfelismerő és a kényszerített illesztő azonos akusztikus modellt használt. A beszédfelismerő szó szintű kimenetet adott, ezért volt szükséges a kényszerített illesztést külön lépésben futtatnom. A nyelvi modellt a kényszerített illesztő esetében maga a fonetikus átirata adta.

6.4. Felügyelt beszélőadaptáció félspontán beszédkorpusszal

A 6.2. fejezetben bemutatott félspontán beszéd tulajdonságai jelentősen eltérnek a tervezett beszéd jellemzőitől. A felügyelet nélküli beszélőadaptációval foglalkozó kutatás kezdetén fontosnak találtam megállapítani, hogy lehetséges-e félspontán adaptációs beszédkorpuszok alapján a tervezett beszéddel való adaptációhoz hasonló minőségű HMM-TTS rendszert készíteni. Ehhez először adaptációs beszédkorpuszt kellett előállítani a rendelkezésre álló nyers hanganyagból.

6.4.1. Az adaptációs beszédkorpusz előállítása

Az adaptációs beszédkorpusz előállítása a nyers hanganyagból négy lépésben történt: beszélő azonosítása, hanganyag szegmentálása, fonetikus átirat elkészítése, adaptációs

beszédkorpusz szelekciója. A *beszélő azonosítása* jelen kutatás során a forrás hangfelvételek alapján adott volt: a hanganyag minden egyes beszélőtől halmazokra bontva állt rendelkezésre. A továbbiakban ezért a beszélőazonosításra már külön nem térek ki.

6.4.1.1. Hanganyag szegmentálása

A *hanganyag szegmentálása* jelen esetben a mondathatárok minél pontosabb megadását jelenti. A kutatás ebben a fázisában a hanganyag szegmentálását kézi úton végeztem, később automatikus módszerekre tértem át.

A félspontán beszéd magában hordozza a tervezett és spontán beszéd jellegzetességeit. Tervezett beszéd esetén a mondathatárok hanganyag alapján történő percepciós jelölése egyöntetűen 90% feletti pontossággal határozható meg [83]. Spontán beszéd esetén azonban már maga a mondathatár fogalma sem egyértelmű. Például a spontán beszédet felépítő szegmens lehet egy mondatnyi egység, de lehet egy hosszabb, összefüggő szövegrész is, az un. „bekezdés" [84]. A mondatok és bekezdések szintaktikailag és szemantikailag különálló egységeket alkotnak és a szupraszegmentális szerkezet jellegzetességei által lehet őket meghatározni. Ilyen jellegzetesség a szünet, az alapfrekvencia változás, az intenzitás csökkenés és a szünet előtti szó lassabb artikulációja [85]. Hird és Kirsner az intenzitás csökkenését és a frekvencia változását figyelte meg a mondathatárokon [86]. A bekezdés – hosszúságából fakadóan – nem előnyös a HMM-TTS beszélőadaptációjához. Célom a hosszabb szövegrészekben (bekezdésekben) a kisebb közlések azonosítása volt. A szakirodalom ezen egységeket *virtuális mondatokként* jelöli [87]. A tanulmány alapján mintegy *700 ms*-os szünethossz felett 80-100% pontossággal lehetséges a virtuális mondathatárokat megállapítani. Ezért a kézi szegmentálás során én is ezt a határértéket vettem alapul.

6.4.1.2. Fonetikus átirat elkészítése

A felügyelt esetben kézileg, a BME-TMIT Beszédtechnológiai Laboratórium egyik munkatársával konszenzusos módon állítottam elő a fonetikus átiratot a BME-TMIT Beszédtechnológiai Laboratórium munkatársai segítségével. A kutatás során ezen fonetikus átiratot tekintettem referenciának, ehhez mérten relatív módon tudtam a felügyelet nélküli rendszerek minőségét meghatározni.

6.4.1.3. Adaptációs beszédkorpusz szelekciója

A félspontán beszéd bizonyos elemei a beszélőadaptáció szempontjából egyértelműen előnytelenek. Hogy az előnytelen minták ne szerepeljenek az adaptációs beszédkorpuszban, a szegmentálás után megvizsgáltam az eredményt, és az előnytelennek vélt mondatokat eltávolítottam. Ezek meghatározásához a szegmentálás során kapott hangmintákat az alábbiak szerint sorolom osztályokba:

1. Hosszú, összetett mondatok: a spontán és félspontán beszéd jellemzője, hogy sok esetben egy-egy gondolatot igen hosszú összetett mondatban fogalmaznak meg a beszélők. Ahol tervezett beszéd esetén vége lenne a mondatnak és új kezdődne, ott a hanganyagban a hangmagasság közel azonos marad, és szünet után folytatódik a mondanivaló. Ezen mondatok a HMM-TTS szempontjából teljesen ismeretlen állapotokat tartalmaznak. A hang, szótag, szó, mondatrész, mondat pozíciójára és méretére vonatkozó környezetfüggő címkék jelentős eltérést mutatnak a tervezett beszéd esetén előforduló elemektől. A jelen kutatásnak nem célja ilyen típusú mondatok modellezése, ezért fontos ezen mondatok

eltávolítása az adaptációs beszédkorpuszból. A szelekció során minden 50 szónál hosszabb mondatot eltávolítottam.
Példa (a félspontán beszéd néhol nyelvtanilag vagy tartalmilag helytelen): „*Nagyon sokszor kérdezik, nagyon sokszor elhangzik, hogy mit kell ezért az országnak fizetni, mi az a támogatási összeg, amit, amit mi kifizetünk, azt kell hogy mondjam, hogy az egyedi, kormány döntéssel nyújtott befektetési támogatások egyértelműen ma már egy jól működő rendszert az európai bizottság által is elismert rendszert alkotnak, nem a névvel, talán egy picit ellentétben nem egyedi döntésekről, hanem egy jól végiggondolt stratégiáról van szó, az egyetlen ország vagyunk a régióban az új tagállamok, tagállamok között akinek nincs visszavont vagy vitatott támogatása, valamennyi támogatásunk megfelel az európai unió előírásainak, nagyon örülünk ennek és ma már azt is elmondhatom itt a slideon látszik, először látják önök ezt az elemzést, először hozzuk nyilvánosságra, hogy ennek a rendszernek az államháztartás szempontjából a megtérülése kettő ezer hét második fele óta pozitív, tehát többet termel ez a rendszer, többet termelnek az egyedi kormánydöntéses támogatások, mint amennyibe kerülnek, azért vagyunk, óvatosan ezzel az adattal, hiszen nagyon sok vállalatnál még nem tudjuk pontosan, hogy a szerződésben vállal kötelezettségek teljesítése az mekkora államháztartási bevételt fog generálni, nagyon csúnyán fogalmazva nem értünk a monitoring időszakok végére, tehát sok esetben becsült adatokkal dolgozunk, de már vannak bár kisebb számban tényadataink arra vonatkozóan, hogy milyen államháztartási bevételt adóbevételt járulékbevételt generálnak ezek a megvalósult beruházások." (forrás: www.parlament.hu)*

2. Rövid, sebes mondatok: a hosszú, összetett mondatok mintegy lezárásaként rövid, sebes ritmusú, néhány szavas mondatok. Mivel az előző kategóriába eső mondatok modellezése nem célom, ezért az azokat követő rövid, sebes mondatokat is eltávolítottam. A szelekció során ennek megfelelően azon mondatokat vetem el, amelyek eleget tesznek az alábbi egyenlőtlenségnek:

$$\frac{{n_{i-1}}^2}{t_{i-1}} > \frac{{n_i}^2}{t_i} d \tag{37}$$

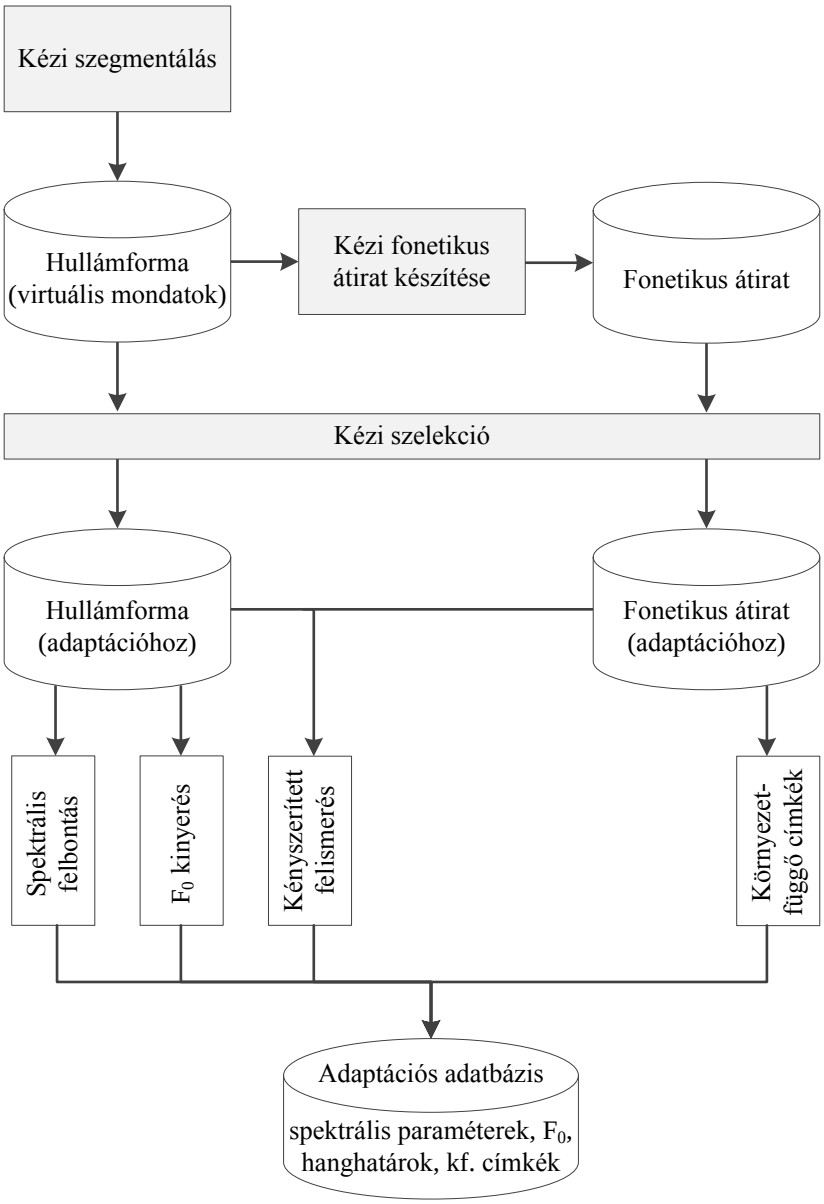
ahol n_i jelöli az *i*-edik mondat fonémáinak a számát, és t_i jelöli az *i*-edik mondat időtartamát. A hányados megadja a felolvasás gyorsaságának és a mondat hosszának az arányát, a *d* arányszámmal pedig az egymást követő mondatok hosszának az arányát lehet szabályozni. A rövid sebes mondatok eldobásához empirikus úton meghatározott *d=20*-at használtam.

3. Félbehagyott, majd újra elkezdett mondatok: a beszélő a mondatot értelmileg nem indokolt helyen fejezi be, majd a gondolatot egy újabb mondatban zárja. A HMM-TTS robosztusságából adódóan ezen mondatok várhatóan nem okoznak problémát, hiszen elsősorban a tartalomra vonatkoznak, a modellben a spektrális, alapfrekvencia és időzítési paramétereket nem befolyásolják negatív módon.
Példa: „*Ezen társaságok esetében sor kerül ez a program elindítására és az értékesítés egy második szakaszában került sor a Szerencsejáték ZRT, a Magyar Posta ZRT, a regionális vízművek és a magyar villamos művek ZRT. Bevonására a programba." (forrás: www.parlament.hu)*

4. A negyedik csoportot a korábbi három csoportba nem sorolható mondatok jelentik. Természetesen Wacha alapján lehetséges volna tovább osztályozni a félspontán beszédet [76], de a HMM-TTS beszélőadaptációja szempontjából előzetesen elegendőnek tartottam a fent ismertetett szelekciót.

6.4.1.4. Az eljárás összefoglalása

A felügyelt, félspontán beszéddel való HMM-TTS adaptáció lépéseit a 19. ábra mutatja. Az ábrán már a beszélőazonosítás utáni lépéseket mutatom be. Az eljárás a következőkben leírtak szerint működik. Először a bemeneti hullámformát a korábban ismertetett módon szegmentáltam, majd elkészítettem az így kapott mondatoknak a fonetikus átiratát. A szelekciót a fonetikus átirat és a hullámforma segítségével végzem el, majd az így kapott adaptációs beszédkorpuszt előkészítettem a HMM-TTS számára. Ennek során a hullámformából kiszámoltam a spektrális és gerjesztési paramétereket, a fonetikus átirat alapján pedig meghatároztam a környezetfüggő címkéket, illetve kényszerített illesztéssel a hanghatárokat. Az ábrán szürke dobozokkal jelöltem azon részeket, melyeket a felügyelet nélküli adaptáció során módosítani fogok. A beszélőadaptációt az 5.4. fejezetben bemutatott módszer segítségével a létrejött adaptációs beszédkorpusszal végeztem el (12. táblázat). A folyamat eredményeként előállt négy félspontán beszéd alapján adaptált HMM-TTS.



19. ábra. A felügyelt beszélőadaptáció lépései félspontán beszéddel.

6.4.2. A létrehozott adaptációs beszédkorpuszok

Az átlaghangot az 5.4. fejezetben ismertetett beszédkorpuszokkal tanítottam. Az adaptációs beszédkorpuszok félspontán (parlamenti) beszédet tartalmaztak. A hanganyag a *http://www.parlament.hu* oldalon található videó felvételekből lett rögzítve. A hanganyagot beszélőnként külön válogattam[9], majd az előző fejezetben ismertetett módon szegmentáltam. Ezután a BME-TMIT Beszédtechnológiai Laboratórium munkatársaival konszenzusos módon elkészítettük a hanganyag kézi fonetikus átiratát (ezt „hiba nélkülinek" tekintettem), végül a HMM-TTS számára előnyös részeket kiválogattam (szelekció). Az így létrejött adatbázisokból mintegy 10 perc hosszúságú adaptációs beszédkorpuszt választottam ki véletlenszerű módon. Az így létrejött adaptációs beszédkorpuszok tulajdonságait a 12. táblázat mutatja be. A táblázatban megadtam a továbbiakban használt jelölést, mely tartalmazza a fontosabb jellemzőket (*FFx:* x-edik férfi beszélő, *FÜ*: felügyelt adaptáció; *RND*: véletlenszerű szelekció).

12. táblázat. Félspontán adaptációs beszédkorpuszok felügyelt beszélőadaptációhoz.

Jelölés	Beszélő	Módszer	Szelekció	Időtartam	PER[10]	WER[11]
FF6-FÜ-RND	Férfi 6.	Felügyelt	Véletlenszerű	11.4 perc	„hiba nélkül"	
FF7-FÜ-RND	Férfi 7.	Felügyelt	Véletlenszerű	9.6 perc	„hiba nélkül"	
FF8-FÜ-RND	Férfi 8.	Felügyelt	Véletlenszerű	10.2 perc	„hiba nélkül"	
FF9-FÜ-RND	Férfi 9.	Felügyelt	Véletlenszerű	9.7 perc	„hiba nélkül"	

6.4.3. Számszerű kiértékelés

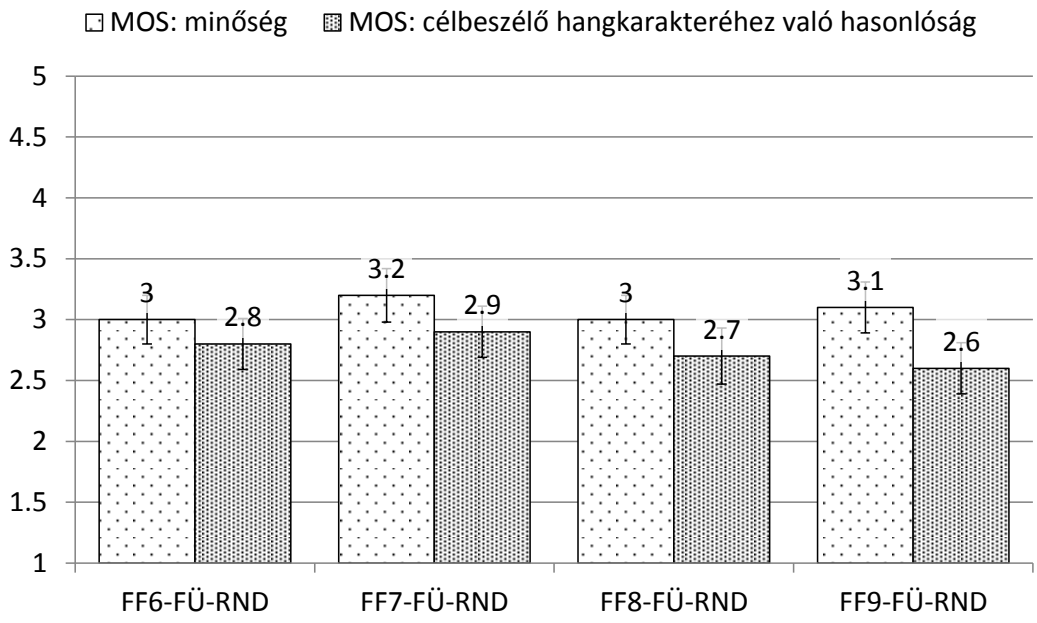
Annak érdekében, hogy az így létrejött, félspontán beszéd alapján adaptált HMM-TTS minőségét megállapítsam, meghallgatásos teszteket végeztem. A kísérlet során a gépi beszéd és az eredeti beszélő közötti hasonlóságot (természetesség), illetve a gépi beszéd minőségét mértem. A meghallgatásos tesztben a 12. táblázatban bemutatott rendszerek vettek részt.

A meghallgatásos teszt MOS jellegű volt. Az első részben minden rendszerből 4-4 (összesen 16) mintát játszottam le, és minőségük alapján a tesztalanyok ezt osztályozták. A teszt második felében a tesztalanyok az adaptált gépi beszédet a célbeszélőtől származó természetes bemondásokkal hasonlították össze. Feladatuk annak meghatározása volt, hogy a gépi hangminták milyen mértékben adják vissza a természetes beszélő hangkarakterét. Páronként a bemondások szövege azonos volt. Minden tesztalanynak az egyes rendszerekből véletlenszerűen választva, egyenletes eloszlást követve 5-5 pár (összesen 20) mintát játszottam le.

A meghallgatásos teszt internet alapú volt. Összesen 25-en végezték el, 19 férfi és 6 nő, közülük tízen beszédtechnológiai szakértők voltak. Az átlagéletkor 35 év volt, a legfiatalabb tesztalany 21, a legidősebb 67 éves volt.

Az eredményeket a 20. ábra szemlélteti. Az ábra alapján megállapítható, hogy mindegyik félspontán beszéddel adaptált rendszer közel azonos minőségű, és a természetes beszélő hangkarakterét hasonló módon adták vissza.

[9] Az automatikus beszélőfelismerő nem volt a kutatás része.
[10] PER: Phone Error Rate, fonéma hibaarány.
[11] WER: Word Error Rate, szó hibaarány.



20. ábra. Felügyelt beszélőadaptációval készített félspontán HMM-TTS minőségének vizsgálata.

6.4.4. Konklúzió

A felügyelet nélküli adaptációt vizsgáló kutatásom kezdetén félspontán típusú bemondásokat választottam beszédkorpusznak, hiszen gyakorlati megvalósítás esetén várhatóan nem tervezett beszéd alapú adaptációs korpusz áll majd rendelkezésre[12]. A kutatás első lépéseként azt vizsgáltam, hogy felügyelt módon lehetséges-e a beszélőadaptáció a korábbi tervezett beszéddel szemben félspontán beszéddel. A meghallgatásos teszt eredményeit figyelembe véve megállapítható, hogy a félspontán hangfelvételek alapján történt beszélőadaptáció a javasolt módszerrel sikeres volt. Továbbá, bár két különböző meghallgatásos teszt konkrét értékeit nem minden esetben lehet pontosan összehasonlítani, mégis a 15. és 20. ábrákat összevetve megfigyelhető, hogy a tervezett beszéd alapú beszélőadaptáció arányaiban azonos minőséget képvisel a jelen pontban ismertetett félspontán adaptációs beszédkorpusz alapján készült HMM-TTS rendszerrel. Fontos megjegyeznem, hogy a félspontán hanganyag HMM-TTS adaptációja szempontjából előnytelen részeit különválogattam, ami megkötést jelent a tervezett beszédhez képest (lásd 6.4.1.3. fejezet).

Az eredmények alapján félspontán beszédkorpusszal lehetséges a felügyelt beszélőadaptáció HMM-TTS rendszerekben. Következő lépésként eljárást dolgoztam ki a HMM-TTS rendszer emberi beavatkozás nélküli beszélőadaptációjára.

[12] Ahogyan korábban már szó volt róla, ha tervezett beszéd lenne az adaptációs beszédkorpusz, az azt feltételezné, hogy rendelkezésre állnak a bemondások szöveges átiratai, így a felügyelet nélküli beszélőadaptáció értelmét veszti.

6.5. Felügyelet nélküli beszélőadaptáció félspontán beszédkorpusszal

A kutatás folytatásaként a 6.4. fejezetben bemutatott eljárást automatizáltam. Első lépésként az előző fejezetben bemutatott adaptációs beszédkorpuszok felügyelet nélküli változatait készítettem el. Az így létrejött beszédkorpuszok a felügyelt, referenciának tekintett esetekkel összehasonlítva 10-40% közötti fonémahiba-aránnyal rendelkeztek. Ezután kiválasztottam egy beszélőt, akinél rosszabb felismerési pontosságú, akár még 90% fonémahiba-arány körüli korpuszokat is vizsgáltam.

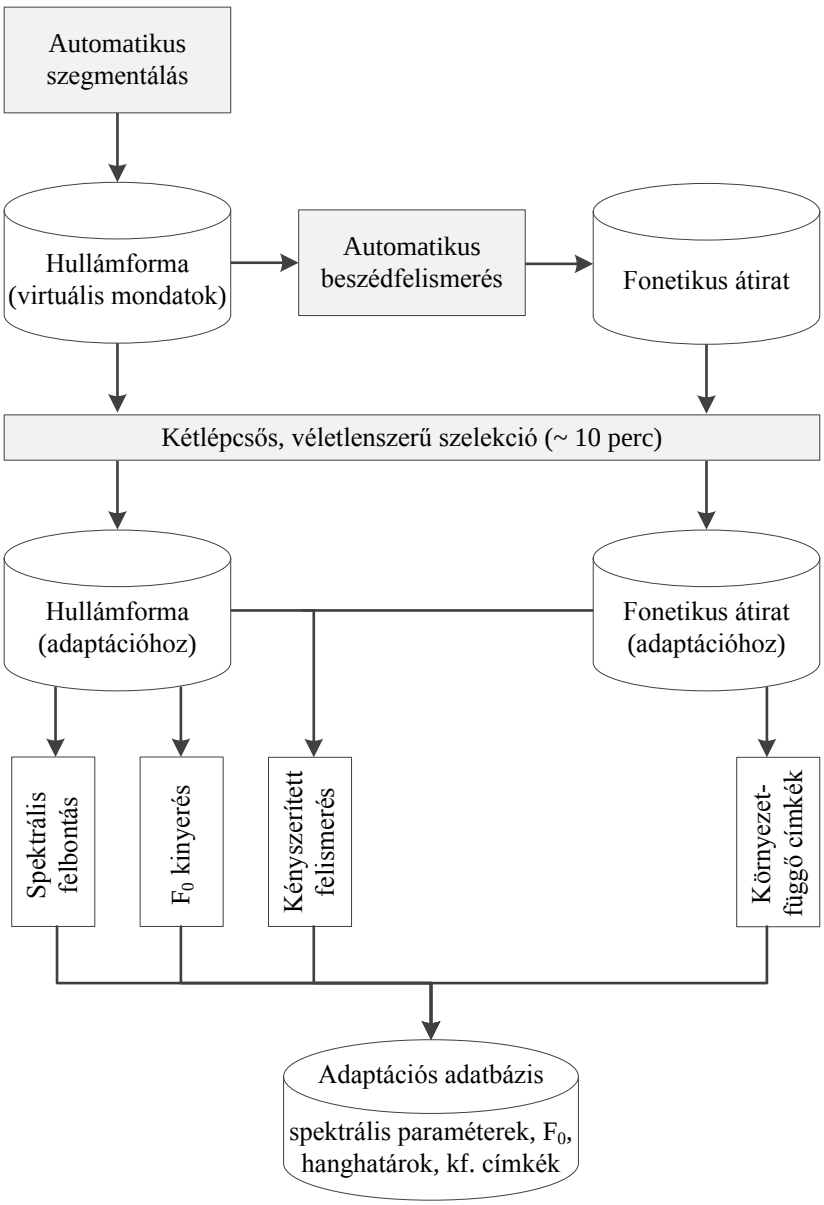
6.5.1. Az adaptációs beszédkorpusz előállítása

Az adaptációs beszédkorpusz előállítása a 6.4.1.-ben bemutatott lépések alapján történt (beszélőazonosítás, szegmentálás, fonetikus átirat elkészítése és szelekció). A felügyelet nélküli működés céljából több lépést módosítottam, illetve automatizáltam. *Automatikus szegmentálásra* a korábbi kutatások során számos megoldás született. Az automatikus beszédfelismerés eredményei alapján a szünetek, az alapfrekvencia, a szótag és szóintenzitás és időtartam befolyásolhatják a szegmentálást [83]. Vannak olyan megoldások, melyek csupán a szöveges átirat (felismert beszéd) alapján határozzák meg a mondathatárokat [88,89]. Egy másik megközelítés a szöveges átirat mellett a prozódiai információkat is figyelembe veszi, és a rejtett Markov-modellek segítségével adja meg a mondathatárokat [90]. Gotoh és Renals cikke alapján már önmagában a szünethossz vizsgálatával is jó eredményt lehet elérni, mely nyelvi modell használatával tovább javítható [91]. A nemzetközi eredmények alapján úgy döntöttem, hogy a szünetek meghatározására félspontán beszéd esetén a beszédfelismerő szünet modelljét használom, és a szünetek hosszát vizsgálva szegmentálom a hanganyagot. A 6.4.1.1. fejezet szerint mondathatárnak a hullámformában azon részeket jelöltem, ahol a beszédfelismerő szünetet jelzett, és a szünet hossza hosszabb, mint *700 ms*. Ez a fajta mondat- és tagmondathatár meghatározás várhatóan nem okoz problémát, mert a HMM-TTS átlaghangja már megtanulta a nyelvre jellemző tagmondat- és mondathatárokon jelentkező tulajdonságokat (pl. F_0, hangerő változás), és ezt a HMM-TTS robosztusságából fakadóan az adaptációs beszédkorpuszban megjelenő esetleges hibás mondathatárok nem rontják le. A HMM-TTS rendszerben a mondatok közötti szünetek modellezése céljából a szegmentálás során a virtuális mondatok hanganyagához hozzávettem az előttük és utánuk szereplő szünetek felét.

A szegmentálás után a hanganyagon magyar nyelvű, nagy szótáras beszédfelismerőt futtattam le [42]. A triád alapú akusztikus modell 500 beszélőtől származó mintegy 5 óra hanganyaggal, a morféma alapú trigram nyelvi modell 1.2 millió szóval lett tanítva. A szavak nagyobb része internetről gyűjtött híreket, másik része politikai híreket tartalmazott. A felismerés pontossága növelése érdekében azonos témájú adatbázisból tanított beszédfelismerőt választottam, mint amivel a HMM-TTS beszédhangját adaptáltam (politikai szövegek). Következő lépésként a beszédfelismerő ortografikus kimenetét fonetikus átirattá alakítottam át. Ezt a korábbiakhoz hasonlóan a ProfiVox fonetikus átíróval tettem meg. A kutatás jelen szakaszában a 6.4.1.3. fejezetben bemutatott szelekciós eljárást az automatikus fonetikus átirat és a hullámformák figyelembe vétele alapján automatizáltam. Ezt használtam *szelekciós eljárásként*. Az eljárás eredményeképp a kísérletben részt vevő minden beszélőtől mintegy 60 percnyi beszédkorpusz állt elő. Az így keletkezett virtuális mondatokból véletlenszerűen mintegy 10 percnyi adaptációs beszédkorpuszt választottam ki.

Az fentebb ismertetett elemekből álló felügyelet nélküli adaptációs eljárást a 21. ábra szemlélteti. A felügyelt esethez képest szürke háttérrel jelöltem az eljárás azon részeit,

melyeket automatizáltam: a szegmentálást, a fonetikus átirat elkészítését, a szelekciót. Az eljárás során a bemeneti hullámformát automatikus úton szegmentáltam. Ezt következőnek az így létrejött virtuális mondatok szövegét automatikus beszédfelismerővel határoztam meg, majd ezt a ProfiVox fonetikus átírójának a segítségével fonetikus átirattá alakítottam. Az így létrejött fonetikus átiratokat és a hozzájuk tartozó hullámformákat két lépcsőben szelektáltam: először az adaptáció szempontjából előnytelen elemeket dobtam el, majd az így maradt hanganyagból véletlenszerű módon mintegy 10 percnyi adaptációs beszédkorpuszt válogattam össze. Ezután az adaptációs beszédkorpuszt a HMM-TTS beszélőadaptációjára készítettem elő: kinyertem a spektrális és gerjesztési paramétereket, kényszerített illesztéssel meghatároztam a fonéma határokat és a fonetikus átiratból elkészítettem a környezetfüggő címkéket. Utolsó lépésként a beszélőadaptációt végzem el, aminek eredményeként felügyelet nélkül jött létre félspontán HMM-TTS.



21. ábra. A felügyelet nélküli beszélőadaptáció lépései félspontán beszéddel.

6.5.2. Adaptációs beszédkorpusz

Az átlaghangot ebben az esetben is az 5.4. fejezetben ismertetett beszédkorpuszokkal tanítottam. A 12. táblázatban bemutatott, kézi címkézéssel készült beszédkorpuszokat tekintettem referenciának, majd az előző fejezetben ismertetett módszerekkel ezek felügyelet nélküli változatait készítettem el. A beszédfelismerésben a BME-TMIT Beszédtechnológiai Laboratórium munkatársai voltak segítségemre. Az így létrejött adaptációs beszédkorpuszokat a 13. táblázatban ismertetem. A jelölés nagy része a korábbiakkal azonos, az *FN* pedig a felügyelet nélküli adaptációra utal.

A beszédfelismerő kimenete a jó minőségű, témaspecifikus tartalomnak köszönhetően olyan jó volt, hogy ezt a minőséget mesterséges úton le kellett rontanom ahhoz, hogy a fonéma tévesztések hatásait rosszabb felismerési pontosság esetén is vizsgálni tudjam. Ezért 0-gramos nyelvi modellekkel[13], különböző szintű fehér zajjal terhelt hanganyagon történt a további beszédfelismerés *Férfi 8.* beszélő esetén. A gyakorlatban sokszor nem áll rendelkezésre jó minőségű hanganyag és a szabad témájú hanganyagok felismerési pontossága is változó, ezért fontos volt, hogy a fenti módon, igen tág megkötésekkel dolgoztam. Ezzel a módszerrel a 14. táblázatban található adatbázisokat hoztam létre. A táblázatban szereplő adatbázisok mindegyikét felügyelet nélkül készítettem (*FN*-el jelölöm) és véletlenszerűen válogattam ki a mintegy 10 percnyi részt (*RND*-vel jelölöm). A jelölésben a *0G* a 0-gram nyelvi modellre utal, a *ZAJ* és *ZAJ2* pedig különböző szintű fehér zajok esetét jelöli: a maximális kivezérlés mondatonként 0 dB-re lett normalizálva, majd a teljes kivezérléshez képest -50 dB (ZAJ) és -25 dB (ZAJ2) fehér zajt kevertem a jelhez.

13. táblázat. Félspontán adaptációs beszédkorpuszok felügyelet nélküli beszélőadaptációhoz.

Jelölés	Beszélő	Módszer	Szelekció	Időtartam	PER	WER
FF6-FN-RND	Férfi 6.	Felügyelet nélküli	Véletlenszerű	11.4 perc	42%	87%
FF7-FN-RND	Férfi 7.	Felügyelet nélküli	Véletlenszerű	9.6 perc	21%	74%
FF8-FN-RND	Férfi 8.	Felügyelet nélküli	Véletlenszerű	10.2 perc	17%	57%
FF9-FN-RND	Férfi 9.	Felügyelet nélküli	Véletlenszerű	9.7 perc	10%	44%

14. táblázat. Egy beszélőtől származó, rossz felismerési eredmények szimulálásával készült félspontán adaptációs beszédkorpuszok felügyelet nélküli beszélőadaptációhoz.

Jelölés	Beszélő	Nyelvi modell	Zaj	Időtartam	PER	WER
FF8-FN-0G-RND	Férfi 8.	0-gram	-	9.5 perc	55%	100%
FF8-FN-0G-RND-ZAJ	Férfi 8.	0-gram	-50 dB	8.9 perc	70%	100%
FF8-FN-0G-RND-ZAJ2	Férfi 8.	0-gram	-25 dB	9.7 perc	89%	100%

6.5.3. Számszerű kiértékelés

6.5.3.1. 50% PER alatti félspontán adaptációs beszédkorpusz

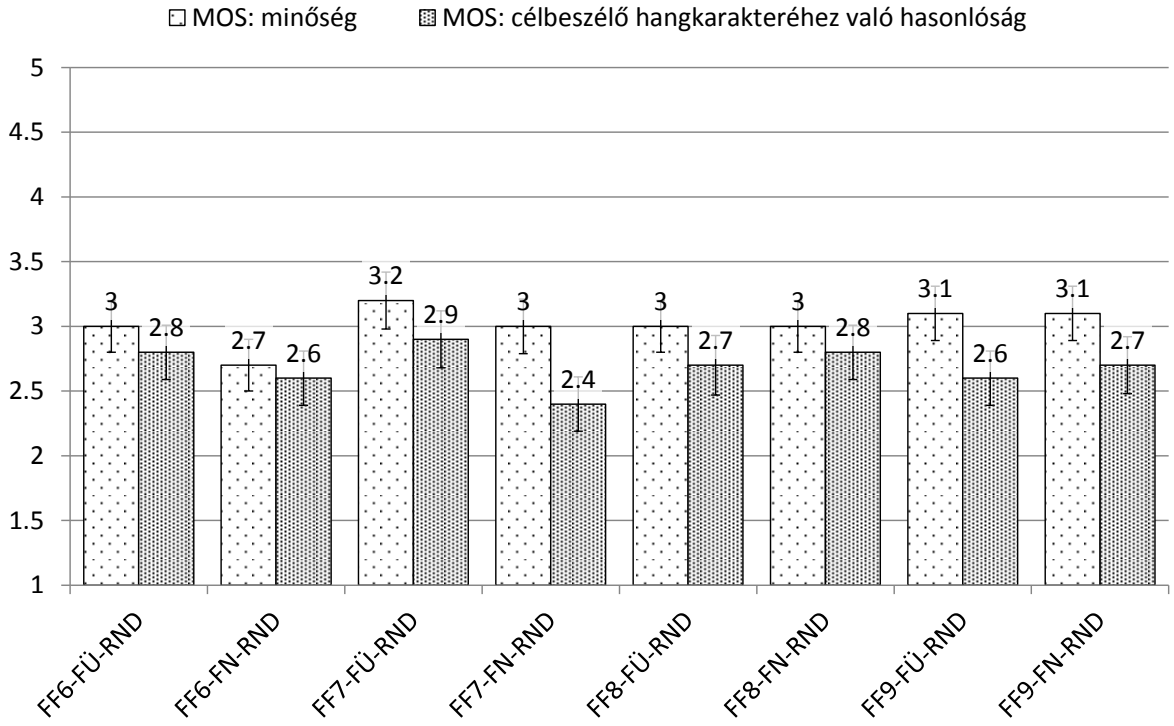
Az adaptációt a 12. és 13. táblázatban szereplő nyolc adaptációs beszédkorpusszal végeztem el. A 12. táblázatban szereplő négy korpusz felügyelt módon, a 13. táblázatban feltüntetettek pedig felügyelet nélküli módon készültek. Az így előállt félspontán HMM-TTS rendszerek minőségét meghallgatásos tesztekkel vizsgáltam. Elsődleges célom az volt, hogy

[13] 0-gram esetén minden morféma egyszer, azonos valószínűséggel szerepel a nyelvi modellben.

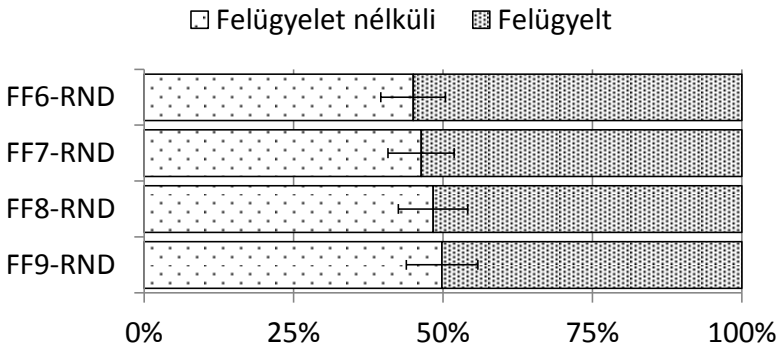
megállapítsam, hogy 50%-os fonémahiba-arány alatt hogyan alakul a felügyelet nélküli beszélőadaptáció minősége.

A meghallgatásos teszt három részből állt: egy a minőségre és egy a természetességre vonatkozó MOS tesztből, és egy CMOS pár összehasonlításból. A meghallgatásos teszt *első részében* a tesztalanyoknak 32 gépi beszédminta minőségét kellett osztályozniuk. Minden rendszerből 4-4 mintát játszottam le. A *második részben* a tesztalanyoknak azt kellett az előző részhez hasonló módon 1-től 5-ig terjedő skálán meghatározniuk, hogy mennyire emlékezteti őket a lejátszott minta az eredeti beszélő hangjára. Ehhez minden minta előtt egy referencia mintát, az eredeti beszélőtől származó természetes bemondást játszottam le. A referencia minta és a gépi bemondás szövege minden esetben azonos volt. Minden rendszerből 5 mintát játszottam le, így egy tesztalany összesen 40 mintát értékelt. Az *harmadik részben* a tesztalanyoknak 5 elemű skálán kellett megállapítaniuk, hogy a lejátszott két minta hogyan viszonyul egymáshoz. Az 1-es azt jelentette, hogy az első minta sokkal jobb, mint a második minta, a 2-es, hogy kicsit jobb, a 3-as, hogy azonos, a 4-es, hogy a második minta jobb, az 5-ös pedig, hogy a második mintát sokkal jobbnak érzi a tesztalany. A minták szövege minden esetben azonos volt. A teszt első felében azonos beszélőtől, de különböző adaptációs módszerrel (felügyelt ↔ felügyelet nélküli) készült rendszerből összesen 8 mintapár vett részt.

A teszt minden részében a minták egy nagyobb halmazból, ál-véletlen módon lettek kiválasztva. A kiválasztás egyenletes eloszlást követett: minden rendszerből azonos mennyiségű mintát választottam ki. A meghallgatásos teszt internet alapú volt. Összesen 25-en végezték el, 19 férfi és 6 nő. Az átlagéletkor 35 év volt, a legfiatalabb tesztalany 21, a legidősebb 67 éves volt. 10 tesztalany beszédtechnológiai szakértő volt.



22. ábra. Felügyelet-nélküli félspontán HMM-TTS minőségének és természetességének vizsgálata 50% PER alatti adaptációs beszédkorpuszok esetén MOS pár összehasonlítással.



23. ábra. Felügyelet-nélküli félspontán HMM-TTS minőségének vizsgálata 50% PER alatti adaptációs beszédkorpuszok esetén CMOS pár összehasonlítással.

A meghallgatásos teszt első és második részének az eredményeit a 22. ábra, a harmadik részének az eredményeit a 23. ábra mutatja. A 22. ábra külön, 2-2 oszloppal jelöli a minőségre és a célbeszélő hangkarakteréhez való hasonlóság MOS értékeit felügyelt (*FÜ*) és felügyelet nélküli (FN) beszélőadaptációk esetén. A magasabb érték jelenti a jobb minőséget és természetességet. Az eredmények alapján megállapítható, hogy sem minőségben, sem természetességben sincs szignifikáns különbség a felügyelt és a felügyelet nélküli esetek között. A 23. ábra is ezt mutatja: a felügyelt és felügyelet nélküli adaptációból származó minták között a tesztalanyok mind a négy beszélő esetén nagymértékű hasonlóságot érzékeltek.

Az eredmények alapján a következő következtetést vontam le: ahogyan nő a PER, úgy romlik az adaptáció minősége, azonban még 42%-os PER esetén sincs az ANOVA analízis szerint szignifikáns különbség a felügyelt és felügyelet nélküli eset között. Ezért kiválasztottam az egyik beszélőt (FF8), és magasabb fonémahiba-arányokkal további meghallgatásos teszteket végeztem. Célom az volt, hogy megállapítsam, az általam kidolgozott eljárás szignifikáns minőségromlás nélkül milyen hibahatárig használható.

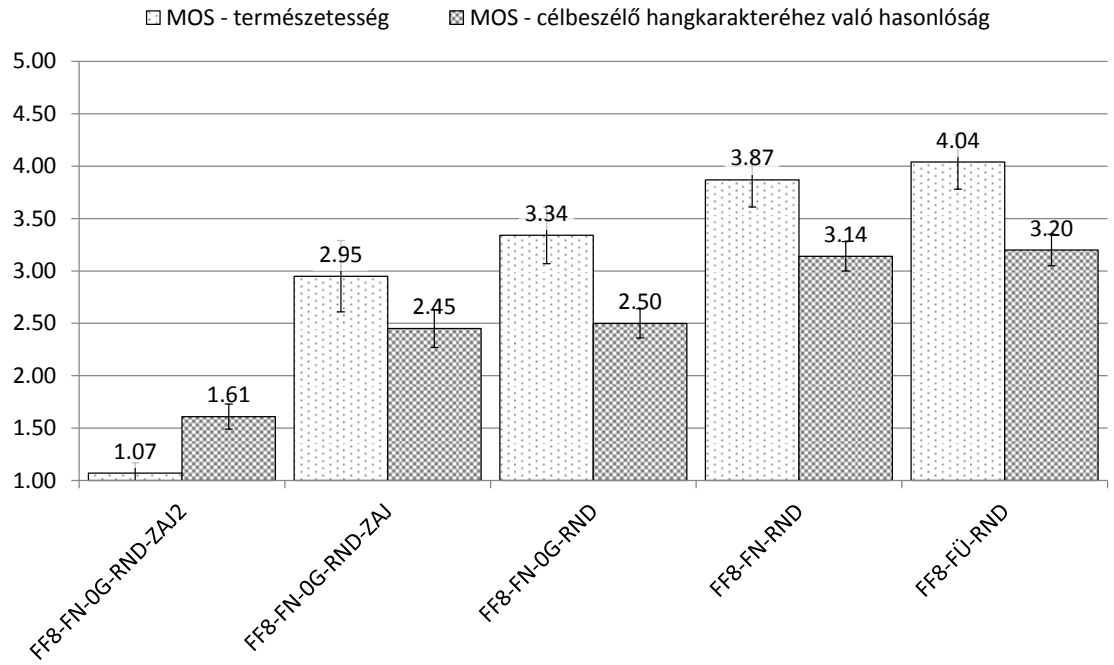
6.5.3.2. 50% PER feletti félspontán adaptációs beszédkorpusz

Az előző pontban végzett kísérlet eredményei alapján további meghallgatásos tesztet terveztem. Ezzel az volt a célom, hogy egy beszélő esetén megállapítsam, hogy 50%-os fonémahiba-arány felett hogyan alakul a gépi beszéd minősége és milyen mértékű hibaarány esetén kell már szignifikáns minőségromlással számolni.

A kísérletet a korábbi eredmények ellenőrzése és egy viszonyítási pont céljából a hiba nélkülinek tekintett kézi átiratra, és az előző pontban létrehozott rendszerekre is kiterjesztettem. További meghallgatásos tesztekhez az előző kísérletben résztvevő FF8 beszélőt választottam ki, mert a rendelkezésre álló átiratok között előnyös összehasonlítási pontnak találtam a 17%-os fonémahiba-arányt. A rosszabb minőségű felismerés szimulálása céljából 0-gram nyelvi modellt és 3000 morfémát használtam, illetve a hullámformához kisebb (ZAJ) és nagyobb (ZAJ2) mértékben zajt kevertem a felismerés előtt. A tesztben az *FF8-FÜ-RND* (férfi 8. beszélő, felügyelt adaptáció, véletlenszerű szelekció), *FF8-FN-RND* (férfi 8. beszélő, felügyelet nélküli adaptáció, véletlenszerű szelekció) és a 14. táblázatban bemutatott adaptációs beszédkorpusszal készült rendszerek vettek részt.

A meghallgatásos teszt két részből állt: a rendszerek természetességét és az eredeti beszélő hangkarakteréhez való hasonlóságát MOS típusú meghallgatásos teszttel mértem. Az *első részben* a tesztalanyoknak egy 5 elemű skálán kellett meghatározniuk, hogy az adott hangminta mennyire természetes. A *második részben* szintén 5 elemű skálán határozták meg

az alanyok, hogy a lejátszott minta mennyire emlékezteti őket az eredeti beszélő hangkarakterére. Mindkét részben egy tesztalany négy hangmintát hallgatott meg, így összesen 20-20 hangmintát osztályoztak. A meghallgatásos tesztben 29-en vettek részt, 12 nő és 17 férfi. A tesztalanyok magyar anyanyelvűek voltak és halláskárosodásuk nem volt. Az átlagéletkor 29 volt, a legfiatalabb tesztalany 19, a legidősebb 60 éves volt. 7 tesztalany beszédtechnológiai szakértő volt. A meghallgatásos teszt internet alapon működött. Az eredményeket a 24. ábra mutatja. Az ábrán a természetességet és a célbeszélőhöz való hasonlóságot mutatom be. Az eredmények alapján megfigyelhető, hogy ahogyan csökken a fonémahiba-aránya, úgy nő a gépi beszéd minősége. Az előző pontban megfogalmazottakat továbbá az is igazolta, hogy *FF8-FN-RND* és *FF8-FÜ-RND* rendszerek között itt sincs szignifikáns minőségbeli különbség, magasabb fonéma hibaarány esetén azonban már számolni kell szignifikáns minőségbeli különbséggel.



24. ábra. Beszédfelismerő (*FF8-FN-**) és kézi átirat (*FF8-FÜ-RND*) alapú beszélőadaptáció vizsgálata szubjektív MOS meghallgatásos tesztekkel.

6.5.4. Konklúzió

Jelen fejezetben eljárást dolgoztam ki felügyelet nélküli félspontán HMM-TTS adaptációra. Az eljárás hatékonyságát meghallgatásos tesztekkel mértem. A tesztek alapján 50% fonéma hibaarány alatt lehetséges a kézi módszertől szignifikánsan nem eltérő minőségű HMM-TTS hang létrehozása. Ez az eredmény az 5.5. fejezetben megfogalmazottaknak a kiterjesztése, hiszen itt már nem csak a fonetikus átiratot és a szegmentálást végeztem automatikusan, hanem a hanganyag szöveges átiratát is automatikus módszerekkel határoztam meg. Rosszabb felismerési pontosság esetén ahogyan nő a fonéma hibaarány, úgy romlik a gépi beszéd minősége. Felügyelet nélküli adaptáció esetén előfordulhat, hogy a felismerő rosszul teljesít (pl. rossz minőségű hangfelvétel, a felismerő tanításától eltérő tématerület). Fontos volt számomra, hogy ezekben az esetekben is minél jobb minőségű gépi hangot lehessen felügyelet nélküli adaptációval létrehozni. Éppen ezért a további kutatásom során olyan eljárást dolgoztam ki, mellyel magasabb fonémahiba-arány esetén is minőségjavulást lehet elérni.

6.6. Felügyelet nélküli beszélőadaptáció hatékonyságának növelése

Az előző fejezetben felügyelet nélküli HMM-TTS alapú beszélőadaptációs eljárást dolgoztam ki félspontán hanganyaggal. Az eljárással jobb (*PER<50%*) minőségű beszédkorpuszok esetén lehetséges a kézi címkézéstől minőségben szignifikánsan nem különböző gépi beszédhangot létrehozni. Jelen fejezetben azt a kérdést vizsgálom, hogy milyen módon lehetne az eljárás hatékonyságát rosszabb (*PER>50%*) minőségű beszédkorpuszok esetén növelni.

6.6.1. Az adaptációs beszédkorpusz előállítása

Az adaptációs beszédkorpuszt a 6.5.1. fejezetben bemutatott módon szegmentáltam és a fonetikus átiratot is a korábban leírtaknak megfelelően készítettem el. Az előző fejezetben a rendelkezésre álló hanganyagból az adaptációs beszédkorpuszt véletlenszerű módszerrel választottam ki. Következő lépésként olyan szelekciós eljárás kidolgozását tűztem ki célul, mellyel lehetséges a felügyelet nélkül adaptáció hatékonyságát növelni.

A 6.3. fejezetben ismertettem a kényszerített illesztést és a beam fogalmát. Korábbi munkám során a kényszerített illesztést egy állandó, a gyakorlatban alapbeállításnak számító beam-el végeztem el. Széles beam esetén a Viterbi keresés mindig be tud lépni a stop állapotba. Ez azt jelenti, hogy ha a szöveges átirat nagymértékben különbözik a hanganyagtól, az eljárás akkor is kijelöl hanghatárokat. A beam szélességének szűkítésével csak a jobban illeszkedő esetek jutnak el a végállapotba. Szűkre szabott beam esetén a Viterbi keresés pedig már nem tud a végállapotba eljutni a legjobb útvonalon sem, amennyiben a szöveges átirat (ami ez esetben az ASR kimenete) jelentősen különbözik a hanganyag tartalmától. A kísérleti rendszerben használt kényszerített illesztővel végzett előzetes tesztek során a fenti működést gyakorlatban is vizsgáltam. A vizsgálatok során, ahogy szűkítettem a beam-et, a rendelkezésre álló félspontán beszéd virtuális mondatait tartalmazó hangfájlok egyre kisebb részén futott le sikeresen a kényszerített illesztés.

Tágabb értelemben véve a beam szélességének módosításával tulajdonképpen a szöveges átirat (felismerő kimenete) és a hanganyag közötti hasonlósági mértéket kerestem. Az eljárás olyan beszédfelismerők esetén is használható, melyekben nem érhető el konfidenciamérték. Azt, hogy melyik beam érték számít „szélesnek”, és mi számít „keskenynek” a hanganyag és a felismerés pontossága is befolyásolja. Továbbá célom egy tetszőleges méretű beszédkorpuszból mintegy 10 percnyi hanganyag kiválasztása volt. A felügyelet nélküli adaptáció során változó minőségű hanganyagokra kell felkészülnünk, melyek esetén különböző módon teljesít a beszédfelismerő, továbbá a kiindulási hanganyag hossza is minden esetben más, ezért nem lehet empirikus úton egzakt beam értéket meghatározni. Ezért eljárást dolgoztam ki a mintegy 10 percnyi (*t_limit*) adaptációs hanganyag automatikus kiválasztásához.

Az eljárás során a félspontán beszéd virtuális mondatait hangfájlokban tárolom (egy virtuális mondatot egy fájlban). Ezekre a fájlokra futtatom le a kényszerített illesztést adott beam szélességgel, és azoknak a hangfájloknak az együttes időbeli hosszát vizsgálom, melyeken sikeresen lefutott a kényszerített illesztés. Ezt a hosszt a továbbiakban *t_adaptation_corpus*-al jelölöm. A beam szélességet iteratív módon úgy állítom, hogy megtaláljam azt az értéket, ami esetén mintegy 10 perc azon hangfájlok időbeli hossza, melyeken sikeresen lefutott a kényszerített illesztés. Az eljárás első változatában maximális beam szélességből indulok ki, és egyesével *lineárisan csökkentem* azt mindaddig, míg az

újabb lépésben nem kerülök „távolabb” a 10 percnyi adaptációs beszédkorpusz hossztól. Ezek alapján a leállási feltétel az, hogy az *i*-edik lépésben a kiválasztott adaptációs beszédkorpusz hossza távolabb esik a 10 perctől, mint az *i-1*-edik, megelőző lépésben. Az eljárás alapját pszeudókód formájában ismertetem:

```
1. i=0
2. t_limit=10 minutes
3. beam[0]=maximum beam width
4. CALL forced alignment WITH beam[0] on each wave file
   RETURNING t_adaptation_corpus[0]
5. DO
6.    i++
7.    beam[i]=beam[i-1]-1
8.    CALL forced alignment WITH beam[i] on each wave file
      RETURNING t_adaptation_corpus[i]
9. WHILE |t_adaptation_corpus[i]-t_limit| ≤ |t_adaptation_corpus[i-1]-t_limit|
```

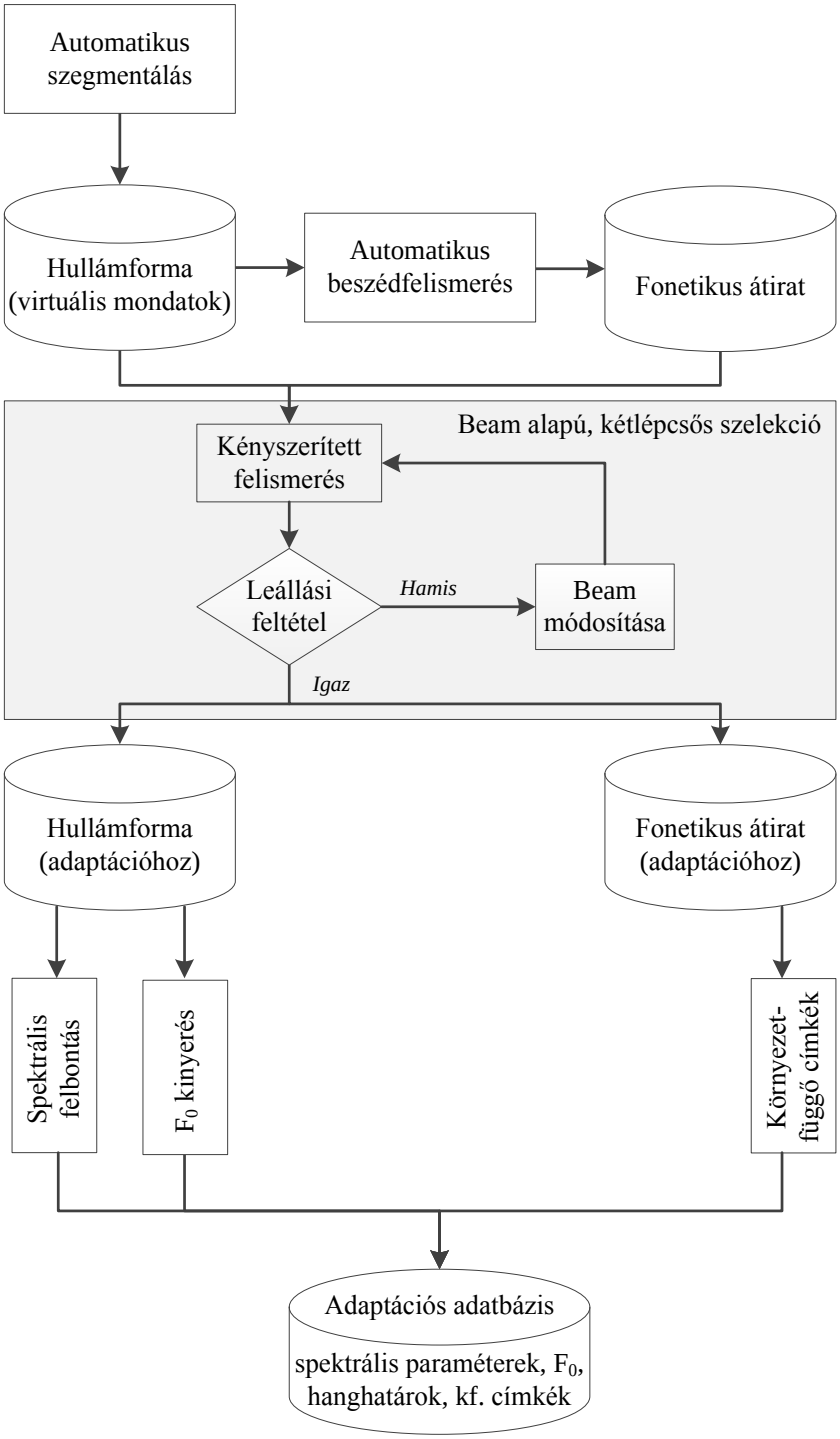
Ha a leállási feltétel teljesül, akkor az eggyel korábbi, *i-1*-edik iteráció során kiválasztott hangfájlokat használom fel a beszélőadaptációhoz.

A beam iterációnként eggyel való csökkentése igen lassú megoldás, mivel a kényszerített illesztés során használt beam szélesség a gyakorlatban sokkal kisebb a maximális beam szélességnél. A gyorsabb működés érdekében az eljárás második változatában *intervallumfelezéssel* keresem az optimális beam értéket. Az eljárás magját ez esetben is pszeudókód formájában ismertetem:

```
 1. i=0
 2. beam_max=beam[0]=maximum beam width
 3. beam_min=0
 4. t_limit=10 minutes
 5. DO
 6.    CALL forced alignment WITH beam[i] on each wave file
       RETURNING t_adaptation_corpus[i]
 7.    IF t_adaptation_corpus[i]>t_limit THEN
 8.       beam_max=beam[i]
 9.       beam[i+1]=beam[i]-floor((beam[i]-beam_min)/2)
10.    ELSE
11.       beam_min=beam[i]
12.       beam[i+1]=beam[i]+floor((beam_max-beam[i])/2)
13.    END IF
14.    i++
15. WHILE beam[i] != beam[i-1]
```

A leállási feltétel az, hogy a beam szélesség két egymást követő lépésben azonos. Ez után még megvizsgálom, hogy a kiválasztott beam értéktől eggyel különböző beamekhez tartozó hangfájlok hossza nincs-e közelebb a 10 perchez. Amennyiben valamelyik esetben közelebb van, akkor azt használom fel az adaptációhoz.

A beam-alapú szelekciós eljárással megvalósított félspontán felügyelet nélküli beszélőadaptáció sematikus felépítését a 25. ábra szemlélteti. Az eljárás működése a szelekciós eljárás (és ezáltal a kényszerített illesztés) kivételével a 6.5.2. fejezetben bemutatottakkal azonos maradt. Az ábrán az előző fejezetekhez képest az új elemeket szürke háttérrel jelölöm. Az ábra visszacsatolással reprezentálja az iterációs eljárás működését. A szelekciós eljárás módosításával gyorsítottam az iteráción.



25. ábra. A felügyelet nélküli, beam-alapú szelekción alapuló beszélőadaptáció felépítése.

6.6.2. Adaptációs beszédkorpusz

A korábbi felügyelet nélküli esethez hasonlóan az átlaghangot az 5.4. fejezetben ismertetett beszédkorpuszokkal tanítottam. Az előző fejezet alapján készített adaptációs beszédkorpuszok tulajdonságait a 15. táblázatban mutatom be. A beszédkorpuszok a 14. táblázatban szereplőktől elsősorban a szelekciós eljárásban térnek el egymástól: korábban véletlenszerűen választottam ki a mintegy 10 percnyi hanganyagot, jelen esetben pedig a beam-alapú szelekciós eljárással. A táblázatban ezt a *BBS* jelöli (*Beam Based Selection*).

15. táblázat. Egy beszélőtől származó, rossz felismerési eredmények szimulálásával készült félspontán adaptációs beszédkorpuszok felügyelet nélküli beszélőadaptációhoz.

Jelölés	Beszélő	Nyelvi modell	Zaj	Időtartam	PER	WER
FF8-FN-0G-BBS	Férfi 8.	0-gram	-	10 perc	52%	100%
FF8-FN-0G-BBS-ZAJ	Férfi 8.	0-gram	-50 dB	9.4 perc	68%	100%
FF8-FN-0G-BBS-ZAJ2	Férfi 8.	0-gram	-25 dB	10.2 perc	88%	100%

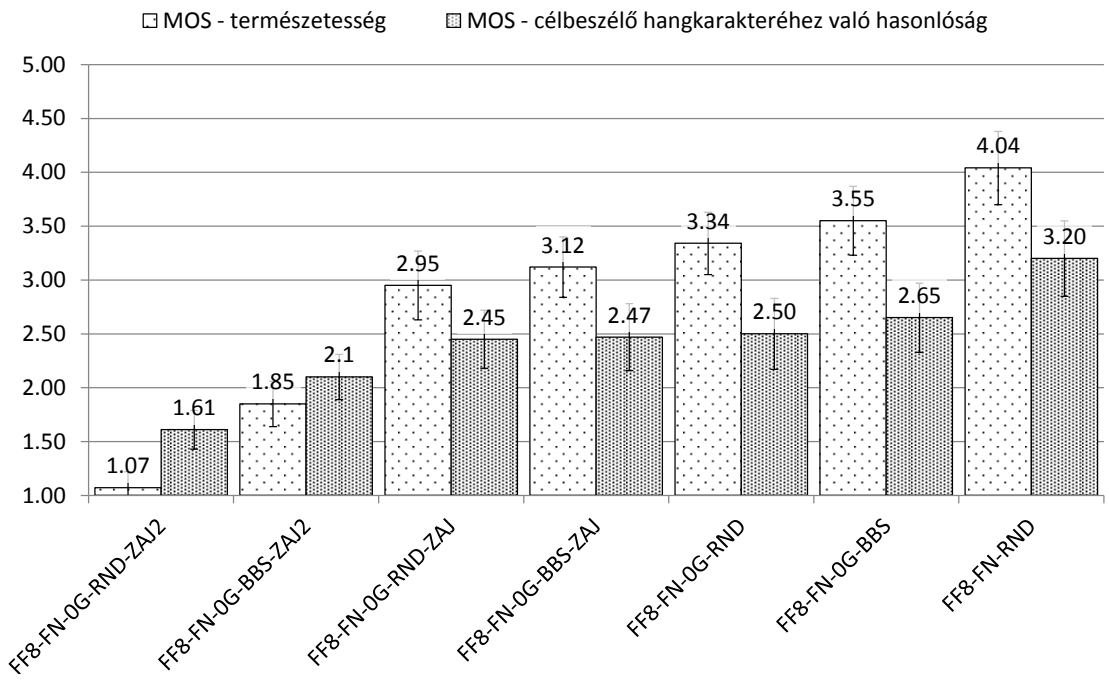
6.6.3. Számszerű kiértékelés

Az eljárás gépi beszéd minőségére gyakorolt hatását szubjektív meghallgatásos tesztekkel mértem. A tesztek során a korábbiaknak megfelelően *RND-vel (Random, véletlenszerű)* jelöltem azt az esetet, amikor véletlenszerűen, *BBS-el (Beam Based Selection, Beam-alapú kiválasztás)* jelöltem, amikor a beam-alapú eljárással válogattam ki a t_{limit} hosszúságú adaptációs beszédkorpuszt. A teszt három részből állt: az első rész a rendszerek természetességét vizsgálta MOS teszttel, a második az eredeti beszélő hangjához való hasonlóságát, szintén MOS teszttel, a harmadik része pedig egy CMOS alapú tesztből állt, ahol az *RND* és *BBS* rendszerek hangminőségét hasonlítottam össze egymással.

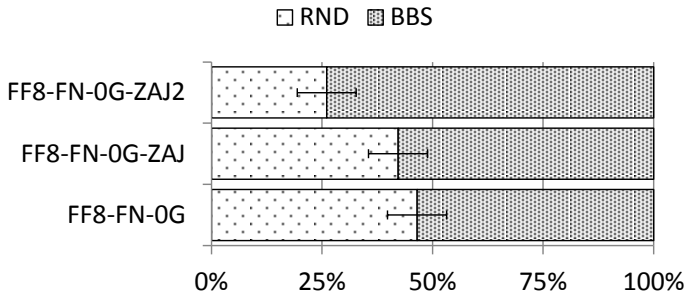
A meghallgatásos teszt *első részében* a tesztalanyoknak természetességük szerint 28 mintát (rendszerenként 4 mintát) kellett osztályozniuk. A *második részben* szintén 28 mintának az eredeti beszélő hangkarakteréhez való hasonlóságát kellett az alanyoknak megállapítaniuk. A meghallgatásos teszt *harmadik részében* pedig párokat kellett összehasonlítaniuk: 12 pár (24 hangminta) esetén kellett megadniuk, hogy melyik minta hangzása természetesebb. A meghallgatásos tesztben 29-en vettek részt, 12 nő és 17 férfi. A tesztalanyok magyar anyanyelvűek voltak és halláskárosodása egyiküknek sem volt. Az átlagéletkor 29 év volt, a legfiatalabb tesztalany 19, a legidősebb pedig 60 éves volt, és közöttük 7-en beszédtechnológiai szakértők voltak. A meghallgatásos teszt internet alapúan működött.

A meghallgatásos teszt első két részében (26. ábra) a MOS értékek alapján megállapítható, hogy nagyobb számú fonéma tévesztés esetén az általam kidolgozott eljárás a véletlenszerű kiválasztásnál jobb minőséget ér el (*FF8-FN-0G-BBS-ZAJ2* szignifikánsan jobb, mint *FF8-FN-0G-RND-ZAJ2*). A fonémahiba-arány csak kismértékben különbözött ebben az esetben a két eljárás között, minőségben mégis szignifikáns javulás mutatkozott. Ez alapján a BBS eljárás okozta minőségbeli javulást elsősorban nem a fonémahiba-arány csökkenésével lehet magyarázni. Az RND és BBS eljárások más részeit választották ki a beszédkorpusznak. Lehetséges, hogy az utóbbi esetben például gyakori fonémáknak jobb minőségű modellje készült el, így a szintézis során ez jelentősen javította a gépi beszéd minőségét.

Kevesebb fonéma tévesztés esetén az eljárásból származó minőségjavulás nem észlelhető a MOS teszteknél. A szubjektív meghallgatásos CMOS részét megfigyelve (27. ábra) láthatjuk, hogy ahogyan nő a fonéma tévesztések száma, úgy egyre jobban teljesít a kidolgozott eljárás. Az *ZAJ* és *ZAJ2* esetekben szignifikáns minőségbeli különbség mutatkozott az *RND* és *BBS* eljárás között a *BBS* javára. Ahogyan csökkent a fonéma tévesztések száma, a két metódus közötti különbség is csökkent (*FF8-FN-0G-BBS* és *FF8-FN-0G-RND*), és szignifikáns minőségbeli különbség már nem volt tapasztalható.



26. ábra. Az RND és BBS alapú rendszerek vizsgálata szubjektív MOS meghallgatásos teszttel.



27. ábra. Az RND és BBS alapú rendszerek vizsgálata szubjektív CMOS meghallgatásos teszttel.

6.6.4. Konklúzió

A 6.5. fejezetben a beszédfelismerő kimenetén alapuló felügyelet nélküli beszélőadaptációt félspontán szövegek esetén mutattam be. Jelen fejezetben ezen eljárás hatékonyságát növeltem azáltal, hogy a kényszerített illesztés során a beam szélességét iteratív módon állítottam. Az iteratív módszer célja az volt, hogy a szelekciós eljárás a kényszerített illesztés szerinti legkedvezőbb adaptációs beszédkorpuszt válassza ki. A módszer hatékonyságát meghallgatásos tesztekkel igazoltam. A tesztek megmutatták, hogy még rossz teljesítményű beszédfelismerővel és/vagy zajos hanganyagok esetén is lehetséges felügyelet nélküli, a véletlenszerű kiválasztásnál szignifikánsan jobb minőségű beszélőadaptáció HMM-TTS rendszerekben.

6.7. Összegzés

Ebben a fejezetben eljárást dolgoztam ki beszédfelismerő kimenetén alapuló felügyelet nélküli HMM-TTS beszélőadaptációra. Felügyelet nélküli beszélőadaptáció esetén emberi beavatkozás nélkül, akár több száz, vagy több ezer új HMM-TTS beszédhang is létrehozható. A kutatást félspontán beszéddel végeztem, amely mind a tervezett beszéd, mind pedig a spontán beszéd jellegzetességeit magába foglalta, az eljárást is ennek megfelelően alakítottam ki. A kísérleteket magyar nyelven végeztem, azonban maga az eljárás nem tartalmaz nyelvfüggő elemeket. A fejezetben bemutatott kutatási eredményeimből megfogalmazott téziseket a 8. fejezetben ismertetem.

Az eredmények alapján megállapítom, hogy a kidolgozott eljárás során generált gépi beszédhang minősége jó minőségű adaptációs hanganyag és beszédfelismerés esetén szignifikánsan nem tér el a felügyelt esetétől (II.1. tézis: 6.4. és 6.5. fejezet). Az eljárás hatékonyságát iteratív beszédkorpusz szelekciós módszerrel javítottam, mely magas fonémahiba-arányú beszédfelismerő kimenet esetén szignifikáns minőségjavuláshoz vezetett (II.2. tézis: 6.6. fejezet).

7. Rejtett Markov-modell alapú szövegfelolvasás illesztése korlátozott erőforrású eszközökre

A HMM-TTS tanítása igen számításigényes folyamat, amely több hétig is eltarthat napjaink csúcsteljesítményű számítógépein. A beszéd előállítása modern asztali számítógépeken impulzus-zaj beszédkódoló esetén valós időnél gyorsabban történik, azonban kisebb erőforrású eszközökön jelentős kezdeti késleltetés és valós időnél lassabb működés várható. Bár igaz, hogy napjaink mobil eszközei nagy teljesítményű, akár több magos processzorral is fel vannak szerelve, de erőforrásaikon számos rendszer- és utólag telepített alkalmazás osztozik. Ezen túl a nagy számítási igény jelentősen csökkenti az akkumulátor töltöttségét, ami hátrányosan hat a mindennapi használat során. Az előző fejezetekben bemutatott kutatásaim eredményeképp számos felügyelt és felügyelet nélküli HMM-TTS rendszert hoztam létre. Célom ezen rendszerek korlátozott erőforrású (például mobil) eszközökre való illesztése volt.

Jelen fejezetben a korlátozott erőforrású eszközökön megvalósított HMM-TTS megoldásokat ismertetem, és új módszereket dolgozok ki a számítási igény csökkentése céljából. Inkrementális lépésekben megoldást kínálok a számítási idők csökkentésére, melyek eredményességét mérésekkel és szubjektív meghallgatásos tesztekkel igazoltam.

7.1. Áttekintés

Számos megoldás létezik gépi szövegfelolvasásra mobil készülékeken. Mivel az adatkapcsolat sokszor nem állandó és külföldön az adatforgalom ára még mindig magas, ezért jelenleg a kliens oldali megoldások a versenyképesek. Az artikulációs szintézis korlátozott erőforrású készülékeken való futtatásához elvi módosítások szükségesek, amelyek kutatói munkám során hasznos iránymutatók voltak [92]. A diád-alapú rendszerek hely- és számításigénye megfelelően alacsony a mobil eszközökön való futtatáshoz, de a beszéd minősége gépies [93]. Az elem összefűzés alapú szövegfelolvasás számításigénye a nagy keresési tér miatt magas, továbbá a jó minőségű gépi beszéd létrehozásához a többi megoldáshoz képest nagy a tárkapacitás igénye (>100 MByte, de akár GByte nagyságrendbe is eshet) [94]. Kisebb adatbázisok esetén (<10 MByte), ahogyan csökken a keresési tér mérete, gyors működésre képesek, azonban így romlik a gépi beszéd minősége. A HMM-TTS alapú megoldás esetén hasonló beszédminőség mellett impulzus-zaj gerjesztésű beszédkódolóval a szükséges tárkapacitás legalább egy nagyságrenddel kisebb lehet (1-2 MByte). Ez rosszabb minőségű gépi beszédet eredményez, mint a korábban bemutatott kevert gerjesztésű megoldás. Ezt a minőségromlást azonban a mobiltelefon hangszórójának a karakterisztikája jelentősen elfedi, és a kevert gerjesztésű beszédkódoló több jellemző paraméterfolyammal dolgozik, ami tovább növelné a számításigényt. Ezért kutatásom ebben a részében impulzus-zaj gerjesztésű beszédkódolóval dolgoztam.

Korábban is foglalkoztak a HMM-TTS korlátozott erőforrású eszközökre való illesztésével. Kim és munkatársai a *mel-kepszt rális* és *spektrumvonal páros* (*LSP*, *Line Spectral Pair*) paramétereket hasonlítják össze, megvizsgálja a kétsávos gerjesztési modelleket és egyesíti a spektrumvonal páros analízist a kétsávos gerjesztési modellekkel [95]. A mel-kepsztrális együtthatók rendjének és az impulzus válasz hosszának csökkentésével szignifikáns minőségromlás nélküli sebességnövekedést és adatbázis méret csökkenést értek el mobil eszközökre szánt HMM-TTS rendszerekben [96]. A reaktív HMM szintézis fonéma szinten generálja a paramétereket, és bár elsődleges célja a paraméterek

valós idejű módosítása, gyorsabb a válaszideje, mint az alap rendszereknek – sajnos minőségromlás mellett [97]. Zen és munkatársai a szükséges tárhely méretet fixpontos számábrázolás bevezetésével és a döntési fák méretének csökkentésével érték el [16]. Eredményeik alapján akár 100 kByte-os futásidejű adatbázis mellett is érthető marad a beszéd, a teljesítménybeli változásokra azonban nem tér ki. Oura és munkatársai a HMM modell paramétereket egységesíti, így csökkentve a HMM-TTS számára szükséges tárhelyet [98].

Ezen cikkek mindegyike a HMM-TTS méretének vagy a paramétergenerálás sebességének optimalizálásával foglalkozik. Korábban azonban tudomásom szerint nem született publikáció a szubjektív beszédminőség kapcsolatának vizsgálatáról a kódtábla alapú zajgenerátorral, a spektrumvonal páros paraméterfolyammal és a döntési fák méretével. Továbbá az ismertetett irodalmak egyike sem foglalkozott korlátozott erőforrású eszköz aktuális terhelését figyelembe vevő, párhuzamos működésű beszédelőállító eljárással.

Első lépésként megállapítottam a leginkább számításigényes részeket, majd ezen számítások sebességének korlátozott erőforrású eszközökön való, szignifikáns minőségbeli romlás nélküli gyorsítását végeztem el.

7.2. Beszédkorpusz

A korlátozott erőforrású eszközökre való HMM-TTS illesztést angol nyelvű beszédkorpusszal végeztem. A tanításhoz a szabadon hozzáférhető, beszédtechnológiai célokra általánosan elfogadott pittsburghi Carnegie Mellon Egyetem Beszédtechnológiai Intézetében rögzített ARCTIC adatbázisok közül az SLT jelű női beszélőt használtam [99]. Kutatásom nem tartalmaz nyelvfüggő elemeket, ezért a benne megfogalmazott eredmények általánosak. Kutatásom impulzus-zaj alapú beszédkódolóval végeztem, amire nehézség nélkül át lehet térni a kevert gerjesztésű módszerről. Ezáltal a korábbi fejezetben bemutatott eredményeim alapján készült rendszereket is lehet a korlátozott erőforrású rendszerekhez illeszteni.

7.3. A beszédelőállítás sebességének mérése

Annak érdekében, hogy az egyes lépések teljesítménybeli növekedését meg tudjam állapítani, fontos volt a beszédelőállítás sebességének pontos mérése. A méréseket automatikus úton, programkódban rögzített időbélyegekkel végeztem. Minden mérést 10 alkalommal ismételtem meg, és a mért értékek számtani átlagát vettem. A kutatás során a beszédelőállítás időtartamának következő három fázisát mértem:

(1) a HMM adatbázisok betöltésének ideje,
(2) a felolvasandó bemeneti szövegre jellemző paraméterfolyamok előállításának ideje,
(3) a paraméterfolyamból beszédkódoló eljárással a hullámforma előállításának ideje a beszédhang megszólalásáig.

A felolvasott szöveg *Lewis Carroll: Alice in Wonderland* című regény első fejezetének az első mondata volt: „*Alice was beginning to get very tired of sitting by her sister on the bank...*”. Normális beszédtempóval mintegy 17 másodpercbe telik felolvasni ezt a mondatot. Azért választottam ilyen hosszú mondatot, mert így jól meg tudtam figyelni az (1), (2) és (3) fázisok időigényét.

A kutatási eredmények igazolására szolgáló méréseket három különböző teljesítményű mobil eszközön végeztem el (Apple iPhone, Samsung Galaxy Spica GT-i5700, HTC Desire A8181), melyek számítási kapacitással kapcsolatos főbb paramétereit a 16. táblázat mutatja be. A készülékekre a táblázat alapján *Mob1, Mob2* és *Mob3*-ként hivatkozom a továbbiakban.

16. táblázat. Az optimalizálás során használt korlátozott erőforrású eszközök.

Készülék	CPU típus	CPU órajel [MHz]
Mob1 (iPhone)	Samsung ARM 11	412
Mob2 (Spica)	Samsung S3C6410	800
Mob3 (Desire)	Qualcomm QSD8250	1000

7.4. A gépi beszéd minőségét befolyásoló eljárások

A várakozásomnak megfelelően és a mérések alapján a leginkább számításigényes fázis a paramétergenerálás és a beszédkódolás volt. Ezért kutatásom során ezekre koncentráltam. Jelen fejezetben – figyelve a gépi beszéd minőségének változását – azokat a lépéseket ismertetem, melyek hatással vannak a gépi beszéd minőségére. A méréseket inkrementális módon végeztem: minden esetben az előző pont alapján elvégeztem a sebesség mérését és meghallgatásos teszttel mértem a gépi beszédhang minőségét. Ha a gépi beszédhang minőségében nem volt szignifikáns változás, a méréseket azon rendszerrel előállított mintákkal folytattam. A fejezet végén minden elkészült rendszerrel is végeztem egy közös meghallgatásos tesztet, ahol a 4.5. fejezet alapján ANOVA és Tukey módszerekkel vizsgáltam, hogy a kiinduló és végállapot között jelentkezett-e szignifikáns minőségbeli különbség.

7.4.1. Beszédkódolási eljárást érintő módosítások

Első lépésként a beszédkódoló eljárást módosítottam több lépésben. A beszédkódoló eljárás és annak paraméterei hatással vannak a gépi beszéd minőségére, ezért minden egyes lépés során MOS tesztekkel figyeltem, hogy szignifikánsan nem romlik-e a gépi beszéd minősége.

7.4.1.1. Zöngétlen hangok gerjesztési modellje

A 2.1.1. fejezetben ismertetett módon a zöngétlen hangok gerjesztését impulzus-zaj alapú beszédkódoló esetén Gauss-eloszlású fehér zajjal modellezik. A Box-Muller eljárás [100] független, Gauss eloszlású, 0 várható értékű, egységnyi szórású fehér zajt hoz létre. HMM-TTS rendszerekben is ezt az eljárást használják elsődlegesen. A Box-Muller eljárás során feltételezzük, hogy U1 és U2 független, egyenletes eloszlású véletlenszerű változók a (0,1] tartományban, és ekkor

$$X = \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \tag{38}$$

$$Y = \sqrt{-2\ln(U_1)}\sin(2\pi U_2) \tag{39}$$

független, véletlenszerű változók Gauss eloszlásúak lesznek. A fenti megoldás a lebegőpontos számítás, az U1 és U2 változók futás időben való generálása, illetve a trigonometriai

műveletek miatt matematikai segédprocesszorral sokszor nem rendelkező korlátozott erőforrású eszközökön nem nevezhető ideálisnak.

Kódtábla alapú, fixpontos Gauss zaj generátorral már korábbi korlátozott erőforrású rendszerek esetén is a lebegőpontos számábrázolással szemben nagymértékű (mintegy tízszeres) teljesítménynövekedést értek el [101]. Igaz, hogy a kódtábla és a fixpontos számábrázolás az előző eljáráshoz képest pontatlanságot okoz, azonban várhatóan ez a korlátozott erőforrású eszközön nem fog érzékelhető minőségromlást okozni. Ezen indokok miatt korlátozott erőforrású HMM-TTS rendszerek esetén a fenti megközelítést használtam fel. HMM-TTS rendszerekben korábban tudomásom szerint a gyorsabb működés érdekében nem használtak kódtábla alapú Gauss-eloszlású zajgenerátort.

7.4.1.2. Spektrális modellezés

A spektrális együtthatók elemzése területén a HMM-TTS rendszerekben az általánosan elterjedt módszer az *MGC* (*Mel-Generalized Cepstrum*) [44] és az *MGC-LSP* (*Mel-Generalized Cepstrum-Line Spectral Pairs*) [102] megközelítés használata. Az *MGC* a kepsztrum általánosított logaritmusa alapján számolt, a percepciós Mel-skála szerint módosított változata. Az *MGC* alapján a beszéd *H(z)* spektruma *c(m) MGC* együtthatókkal modellezhető a következőképpen:

$$H(z) = \begin{Bmatrix} \left(1 + \gamma \sum_{m=0}^{M} c(m)\psi_\alpha^m(z)\right)^{\frac{1}{\gamma}}, & -1 \le \gamma < 0 \\ exp\left\{\sum_{m=0}^{M} c(m)\psi_\alpha^m(z)\right\}, & \gamma = 0 \end{Bmatrix} \tag{40}$$

ahol $\psi_\alpha^m(z)$ a $\psi_\alpha(z)$ mindent áteresztő szűrő átviteli függvényének *m*-edik hatványa:

$$\psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{41}$$

A fenti egyenletekből látható, hogy $(\alpha, \gamma) = (0, -1)$ esetén a spektrumot (toldalék csövet) all-pole, tehát zérusokat nem tartalmazó formában modellezzük. Ilyenkor a pólusok jellemzően a rezonancia pontok, tehát a spektrum helyi maximumainak felnek meg. $(\alpha, \gamma) = (0,0)$ esetben a fenti egyenlet a kepsztrális megközelítést követi [102].

Amennyiben $\gamma \neq 0$, a (40)-es egyenletet felírhatjuk a következőképpen is:

$$H(z) = \frac{\widetilde{K}}{C^{-\frac{1}{\gamma}}(z)} \tag{42}$$

$$C(z) = 1 + \gamma \sum_{m=1}^{M} c'(m)\,\psi_\alpha^m(z), \quad -1 \le \gamma < 0 \tag{43}$$

$$\widetilde{K} = \{1 + \gamma c(0)\}^{\frac{1}{\gamma}} \tag{44}$$

$$c'(m) = \frac{c(m)}{1 + \gamma c(0)}, \quad m = 1, 2, \ldots, M \tag{45}$$

Ekkor a (43)-as egyenlet egy *LPC* (*Linear Predictive Coding*) polinom, és így az *MGC-LSP* paraméterek meghatározása az *LSP* paraméterek kiszámításával azonos módon történik [103]. Az *MGC* és *MGC-LSP* paraméterek alapján való spektrális formálást HMM-TTS rendszerekben leggyakrabban *MLSA* szűrővel valósítják meg. Az ideális *MLSA* szűrő átviteli függvénye azonban nem valósítható meg, ezért gyakorlatban huszad-rendű Padé becsléssel közelítik [104]. Ez növeli a számítások komplexitását, hiszen ekkor nemcsak a spektrális felbontás rendjével kell számolnunk, hanem a Padé becslés rendjével is.

Amennyiben *MGC* és *MGC-LSP* paraméterekről áttérünk csupán *LSP* paraméterek használatára, a spektrális formálást *LPC* eljárással hajthatjuk végre, így jelentősen egyszerűsödik a rendszerünk, hiszen csak a spektrális felbontás rendjének megfelelő rendű szintézis szűrőre van szükségünk. Továbbá várhatóan, ahogyan csökkentem az *LSP* felbontás rendjét, annál gyorsabban fog működni. Mindeközben rosszabb beszédminőség is várható. Ezért meghallgatásos teszteket végeztem, hogy megkeressem annak a határát, amikor még szignifikánsan nem érzékelhető az *LSP*-re való átállás, és a spektrális felbontás rendjének csökkentése (lásd 7.4.3.).

7.4.2. A döntési fa méretének korlátozása

A HMM-TTS tanítása során a hasonló paraméterhalmazokat az *MDL* kritérium alapján osztályokba soroljuk (lásd 2.4. fejezet). Impulzus-zaj alapú beszédkódoló esetén külön paraméterfolyam vonatkozik a spektrális, a gerjesztési és az időzítési paraméterekre. Ezekhez készíthetünk kapcsolt (*tied*) és különválasztott (*untied*) döntési fákat. A döntési fák kapcsolása csökkenti a szükséges tárhely méretét, azonban a futás idejű számítási teljesítményre nincs különösebb hatással [98], így kutatásom során különválasztott döntési fákkal dolgoztam.

A döntési fák méretének korlátozásával kevesebb *ág* és *levél* jön létre, ezáltal csökken a futás időben bejárandó fa mérete, ami sebességnövekedéshez vezet. A kisebb méretű döntési fák várhatóan a gépi beszéd minőségromlásához vezetnek, hiszen a döntési fa egyes ágaiban több, kevésbé hasonló paraméter kerül egy halmazba. A célom az volt, hogy megállapítsam, hogy milyen mértékben lehet a döntési fa méretét csökkenteni ahhoz, hogy a gépi beszéd minőségét ez szignifikáns mértékben még ne befolyásolja.

7.4.3. Számszerű kiértékelés

A számszerű kiértékelés során a fentiek alapján inkrementális lépésekben a következő HMM-TTS rendszereket vizsgáltam:

- *Kódtábla alapú, fixpontos Gauss-eloszlású fejér zaj* bevezetése a zöngétlen hangok gerjesztéséhez.
- *Spektrális reprezentáció módosítása MGC*-ről *LSP*-re, továbbá a szintézisszűrő módosítása *MLSA*-ról *LPC*-re.
- *Az LSP felbontás rendjének módosítása*: A mérések során a tanítást 24-ed, 22-ed, 20-ad, 18-ad, 14-ed, 12-ed és 10-ed rendű LSP felbontással végeztem. 24-ed, 22-ed és 20-ad rendű szűrők esetén az összes tesztalannyal a meghallgatásos tesztet nem végeztem el, mert a beszédtechnológiai szakemberek által végzett előzetes teszt azt mutatta, hogy ezek minősége nem különbözik a 18-ad rendű LSP felbontástól.

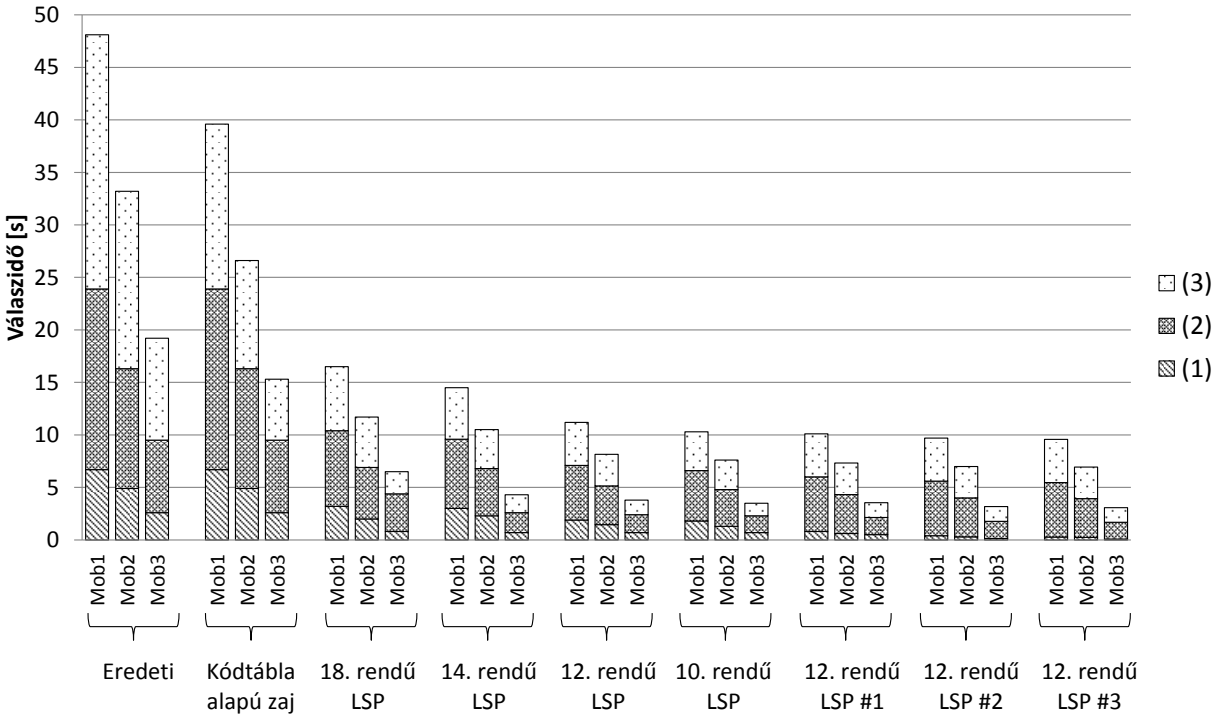
- *A döntési fák méretének csökkentése:* További teljesítménynövekedést okoz a döntési fák méretének a korlátozása. A 17. táblázat alapján 4 különböző méretű döntési fával mértem meg a jelen téziscsoport bevezetőjében található lépések futásához szükséges időket.

17. táblázat. A mobil eszközön való optimalizálás során használt döntési fák méretei.

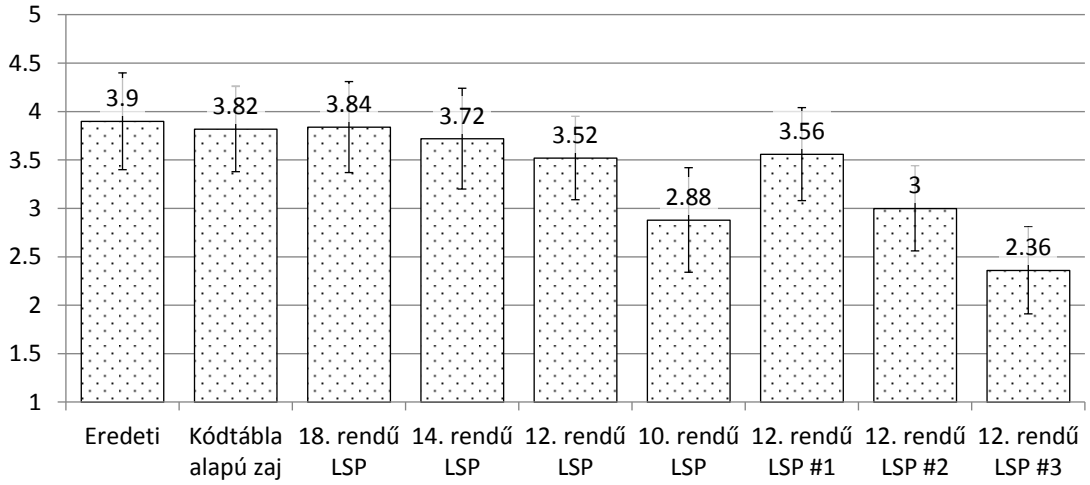
Beállítások	**A döntési fa leveleinek a száma**			**Méret [KByte]**
	LSP	**LogF_0**	**Időtartam**	
Alapkonfiguráció	2883	3545	555	666
#1	2282	2104	376	463
#2	1227	1344	172	214
#3	651	543	79	140

A kiértékelés során a 7.3. fejezet alapján mértem a beszédelőállítás *sebességét*, továbbá MOS alapú meghallgatásos tesztekkel a *gépi beszéd minőségét*. A meghallgatásos tesztek során véletlenszerű módon, egyenletes eloszlással a 16. táblázatban található eszközökön csendes környezetben a tesztalanyok fejével egy magasságban, a tesztalanyoktól mintegy 50 cm távolságban játszottam le hangmintákat. A minták az eszközök hangszóróján szólaltak meg. Minden tesztalany esetén 40 előre elkészített mintából egyenletes eloszlással, véletlenszerűen 10-et választottam ki. A meghallgatásos tesztben 21-en vettek részt, 8 nő és 13 férfi. A tesztalanyok magyar anyanyelvűek voltak legalább középfokú angol nyelvtudással. Közöttük halláskárosult nem volt. Az átlagéletkor 29 év volt, a legfiatalabb tesztalany 18, a legidősebb 68 éves volt. Közülük 5-en beszédtechnológiai szakértők voltak.

A sebességmérés eredményeit a 28. ábra, a meghallgatásos teszt eredményeit pedig a 29. ábra mutatja be. A két ábrán egyazon lépéshez tartozó számítási idő és a gépi beszéd minőségének értékei azonos oszlopban szerepelnek. A 28. ábrán a számítások (1), (2) és (3) részeit egymás felett ábrázoltam, és így az ábráról le lehet olvasni, hogy a gépi beszédhang megszólalásáig mennyi időre volt szükség. A sebességmérést a 7.3. fejezet alapján végeztem. A meghallgatásos teszt esetén ANOVA módszerrel Tukey post hoc teszttel vizsgáltam, hogy van-e a mért eredmények közötti szignifikáns eltérés. (A *12. rendű LSP #1* rendszer minősége a meghallgatásos teszt mérési bizonytalanságai miatt lehetett kicsivel jobb, mint a nagyobb méretű döntési fával rendelkező *12. rendű LSP* rendszer minősége.)



28. ábra. A bevezetett módosítások hatására elért sebességnövekedés korlátozott erőforrású eszközökön futó HMM-TTS rendszerekben.



29. ábra. A bevezetett módosítások hatásának vizsgálata a korlátozott erőforrású eszközökön futó HMM-TTS rendszer minőségére (MOS teszt).

7.4.4. Konklúzió

A fenti eredményekre támaszkodva a tesztelt mobil eszközökön szignifikáns minőségromlás nélkül a futási idő jelentős mértékben csökkent. A 12. rendű LSP spektrális reprezentáció esetén, az #1-es esetnek megfelelő döntési fa korlátozása mellett kódtábla alapú zajgenerálással szignifikáns minőségromlás nélkül a számítás mintegy ötszörösére gyorsult. Az ennél gyorsabb működést mutató esetekben már észlelhető volt szignifikáns minőségbeli eltérés, ezért ezen rendszereket elvetettem.

7.5. A gépi beszéd minőségétől független eljárás

A gépi beszéd minőségét érintő lépések után a gépi beszéd minőségétől független módosítást végeztem a reakcióidő csökkentése céljából. Kutatásom során a rejtett Markov-modellek paramétergeneráló algoritmusát módosítottam oly módon, hogy az az adatbázis méret növekedése és a gépi beszéd minőségének csökkenése nélkül gyorsabb működésre legyen képes. Ehhez a korábban használt rekurzív paramétergenerálás módszerén módosítottam [34,35].

7.5.1. Rekurzív paramétergenerálás

A 2.6.1. fejezetben ismertetett (33)-(35) egyenletek minél gyorsabb kiszámolása kulcsfontosságú szerepet játszik a HMM-TTS korlátozott erőforrású eszközökön való futásához. Az általános megoldás helyett *Cholesky* vagy *QR dekompozíciót* használva a megoldás komplexitás $O(TM^3L^2)$-re módosul, ahol

$$\mathrm{L} = \max_{n\epsilon\{1,2\},s\epsilon\{+,-\}} L_s^{(n)} \tag{46}$$

Diagonális kovariancia mátrix esetén a komplexitás tovább csökken $O(TML^2)$-re. Tovább lehet növelni az állapot- és paramétersorozat kiszámításának sebességét a (33)-as egyenlet tulajdonságainak figyelembe vételével. Az eljárást Tokuda és munkatársai dolgozták ki az alábbiak szerint [34,35]. Helyettesítsük (q_t, i_t) *t*-edik keret q_t állapotának i_t-edik összetevőjét $(\hat{q}_t, \hat{\imath}_t)$-vel. Ekkor

$$\hat{R}\hat{c} = \hat{r} \tag{47}$$

ahol

$$\hat{R} = R + w_t D w_t^T \tag{48}$$

$$\hat{r} = r + w_t d \tag{49}$$

$$D = U_{\hat{q}_t,\hat{\imath}_t}^{-1} - U_{q_t,i_t}^{-1} \tag{50}$$

$$d = U_{\hat{q}_t,\hat{\imath}_t}^{-1}\mu_{\hat{q}_t,\hat{\imath}_t} - U_{q_t,i_t}^{-1}\mu_{q_t,i_t} \tag{51}$$

Az (48) egyenletet hasonló az adaptív szűrőknél használt *legkisebb négyzetek módszere* (*Recursive Least-Squeres, RLS*) idő frissítéséhez, ahol a $w_t D w_t^T$ rangja *3M* és az *R* rangja *TM* [105]. Az *RLS* algoritmus analógiájára rekurzív módon határozható meg *c* értéke $\hat{c}$-ből. Az algoritmust a 18. táblázat ismerteti. Mivel $\mathrm{w_t}$ közel összes eleme nulla, elsősorban az (RLS6.) egyenlet határozza meg a számítás komplexitását, amely $O(T^2M^3)$, diagonális U_{q_t,i_t} esetén $O(T^2M)$. Az emberi beszéd tulajdonsága alapján feltételezve, hogy a várható értéket és a kovarianciát a *t*-edik keret esetén csak az *S* szomszédságú paraméterek befolyásolják, a komplexitás $O(S^2M^3)$-re csökken, diagonális kovariancia esetén $O(S^2M)$-re[14]. Ezen algoritmus segítségével rekurzív módon tudjuk megoldani a (33)-as egyenletet minden (q_t, i_t) al-állapot sorozatra és ezt az optimális al-állapot sorozatot fogjuk az adott keretben használni.

[14] Empírikusan S=30 elegendő.

18. táblázat. RLS alapú algoritmus a (q_t, i_t) al-állapot $(\hat{q}_t, \hat{\imath}_t)$-vel való helyettesítésére; [34,35] alapján.

Az iteráció során az előző lépésben kiszámolt $\hat{\mathbf{c}}$, $\hat{\boldsymbol{P}}$ *és* $\hat{\boldsymbol{\epsilon}}$ *értékeket rendre c, P és* $\boldsymbol{\epsilon}$*–re helyettesítjük be és számoljuk ki az egyenleteket.*	
$\pi = Pw_t$	(RLS1.)
$\nu = w_t^T \pi$	(RLS2.)
$k = \pi\{I_{3M} + (U_{\hat{q}_t,\hat{\imath}_t}^{-1} - U_{q_t,i_t}^{-1})\nu\}^{-1}$	(RLS3.)
$\hat{c} = c + k\{U_{\hat{q}_t,\hat{\imath}_t}^{-1}(\mu_{\hat{q}_t,\hat{\imath}_t} - w_t^T c) - U_{q_t,i_t}^{-1}(\mu_{q_t,i_t} - w_t^T c)\}$	(RLS4.)
$\hat{\varepsilon} = \varepsilon + (\mu_{\hat{q}_t,\hat{\imath}_t} - w_t^T \hat{c})^T U_{\hat{q}_t,\hat{\imath}_t}^{-1}(\mu_{\hat{q}_t,\hat{\imath}_t} - w_t^T c) - -(\mu_{q_t,i_t} - w_t^T \hat{c})^T U_{q_t,i_t}^{-1}(\mu_{q_t,i_t} - w_t^T c)$	(RLS5.)
$P = P - k(U_{\hat{q}_t,\hat{\imath}_t}^{-1} - U_{q_t,i_t}^{-1})\pi$	(RLS6.)
$P=R^{-1}$	(RLS7.)

7.5.2. Párhuzamos működésű beszédelőállítás az aktuális terhelés figyelembevételével

A (33)-as egyenlet előző fejezetben bemutatott megoldása lehetővé teszi, hogy idő-rekurzív módon számoljuk ki a (q_t, i_t) al-állapot sorozatokhoz tartozó paraméterfolyamokat. Ezáltal nem kell megvárni a (33)-as egyenlet teljes megoldását a beszédkódoló eljárás megkezdéséhez, hanem keretenként átadhatjuk a paraméterfolyamot a beszédkódolónak. A beszédkódoló kimenetén megjelenő hullámformát pedig lejátsszuk a mobil eszközön.

Általános szövegfelolvasó architektúrákban annak érdekében, hogy platform-független maradjon a megoldás, a hullámforma lejátszás nincsen megvalósítva. A hullámforma lejátszás bevezetésével igaz, hogy platformfüggővé válik a szövegfelolvasás, azonban a válaszidőt (a felolvasandó szöveg átadása és a gépi beszédhang megszólalása között eltelt időt) a paraméterfolyam, a beszédkódolás és a hullámforma lejátszás párhuzamosításával lehetséges csökkenteni. Nevezzünk szegmensnek *k* darab keretből álló paraméterfolyamot. Ekkor a párhuzamos működés az alábbiak szerint valósítható meg:

1. Paraméterfolyam generálás idő-rekurzív algoritmussal adott szegmenshez (k keret). Ezt átadom a beszédkódoló eljárásnak (2. lépés) és folytatom a paraméterfolyam kiszámítását a következő szegmensre.
2. A szegmenshez tartozó paraméterfolyamból hullámforma készítése beszédkódoló eljárással.
3. Szegmenshez tartozó hullámforma hozzáadása a lejátszási sorhoz.

A szegmens mérete döntő fontosságú; ha túl rövid, akkor szaggatottá válik a gépi beszéd, amennyiben túl hosszú, akkor pedig fölösleges késletetés terheli a rendszert. Ezen túl a szegmens hossza függ az adott rendszer teljesítményétől és aktuális terhelésétől is. Ezen okok miatt a szegmens hosszát futásidőben határoztam meg a hálózaton keresztüli hang lejátszás analógiájára. A Ramjee és munkatársai által kidolgozott eljárás az alábbiak szerint működik [106]. Legyen n^i az *i*-edik audio csomag teljes késleltetése a hálózatban. Minden bejövő csomag esetén számoljuk ki d^i-t, a becsült késleltetést, és v^i-t, a késleltetés szórását a következőképpen:

$$\hat{d}^i = A * \hat{d}^{i-1} + (1 - A) * n^i \tag{52}$$

$$\hat{v}^i = A * \hat{v}^{i-1} + (1 - A) * \left|\hat{d}^i - n^i\right| \tag{53}$$

A (52) és (53) egyenleteket minden csomag esetén kiszámoljuk, de csak szünetek után használjuk fel. Szünet után a következő képlet alapján számoljuk ki, hogy miután beérkezett az *i*-edik csomag, mennyi időn belül kezdődjön annak lejátszása:

$$p^i = \hat{d}^i + B * \hat{v}^i \tag{54}$$

Az (52) és (53) egyenletekben szereplő *A* konstans adja meg a becslés memóriáját, a (54) egyenletben szereplő *B* pedig a késleltetés / csomagvesztés arányát határozza meg. A gyakorlatban *A=0.998002* és *B=4* értékeket használnak.

A fenti eljárást a következőképp módosítottam a HMM-TTS számára: jelölje n^i az *i*-edik szegmens paraméterfolyam generálásának és beszédkódolásának együttes idejét. Ekkor d^i-t, v^i-t és p^i-t az (52)-(54) egyenleteknek megfelelően számolom ki $d^1=n^1$, $k^1=30$, $v^0=0$ kezdeti értékek és *A=0.99, B=4* konstansok mellett (*i>0*). Az *i+1*-edik szegmens kereteinek a számát megadó k^{i+1} értéket 60 keretenként (*S=30* empirikusan, lásd 7.5.1.) az alábbi képlet alapján módosítom, ahol T_{keret} a keret hosszát adja meg (a kísérleti mintarendszerben 25 ms):

$$k^{i+1} = \left\lceil \frac{p^i}{T_{keret}} \right\rceil \tag{55}$$

A párhuzamos működésű beszédelőállítás sematikus blokkdiagramját a 30. ábra szemlélteti.

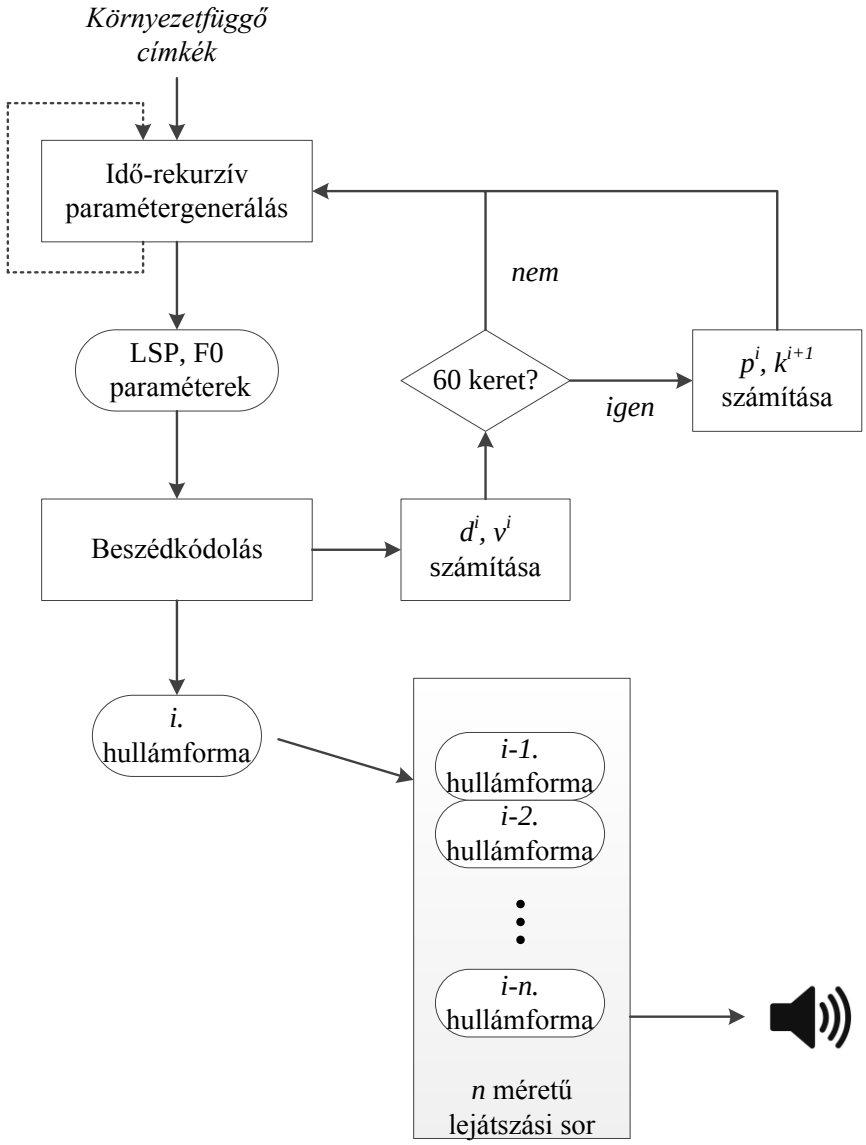
7.5.3. Számszerű kiértékelés

Az ebben a fejezetben ismertetett módszer nincs hatással a gépi beszéd minőségére, ezért nem volt szükség szubjektív meghallgatásos tesztekre.[15] A számításokhoz szükséges időket a 31. ábra mutatja be. Az ábra bal oldalán a 7.4. fejezetben előállt rendszer látható, a jobb oldalán pedig a jelen fejezetben bemutatott módosítások hatása figyelhető meg.

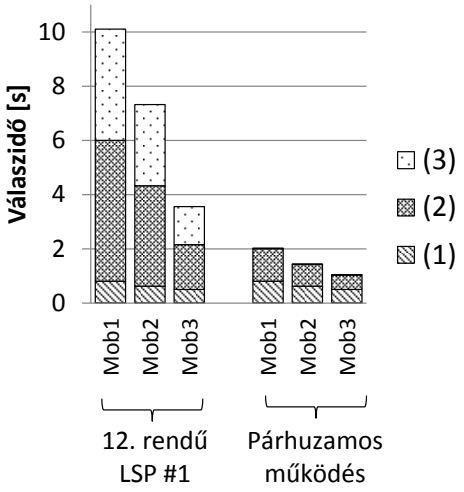
7.5.4. Konklúzió

A paramétergenerálás, beszédkódolás és hullámforma lejátszás párhuzamosításával tovább tudtam csökkenteni a HMM-TTS válaszidejét. A bemutatott módszer az *RLS* szerű idő-rekurzív eljárásra támaszkodva a hálózaton keresztüli hanglejátszás analógiájára futásidőben határozza meg a lejátszás egységét képző szegmensek méretét a HMM-TTS rendszerben. A bemutatott eljárás eredményeképp a 7.4. fejezet során előállított rendszerhez képest a válaszidő mintegy ötszörös javulást mutatott, a 7. fejezet kezdetén használt rendszerhez képest pedig mintegy húszszorosára gyorsult a válaszidő. A jelen fejezetben bemutatott módosítások bevezetésével a rendszer platform függetlenségét a hullámforma lejátszás miatt elveszítettem.

[15] Jelentősen befolyásolná a gépi beszéd minőségének megítélését, amennyiben a lejátszás szaggatottá válna. A 7.5.2. fejezetben ismertetett eljárás bevezetése után nem tapasztaltam ilyen problémát. Amennyiben ez a későbbiekben mégis jelentkezne, a párhuzamosított eljárást javasolt kibővíteni a késleltetési csúcsokat kezelő feltételekkel Ramjee és munkatársai 4-es számú algoritmusa alapján [106].



30. ábra. A paramétergenerálás, beszédkódolás és hullámforma lejátszás az aktuális terhelés figyelembevételével való párhuzamosításának sematikus blokkdiagramja.



31. ábra. A minőséget nem érintő lépések során elért javulása a válaszidőnek.

7.6. Összegzés

Az ebben a fejezetben bemutatott kutatásom arra irányult, hogy minél kisebb számítási kapacitás árán lehessen gépi beszédet korlátozott erőforrású eszközökön előállítani (a kapcsolódó téziseim a következő fejezetben ismertetem). A kutatást két részre osztottam: a gépi beszéd minőségét érintő, és a gépi beszéd minőségétől független lépésekre.

Az első esetben inkrementális módon haladtam, és minden lépés esetén meghallgatásos tesztekkel ellenőriztem, hogy a gépi beszéd minőségét érintő módosítás okozott-e szignifikáns minőségromlást. Amennyiben nem okozott, és sebességnövekedést volt megfigyelhető, akkor az így létrejött rendszerrel folytattam a kutatást. A minőséget érintő módosítások a következőek voltak: spektrális modellezés módosítása, zaj modell módosítása, spektrális felbontás rendjének módosítása és a döntési fák méretének korlátozása. A kutatás eredményeként előállt egy, az eredeti rendszer minőségétől szignifikánsan nem különböző, mintegy ötször gyorsabb HMM-TTS (III.1. tézis: 7.4. fejezet).

A második esetben a paraméterfolyamok idő-rekurzív módon történő generálásával párhuzamosítottam a beszédkódoló eljárást és a hullámforma lejátszást. A hálózaton keresztül történő hang lejátszás analógiájára eljárást dolgoztam ki változó terhelés melletti HMM-TTS működésre mobil eszközökön. Az így létrehozott eljárás válaszideje a gépi beszéd minőségét érintő módszerekkel együtt a kiindulási rendszer válaszidejéhez képest mintegy húszszoros javulást mutatott (III.2. tézis: 7.5. fejezet).

Az eredmények alapján napjaink mobil eszközein is alacsony reakcióidővel futtathatóvá vált a rejtett Markov-modell alapú szövegfelolvasó. A megoldás nyelvfüggetlen, így mind magyar, mind más nyelven is hasonló teljesítménybeli növekedés várható.

8. Összefoglalás és tézisek

Doktori értekezésemben a statisztikai parametrikus beszédszintézis, ezen belül a rejtett Markov-modell alapú gépi szövegfelolvasás területén elért főbb tudományos eredményeimet mutattam be. Dolgozatomban három fő részre osztottam a témakörhöz kapcsolódó kutatói munkámat.

Az *első részben* (5. fejezet) a rejtett Markov-modell alapú szövegfelolvasás magyar nyelvre történő bevezetését ismertettem. A gépi beszéd minőségét környezetfüggő jegyeken alapuló eljárással növeltem. Eljárást dolgoztam ki a magyar nyelvű beszélőadaptációra, és megmutattam, hogy ezáltal készíthető a beszélőfüggő esethez képest szignifikánsan jobb minőségű gépi beszéd. Kísérleti úton igazoltam, hogy a nagy precizitású kézi címkézéssel nem feltétlen érhető el szignifikáns minőségjavulás.

Ezek az eredmények alapozták meg kutatómunkám *második részét* (6. fejezet), melyben a felügyelet nélküli beszélőadaptációval foglalkoztam. Célom olyan automatikus módszer kidolgozása volt, melynek segítségével manuális munka nélkül egy adott beszélőtől származó hanganyagból lehetséges új beszédhangok létrehozása. A felügyelet nélküli adaptációhoz az automatikus beszédfelismerő kimenetét használtam fel, és megmutattam, hogy jelentős fonémahiba-arány esetén is jól teljesít az így létrejött szövegfelolvasó rendszer. A megoldás hatékonyságát magasabb fonémahiba-arányok esetén a kényszerített illesztő beam paraméterének iteratív módosításával növeltem.

Értekezésem *harmadik fő kutatási* területe a rejtett Markov-modell alapú szövegfelolvasó korlátozott erőforrású eszközökre való illesztése (7. fejezet). Több lépésben csökkentettem a szövegfelolvasó számításigényét. A lépések egy része hatással volt a beszédminőségre, ezekben az esetekben szubjektív meghallgatásos tesztekkel figyeltem, hogy szignifikáns minőségromlást ne legyen. A kutatás eredményeképpen a rendszer válaszideje a kiindulási rendszerhez képest mintegy húszszoros javulást mutatott.

Dolgozatom tudományos eredményeit a következő három téziscsoportba rendezve tézisek formájában alább összegzem:

I. téziscsoport: Rejtett Markov-modell alapú szövegfelolvasó kidolgozása magyar nyelvre.

I.1. tézis: [J2, J3, J4, B2a, B3, C6, C7] *Kidolgoztam a magyar nyelv sajátosságainak megfelelő rejtett Markov-modell alapú általános szövegfelolvasó eljárást, és megmutattam, hogy a módszer a magyar nyelvre publikált legjobb minőségű, témaspecifikus szövegfelolvasónál szignifikánsan nem alacsonyabb minőségű beszédet képes előállítani, szignifikánsan kisebb adatbázis-méret mellett.*

I.2. tézis: [C2] *Megkülönböztető jegyeken alapuló eljárást dolgoztam ki rejtett Markov-modell alapú szövegfelolvasóhoz, és kimutattam, hogy ennek segítségével lehetséges javítani a gépi beszéd minőségét.*

I.3. tézis: [J2, B1, B2a, C6, C7] *Módszert dolgoztam ki magyar nyelvű felügyelt beszélőadaptációra rejtett Markov-modell alapú szövegfelolvasó rendszerben, amely az átlaghangból új hangkarakter létrehozásához a beszélőfüggő tanítás beszédkorpuszának kevesebb, mint 10%-át használja, és megmutattam, hogy segítségével előállítható a beszélőfüggő megoldásnál szignifikánsan jobb minőségű gépi beszéd.*

I.4. tézis: [B1] *Kimutattam, hogy a beszédkorpusz címkézés nagy pontosságú kézi javításának elhagyása nem okoz szükségszerűen szignifikáns minőségromlást beszélőfüggő és beszélőadaptált rejtett Markov-modell alapú szövegfelolvasók esetén.*

II. téziscsoport: Felügyelet nélküli, félspontán rejtett Markov-modell alapú szövegfelolvasó beszédhangjának adaptációja.

II.1. tézis: [C1, C5, C6] *Eljárást dolgoztam ki beszédfelismerő kimenetén alapuló felügyelet nélküli félspontán beszélőadaptációra rejtett Markov-modell alapú rendszerekben, és megmutattam, hogy segítségével lehetséges a felügyelt eset minőségétől szignifikánsan nem eltérő minőségű gépi beszéd előállítása.*

II.2. tézis: [C1, C3, C5] *Felügyelet nélküli eljárást dolgoztam ki egy adott beszélőtől származó hullámforma részhalmazának kiválasztására adaptációs beszédkorpusz kialakításának céljából, és megmutattam, hogy segítségével előállítható a véletlenszerű kiválasztásnál jobb minőségű gépi beszéd.*

III. téziscsoport: Rejtett Markov-modell alapú szövegfelolvasás illesztése korlátozott erőforrású eszközökre.

III.1. tézis: [J1, C4] *Modellt dolgoztam ki a rejtett Markov-modell alapú szövegfelolvasó korlátozott erőforrású eszközökön való megvalósítására, és kísérleti úton igazoltam, hogy segítségével a gépi beszéd előállítása szignifikáns minőségromlás nélkül szignifikánsan gyorsabb működésre képes.*

III.2. tézis: [J1, C4] *Eljárást dolgoztam ki a rejtett Markov-modell alapú szövegfelolvasó számításigényes folyamatainak (paramétergenerálás, beszédkódolás) a rendelkezésre álló erőforrások függvényében való párhuzamos működésére, és kísérleti úton megmutattam, hogy hatására azonos beszédminőség mellett a szövegfelolvasó válaszideje szignifikánsan javul.*

8.1. Az eredmények alkalmazhatósága

Az értekezésemben bemutatott új kutatási eredmények nemcsak elméletileg jelentenek új megközelítést, hanem a gyakorlatban való alkalmazhatóságuk is kézenfekvő. Ezt röviden a következőkben foglalom össze.

Az első téziscsoport I.1. és I.2. téziseiben ismertetett eljárással jó minőségű, tartalom független magyar nyelvű szövegfelolvasó hozható létre. Általános felhasználásával számos (pl. képernyő felolvasó vak felhasználók részére, IVR - Interactive Voice Response, prompt generátor és további gépi beszéd alapú asztali számítógépen futó) beszédtechnológiai alkalmazás valósítható meg. Az I.2. tézisben bemutatott megoldás nemcsak magyar, hanem más nyelvekre is kiterjeszthető. Az I.3. tézis eredményeire támaszkodva 10-15 percnyi felvétel alapján beszélőadaptáció segítségével a korábbiaknál jobb minőségű új beszédhangok hozhatók létre. Az I.4. tézis alapján pedig az új beszédhangok létrehozásához beszélőfüggő és beszélőadaptált esetben sem szükséges az automatikus címkézés kézi ellenőrzése. Ennek az eredménynek köszönhetően jelentős többletmunkát takaríthatnak meg a HMM-TTS alapú gépi beszéddel foglalkozó mérnökök. A bemutatott eredmények idegen nyelvekre is alkalmazhatóak.

A második téziscsoportomban bemutatott eredmények elsődleges újdonsága, hogy segítségükkel teljesen automatikusa módon új, adott célbeszélőkre jellemző hangkarakterisztikák létrehozása lehetséges. Ezzel a megoldással lehetőség van például telefonos adatbázisokból nagyszámú beszédhangot tartalmazó szövegfelolvasó rendszer automatikus létrehozása. Emellett arra is lehetőséget biztosít, hogy az adott HMM-TTS rendszert automatikusan a felhasználó hangjára lehet szabni. Például a mobil eszköz adott idő után emberi beavatkozás nélkül „megtanul" a tulajdonosa hangján beszélni. Tovább vizsgálandó kérdés, hogy a megoldás mennyire terjeszthető ki idegen nyelvekre (hipotézis: kiterjeszthető). A II.1. tézisben bemutatott eljárás jó minőségű beszédfelismerés esetén, míg a II.2. tézis rosszabb minőségű beszédfelismerés esetén nyújt megoldást felügyelet nélküli HMM-TTS beszélőadaptációra.

A harmadik téziscsoportom eredményeire támaszkodva az első téziscsoportban bemutatott szövegfelolvasó korlátozott erőforrású eszközökön is futtathatóvá válik, és a módszer a gépi hang megszólaltatása során figyelembe veszi az eszköz aktuális számítási terheltségét. A kutatást napjaink mobil eszközein végeztem. Az kutatás eredményeként létrejött rendszert a Google Android telefonok rendszer szintű TTS-eként lehet használni, amely szélesebb körű felhasználási lehetőségeket nyújt. Ilyen lehetőség például az SMS / E-mail felolvasás, e-könyv felolvasás, a hívó fél nevének felolvasása, navigáció során a hangos visszajelzés, mobil képernyő felolvasó vak és gyengénlátó felhasználók számára, stb. Az ebben a téziscsoportban bemutatott eredmények angol nyelvű HMM-TTS-el készültek, de mivel nem tartalmaznak nyelv specifikus részeket, így más nyelvekre is alkalmazhatóak. A kutatás részeként a magyar nyelvű változatot is elkészítettem. Továbbá a mobil HMM-TTS hangkarakterisztikáját az értekezésem korábbi részében bemutatott módszerekkel szintén lehetséges módosítani.

Mindegyik téziscsoportom eredményeit mintarendszerekben alkalmaztam.

Köszönetnyilvánítás

Köszönöm konzulenseim, Dr. Németh Géza és Dr. Olaszy Gábor nélkülözhetetlen segítségét és iránymutatását kutatói munkám során. Szakmai vezetésük mellett megismerkedtem a gépi beszédkeltés tudományával és inspiráló légkörben végezhettem a kutatást. Pozitív, emberközpontú szemléletüknek köszönhetően munkámat a BME-TMIT Beszédtechnológiai Laboratóriumban a munkatársaimmal baráti légkörben végezhettem.

Köszönöm a BME-TMIT Beszédtechnológiai Laboratórium munkatársainak, Bartalis Mátyásnak, Dr. Bőhm Tamásnak, Csapó Tamásnak, Dr. Zainkó Csabának a kutatás során az elméleti és gyakorlati segítségét.

Köszönöm Fegyó Tibor, Dr. Mihajlik Péter és Tarján Balázs közreműködését és segítségét, hogy a beszédfelismerés fontos részét képezhette kutatói munkámnak.

Köszönöm Dr. Siptár Péter fonológiai és Dr. Markó Alexandra nyelvészeti problémák terén nyújtott értékes segítségét és támogatását.

Köszönöm továbbá Dr. Henk Tamás tanszékvezető úrnak, hogy a doktori munkámat a vezetése alatti tanszéken végezhettem, valamint köszönöm a disszertáció és tézisfüzet megírásának és a doktori eljárás menetének segítő felügyeletét.

Köszönöm Dr. Gordos Géza ösztönző tanácsait és támogatását doktori munkám során.

Köszönöm Dr. Takács Györgynek és Dr. Tóth Lászlónak, hogy értékes észrevételeikkel segítettek a téma néhány fontos pontjának újragondolásában és az értekezés jobbá tételében.

Szeretném megköszönni családom doktori tanulmányaim során nyújtott példamutatását és segítségét. Külön köszönöm Édesapámnak, Dr. Tóth Pál Péternek a disszertációval kapcsolatos megjegyzéseit, észrevételeit, és Édesanyám, Dr. Gyires Klárának a tudományos kutatás általános kérdéseiben nyújtott segítségét. Köszönöm nővéremnek, Dr. Tóth Veronikának mindennapokban való önzetlen segítségnyújtását. Köszönöm kedvesemnek, Deák Robertának megértését és támogatását kutatói munkám során.

A doktori értekezést Nagyapám, Dr. Gyires Béla akadémikus emlékének ajánlom.

A kutatói munkámat a NAP (OMFB-00736/2005), a Teleauto (OM-00102/2007), a BelAmi (ALAP2-00004/2005), az ETOCOM (TÁMOP-4.2.2-08/1/KMR-2008-0007), a TÁMOP-4.2.1/B-09/1/KMR-2010-0002, a CESAR (No271022), az EITKIC_12-1-2012-0001 és a Paelife (Grant No AAL-08-1-2011-0001) projektek támogatták.

Irodalomjegyzék

[1] Mermelstein, P.: Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53 (4), 1070-1082 (1973)

[2] Kiss, G., Olaszy, G.: Interaktív beszédszintetizáló rendszer számítógéppel és OVE III beszédszintetizátorral. Magyar Fonetikai Füzetek 10., 21-45 (1982)

[3] Klatt, D. H., Klatt, L. C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America vol. 87., issue 2, 820-857 (1990)

[4] Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communications 9., 453–467 (1990)

[5] Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., Gordos, G.: Profivox – a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4., 201-215 (2000)

[6] Möbius, B.: Corpus-based speech synthesis: methods and challenges. Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition, 79–96 (2000)

[7] Németh, G., Olaszy, G., Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás 2006, 183-196 (2006)

[8] Németh, G., Olaszy, G., eds.: A magyar beszéd. Akadémiai Kiadó, Budapest (2010)

[9] Kempelen, F.: Az emberi beszéd mechanizmusa, valamint a szerző beszélőgépének leírása. Szépirodalmi Könyvkiadó, Budapest (1989)

[10] Dudley, H. W.: Remaking speech. Journal of the Acoustical Society of America 17, 1969-1977 (1939)

[11] Gold, B., Rabiner, L. R.: Analysis of Digital and Analog Formant Synthesizers. IEEE Transactions on Audio and Electroacoustics 16(1), 81-94 (1968)

[12] Gordos, G., Takács, Gy.: Digitális beszédfeldolgozás. Műszaki Könyvkiadó, Budapest (1983)

[13] Lei, M., Yamagishi, J., Richmond, K., Ling, Z., King, S., Dai, L.: Formant-controlled HMM-based Speech Synthesis. Proc. of Interspeech, 2777-2780 (2011)

[14] Remez, R. E., Rubin, P. E., Pisoni, D. B., Carrell, T. D.: Speech perception without traditional speech cues. Science 212, 947-950 (1981)

[15] Atal, B.: A new model of LPC excitation for producing natural-sounding speech at low bit rates. Proc. of ICASSP, 614-617 (1982)

[16] Zen, H., Tokuda, K., Black, A. W.: Statistical parametric speech synthesis. Speech Communication vol. 51, 1039-1064 (2009)

[17] Yamagishi, J., Nose, T., Zen, H., Toda, T., Tokuda, K.: Performance evaluation of the speakerindependent HMM-based speech synthesis system HTS-2007 for the Blizzard Challenge 2007. Proc. of ICASSP, 3957-3960 (2008)

[18] Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. Proc of ICASSP, 805-808 (2001)

[19] Ogata, K., Tachibana, M., Yamagishi, J., Kobayashi, T.: Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis. Proc. of ICSLP, 1328-1331 (2006)

[20] Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., Kitamura, T.: Speaker interpolation for HMM-based speech synthesis system. Journal of the Acoustical Society of Japan (E) vol. 21, issue 4., 199-206 (2000)

[21] Nose, T., Tachibana, M., Kobayashi, T.: HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation. IEICE Transactions on Information and Systems vol. E92-D, issue 3., 489-497 (2009)

[22] Tokuda, K., Zen, H., Black, A. W.: An HMM-based speech synthesis system applied to English. Proc. of IEEE SSW, 227-230 (2002)

[23] Drugman, T., Moinet, A., Dutoit, T., Wilfart, G.: Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. Proc. of ICASSP, 3793-3796 (2009)

[24] Chu, W. C.: Speech Coding Algorithms: Foundation and Evolution. Wiley-Interscience (2003)

[25] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Mixed Excitation for HMM-based Speech Synthesis. Proc. of Eurospeech, 2263-2266 (2001)

[26] Imai, S., Sumita, K., Furuichi, C.: Mel log spectrum approximation (MLSA) filter for speech synthesis. Transactions IEICE, vol. J66-A, 122-129 (1983)

[27] Yoshimura, T.: Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, Ph.D. thesis. (2002)

[28] Tokuda, K.: Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proc. of ICASSP, 229-232 (1999)

[29] Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE 77(2), 257-286 (1989)

[30] Zen, H., Masuko, T., Yoshimura, T., Tokuda, K., Kobayashi, T., Kitamura, T.: State duration modeling for HMM-based speech synthesis. IEICE Trans. on Inf. & Syst. E90-D(3), 692-693 (2007)

[31] Shinoda, K., Watanabe, T.: Acoustic modeling based on the MDL principle for speech recognition. Proc. of EUROSPEECH, 99-102 (1997)

[32] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K.: The HMM-based Speech Synthesis System (HTS) Version 2.0. Proc. of ISCA SSW6, 294-299 (2007)

[33] Yamagishi, J.: Average-Voice-Based Speech Synthesis, Ph.D. thesis., 178 (2006)

[34] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., Imai, S.: An Algorithm For Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features. Proc. of Eurospeech, 757-760 (1995)

[35] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. Proc. of ICASSP, 1315-1318 (2000)

[36] Brugnara, F., Falavigna, D., Omologo, M.: Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. Speech Communication 12(4), 357-370 (1993)

[37] Zen, H., Oura, K., Nose, T., Yamagishi, Y., Sako, S., Toda, T., Masuko, T., Black, A. W., Tokuda, K.: Recent development of the HMM-based speech synthesis system (HTS). Proc. of Asia-Pacific Signal and Information Processing Association, 121-130 (2009)

[38] Tokuda, K., Oura, K., Tamamori, A., Sako, S., Zen, H., Nose, T., Takahashi, T., Yamagishi, J., Nankaku, Y.: Speech Signal Processing Toolkit (SPTK), Version 3.5. (Accessed 2013) Available at: http://sp-tk.sourceforge.net/

[39] Sjolander, K.: The Snack Sound Toolkit. (Accessed 2013) Available at: http://www.speech.kth.se/snack/

[40] Kawahara, H., Masuda-Katsuse, I., Cheveign´e, A.: Restructuring speech representations

using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech Communication vol. 27, 187-207 (1999)

[41] Halácsy, P., Kornai, A., Oravecz, C.: HunPos - an open source trigram tagger. Proc. of 45th ACL, 209-212 (2007)

[42] Mihajlik, P., Fegyó, T., Tüske, Z., Ircing, P.: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages - like Hungarian. Proc. of Interspeech, 1497-1500 (2007)

[43] Boersma, P., Weenink, D.: Praat: doing phonetics by computer. (Accessed 2013) Available at: http://www.praat.org/

[44] Tokuda, K., Kobayashi, T., Masuko, T., Imai, S.: Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. Proc. of ICASSP, 1043-1046 (1994)

[45] ITU-T: P.800: Methods for subjective determination of transmission quality. (1996)

[46] Van Santen, J.: Perceptual experiments for diagnostic testing of text-to-speech systems. Computer Speech and Language, 49-100 (1993)

[47] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proc. of Eurospeech, 2347-2350 (1999)

[48] Bennett, C. L.: Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005. Proc. of Interspeech, 105-108 (2005)

[49] Bennett, C., Black, A. W.: Blizzard Challenge 2006. Proc. of Blizzard Challenge Workshop, Interspeech - ICSLP satellite event (2006)

[50] Krstulovic, S., Hunecke, A., Schröder, M.: An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. Proc. of Interspeech, 1897-1900 (2007)

[51] Gonzalvo, X., Iriondo, I., Socoro, A. F., Monzo, C.: HMM-based Spanish speech synthesis using CBR as F0 estimator. Proc. of NOLISP, 7–10 (2007)

[52] Martincic, S., Ipsic, I.: Croatian HMM-based Speech Synthesis. Journal of Computing and Information Technology 14(4), 307–313 (2006)

[53] Vicsi, K., Tóth, L., Kocsor, A., Gordos, G., Csirik, J.: MTBA - magyar nyelvű telefonbeszéd-adatbázis. Hiradastechnika Vol. LVII, NO.8, 35-43 (2002)

[54] Tóth, L., Kocsor, A.: Az MTBA magyar telefonbeszéd-adatbázis kézi feldolgozásának tapasztalatai. Beszédkutatás, 134-146 (2003)

[55] Zsigri, G., Tóth, L., Kocsor, A., Sejtes, G.: Az automata és kézi szegmentálás ejtésvariációk okozta problémái., 327-334 (2004)

[56] Gósy, M.: Fonetika, a beszéd tudománya. Osiris kiadó (2004)

[57] Akadémiai Kiadó: A magyar helyesírás szabályai 11th edn. Akadémiai Kiadó (1984)

[58] Durand, J., Siptár, P.: Bevezetés a fonológiába. Osiris Kiadó, Budapest (1997)

[59] Kiefer, F.: Strukturális magyar nyelvtan II. Fonológia. Akadémiai Kiadó, Budapest (1994)

[60] Siptár, P., Törkenczy, M.: The phonology of Hungarian. Oxford University Press, Oxford & New York (2000)

[61] Flynn, D.: Articulator Theory: An Introduction to Segmental Phonology. Textbook ms. (2006)

[62] Goldsmith, J. A.: The handbook of phonological theory. Wiley-Blackwell (1996)

[63] Menyhárt, K.: Zöngésedési és zöngétlenedési folyamatok a /j/ fonéma realizációiban.

Beszédkutatás 2003, 75-89 (2003)

[64] Isogai, J., Yamagishi, J., Kobayashi, T.: Model adaptation and adaptive train-ing using ESAT algorithm for HMM-based speech synthesis. Proc. of Interspeech, 2597–2600 (2005)

[65] Yi, J. R.-W.: Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis, Ph.D. thesis. Massachusetts Institute of Technology, Massachusetts, USA (2003)

[66] Wang, L., Zhao, Y., Chu, M., Zhou, J., Cao, Z.: Refining segmental boundaries for TTS database using fine contextual-dependent boundary models. Proc. of ICASSP, 641-644 (2004)

[67] Kang, S., Wu, Z., Cai, L., Shuang, Z., Qin, Y.: Comparison of Syllable/Phone HMM Based Mandarin TTS. Proc. of ICPR, 4496-4499 (2010)

[68] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (Version 3.4.). Cambridge University Engineering Department., Cambridge (2006)

[69] Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T.: Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing. IEICE Trans. Inf. Syst. E88-D(11), 2484-2491 (2005)

[70] King, S., Tokuda, K., Zen, H., Yamagashi, J.: Unsupervised adaptation for HMM-based speech synthesis. Proc. of Interspeech, 1869–1872 (2008)

[71] Gibson, M.: Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models. Proc. of Interspeech, 1791-1794 (2009)

[72] Gibson, M., Hirshimaki, T., Karhila, R., Kurimo, M., Byrne, W.: Unsupervised Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis Using Two-Pass Decision Tree Construction. Proc. of ICASSP, 4642-4645 (2010)

[73] Wu, Y., King, S., Tokuda, K.: Cross-lingual speaker adaptation for HMM-based speech synthesis. Proc. of ISCSLP, 1-4 (2008)

[74] Oura, K., Tokuda, K., Yamagishi, J., King, S., Wester, M.: Unsupervised Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis. Proc. of ICASSP, 4594-4597 (2010)

[75] Liang, H., Dines, J., Saheer, L.: A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-based Speech Synthesis. Proc. of ICASSP, 4598-4601 (2010)

[76] Wacha, I.: Az elhangzó beszéd főbb akusztikus stíluskategóriáiról., 203-216 (1974)

[77] Cosi, P., Falavigna, D., Omologo, M.: A preliminary statistical evaluation of manual and automatic segmentation discrepancies. Proc. of Eurospeech, 693-696 (1991)

[78] Leung, H. C., Zue, V. W. A.: Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech. Proc. of ICASSP, 271-274 (1984)

[79] Ljolje, A., Hirschberg, J., Van Santen, J. P. H.: Automatic Speech Segmentation for Concatenative Inventory Selection. In Van Santen, J. P. H., Sproat, R. W., Olive, J., Hirschberg, J., eds. : Progress in Speech Synthesis. Springer-Verlag, New York (1997) 305–312

[80] Van Santen, J. P. H., Sproat, R. W.: High-Accuracy Automatic Segmentation. Proc. of Eurospeech, 2809-2812 (1999)

[81] Zsigri, G., Kocsor, A., Tóth, L., Sejtes, G.: Phonetic Level Annotation and Segmentation of Hungarian Speech Databases., 659-673 (2004)

[82] Elenius, K., Takács, Gy.: Phoneme Recognition with an Artificial Neural Network. Proc. of Eurospeech, 121-124 (1991)

[83] Batliner, A., Kompe, R., Kiessling, A., Mast, M., Niemann, H., Nöth, E.: Syntax + Prosody:

A syntactic-prosodic labelling scheme for large spontaneous speech databases. Speech Communication 25, 193-222 (1998)

[84] Lehiste, I.: Perception of sentence and paragraph boundaries. Lindblom, B., Öhman, S., eds. : Frontiers of Speech Communication Research, 191-201 (1979)

[85] Schafer, A. J., Speer, S. R., Waren, P., White, S. D.: Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research 29, 169-182 (2000)

[86] Hird, K., Kirsner, K.: The relationship between prosody and breathing in spontaneous discourse. Brain and Language 80, 536-555 (2002)

[87] Gósy, M.: Virtuális mondatok a spontán beszédben. Beszédkutatás 2003, 19-43 (2003)

[88] Hillard, D., Ostendorf, M., Stolcke , A., Liu, Y., Shriberg, E.: Improving automatic sentence boundary detection with confusion networks. Proc. of HLT/NAACL, 69-72 (2004)

[89] Gillick, D.: Sentence boundary detection and the problem with the U.S. Proc. of HLT, 241-244 (2009)

[90] Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E., Stolcke, A.: A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer, Speech and Language 20(4), 468-494 (2006)

[91] Gotoh, Y., Renals, S.: Sentence Boundary Detection in Broadcast Speech Transcripts. Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium, 228-235 (2000)

[92] Hill, D. R., Manzara, L., Taube-Schock, C.-R.: Real-time articulatory speech-synthesis-by-rules. Proc. of AVIOS, 27-44 (1995)

[93] Black, A. W., Lenzo, K. A.: Flite: a small fast run-time synthesis engine. Proc. of 4th ISCA ETRW on Speech Synthesis, 157-162 (2001)

[94] Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., Raptis, S.: Embedded unit selection text-to-speech synthesis for mobile devices. IEEE Transactions on Consumer Electronics, vol. 55, issue:2, 613-621 (2009)

[95] Kim, S. J., Kim, J. J., Hahn, M. S.: HMM-based Korean speech synthesis system for hand-held devices. IEEE Transactions Consumer Electronics, vol. 52, issue 4., 1384-1390 (2006)

[96] Chinathimatmongkhon, N., Suchato, A., Punyabukkana, P.: Implementing Thai Text-to-Speech Synthesis for Hand-held Devices. Proc. of ECTI-CON, 545-548 (2008)

[97] Astrinaki, M., Alessandro, N., Picart, B., Drugman, T., Dutoit, T.: Reactive and Continuous Control of HMM-based Speech Synthesis. Proc. of SLT, 252-257 (2012)

[98] Oura, K., Zen, H., Nankaku, Y., Lee, A., Tokuda, K.: Tying covariance matrices to reduce the footprint of HMM-based speech synthesis system. Proc. of Interspeech, 1759-1762 (2009)

[99] Kominek, J., Black, A. W.: The CMU Arctic speech databases. Proc. of 5th ISCA Speech SynthesisWorkshop, 223-224 (2004)

[100] Jeruchim, M. C., Balaban, P., Shanmugan , K. S.: Simulation of Communication Systems: Modeling, Methodology and Techniques., 383-384 (2000)

[101] Chu, P. L.: Fast Gaussian Noise Generator. IEEE Transactions on Acoustics, Speech and Signal Processing 37(10), 1593-1596 (1989)

[102] Zen, H., Toda, T., Tokuda, K.: The Nitech-NAIST HMM-Based Speech Synthesis System for the Blizzard Challenge 2006. Journal IEICE - Transactions on Information and Systems E91-D(6), 1764-1773 (2008)

[103] Itakura, F.: Line spectrum representation of linear predictive coefficients of speech signals. Journal of the Acoustical Society of America 57(1), 35 (1975)

[104] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P.: Padé Approximants. Numerical Recipes: The Art of Scientific Computing, 245-247 (2007)

[105] Haykin, S.: Adaptive Filter Theory. N.J.: Prentice-Hall, Englewood Cliffs (1991)

[106] Ramjee, R., Kurose, J., Towsley, D., Schulzrinne, H.: Adaptive playout mechanisms for packetized audio applications in wide-area networks. Proc. of IEEE Infocomm, 680-688 (1994)

A szerző tudományos közleményei

A tézispontokhoz kapcsolódó tudományos közlemények

Folyóiratcikkek

[J1] Tóth, B., Németh, G.: Optimizing HMM Speech Synthesis for Low Resource Devices, Journal of Advanced Computational Intelligence & Intelligent Informatics, Vol. 16, No. 2., 327-334 (2012)
(BME-PA pontszám: 6, Scopus ID: 84859253582)

[J2] Tóth, B., Németh, G.: Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis, Acta Cybernetica, 19(4), 715-731 (2010)
(BME-PA pontszám: 4, Scopus ID: 78649885372)

[J3] Tóth, B., Németh, G.: Hidden Markov Model Based Speech Synthesis System in Hungarian, Infocommunications Journal, Volume LXIII, 2008/7, 30-34 (2008)
(BME-PA pontszám: 4)

[J4] Tóth, B., Németh, G.: Rejtett Markov-Modell Alapú Mesterséges Beszédkeltés Magyar Nyelven, Híradástechnika, Volume LXIII., 2-6 (2008)
(BME-PA pontszám: 2)

Cikkek szerkesztett könyvekben

[B1] Tóth, B., Németh, G., Olaszy G.: Beszédkorpusz tervezése magyar nyelvű, rejtett Markov-modell alapú szövegfelolvasóhoz, Gósy M.: Beszédkutatás 2012, MTA Nyelvtudományi Intézet, 278-295 (2012)
(BME-PA pontszám: 1)

[B2a] Tóth, B., Németh, G.: A rejtett Markov-modellen alapuló gépi szövegfelolvasás, Németh, G., Olaszy, G. (eds.), A magyar beszéd, 512-518 (2010)
(BME-PA pontszám: 3)

[B3] Tóth, B., Németh, G.: Rejtett Markov-modell alkalmazása magyar nyelvű gépi szövegfelolvasóhoz, Gósy, M.: Beszédkutatás 2008, MTA Nyelvtudományi Intézet, 182-193 (2008)
(BME-PA pontszám: 1)

Konferenciacikkek

[C1] Székely, É., Csapó, T-G., Tóth, B., Mihajlik, P., Carson-Berndsen J.: Synthesizing Expressive Speech from Amateur Audiobook Recordings, Proc. of IEEE Workshop on Spoken Language Technology, Miami, USA, 297-302 (2012)
(BME-PA pontszám: 0.6)

[C2] Tóth, B., Berki, S., Németh, G.: Distinctive Features in a Hungarian Hidden Markov Model Based TTS System, Proc. of 53rd International Symposium ELMAR-2011, Zadar, Croatia, 213-216 (2011)
(BME-PA pontszám: 1.5)

[C3] Tóth, B., Fegyó, T., Németh, G.: The Effects of Phoneme Errors in Speaker Adaptation for HMM Speech Synthesis, Proc. of 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, 2805-2808 (2011)
(BME-PA pontszám: 1.5)

[C4] Tóth, B., Németh, G.: Some Aspects of HMM Speech Synthesis Optimization on Mobile Devices, Proc. of 2nd International Conference on Cognitive Infocommunications, Budapest, Hungary, 1-5 (2011)
(BME-PA pontszám: 3)

[C5] Tóth, B., Fegyó, T., Németh, G.: Some Aspects of ASR Transcription based Unsupervised Speaker Adaptation for HMM Speech Synthesis, Proc. of 13th International Conference on Text, Speech and Dialogue (TSD), Brno, Czech Republic, 408-415 (2010)
(BME-PA pontszám: 1.5)

[C6] Tóth, B., Németh, G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel, Proc of. VI. Magyar Számítógépes és Nyelvészeti Konferencia (MSZNY), Szeged, Hungary, 246-256 (2009)
(BME-PA pontszám: 1)

Konferencia előadás kivonat

[C7] Tóth, B., Németh, G.: Hidden Markov Model Based Speaker Dependent and Adaptive Training of Hungarian Text-to-Speech System, Proc. of International Conference Probability and Statistics with Applications, Debrecen, Hungary, abstract (2009)

A szerző további tudományos közleményei

Cikkek szerkesztett könyvekben

[B2b] Tóth, B., Németh, G., Kiss, G.: Mobiltelefonba épített SMS felolvasó, Németh G., Olaszy G.: A magyar beszéd, 560-561 (2010)

[B2c] Viktóriusz, Á., Németh, G., Tóth, B.: NaviSpeech – beszélő navigátor látássérült gyalogosoknak, Németh G., Olaszy G.: A magyar beszéd, 591-595 (2010)

[B2d] Tóth, B., Németh, G.: Beszédkommunikátor beszédsérültek segítésére, Németh G., Olaszy G.: A magyar beszéd, 620-623 (2010)

[B4] Németh, G., Kiss, G., Zainkó, Cs., Olaszy, G., Tóth, B.: Speech Generation in Mobile Phones. In: Gardner-Bonneau, D., Blanchard, H. (eds.), Human Factors and Interactive Voice Response Systems, New York: Springer, 163-191 (2008)

[B5] Németh, G., Kiss, G., Tóth, B.: Cross Platform Solution of Communication and Voice/Graphical User Interface for Mobile Devices in Vehicles, In: Abut, H., Hansen, J. H. L., Takeda, K. (eds.), Advances for In-Vehicle and Mobile Systems: Challenges for International Standards, Springer, 237-250 (2007)

Konferenciacikkek

[C8] Tóth, B., Nagy, P., Németh, G.: New Features in the VoxAid Communication Aid for Speech Impaired People, Proc of. Computers Helping People with Special Needs: Lecture Notes in Computer Science. Linz, Ausztria, 295-302 (2012)

[C9] Németh, G., Csapó, T., Tóth, B.: Improving the Quality of Unit Selection and HMM based Speech Synthesis, Proc of. FuturICT, Budapest, Hungary (2009)

[C10] Tóth, B., Németh, G.: XML Based Multimodal Interfaces on Mobile Devices in an Ambient Assisted Living Scenario, Proc of. Workshop on Intelligent User Interfaces for Ambient Assisted Living, International Conference on Intelligent User Interfaces, Maspalomas, Gran Canaria, January 13-16 (2008)

[C11] Tóth, B., Németh, G.: Speech Enabled GPS Based Navigation System for Blind People on Symbian Based Mobile devices in Hungarian, Proc. of Regional Conference on Embedded and Ambient Systems, Budapest, Hungary, 69-74 (2007)

[C12] Tóth, B., Németh, G.: Challenges of Creating Multimodal Interfaces on Mobile Devices, Proc. of 49th International Symposium ELMAR-2007 focused on Mobile Multimedia, Zadar, Croatia, 171-174 (2007)

[C13] Tóth, B., Németh, G.: Creating XML Based Scalable Multimodal Interfaces for Mobile Devices, Proc. of 16th IST Mobile and Wireless Communications Summit, Budapest, Hungary, CD-ROM Proceedings (2007)

[C14] Németh, G., Kiss, G., Tóth, B.: Proposals for Extending the Speech Synthesis Markup Language (SSML) 1.0 from the Point-of-View of Hungarian TTS Developers, Proc. of W3C Second Workshop on Internationalizing SSML, Crete, Greece, (2006)

[C15] Tóth, B., Németh, G.: VoxAid 2006: Telephone Communication for Hearing and/or Vocally Impaired People, Proc. of 10th International Conference on Computer Helping People with Special Needs, Springer, Linz, Austria (2006)

[C16] Németh, G., Kiss, G., Tóth, B.: Cross Platform Solution of Communication and Voice / Graphical User Interface for Mobile Devices in Vehicles, Proc. of Biennial on DSP for in-Vehicle and Mobile Systems, Sesimbra, Portugal, CD-ROM Proceedings (2005)

Függelék: Rövid fogalommagyarázat a fontosabb kulcsszavakhoz

0G: 0-gram nyelvi modell. A beszédfelismerő nyelvi modelljében minden morféma egyszer, azonos valószínűséggel szerepel.

ASR: Automatikus beszédfelismerés (Automatic Speech Recognition)

Átlaghang: Ha a HMM-TTS rendszert sok beszélő személy hangjával tanítjuk, akkor a szintetizált beszéd valamennyi beszélő sajátosságait kisebb-nagyobb mértékben tartalmazza. Ebből kiindulva lehet egy további tanítással egy célbeszélő hangjához igazítani a gépi hangot.

BA-FF*x*: Beszélőadaptált tanítás, *x*-edik férfi beszélő hangjával.

Beam: A beam a kényszerített illesztés paramétere. A kényszerített illesztés alapját képző Viterbi algoritmus egy adott (nem végső) időpillanatához számos részleges felismerési hipotézis tartozik. A felismerési hipotézisek azért részlegesek, mert az algoritmus nem ért a jellemzővektor sorozat végéhez. Ezen részleges felismerési hipotézisekből igen számításigényes (és általában értelmetlen) volna mindet megtartani. Ezért vagy az adott időpillanatban a legjobbnak ítélt felismerési hipotézishez képest adott mértékben lemaradó felismerési hipotéziseket tartjuk meg a likelihood érték alapján, vagy pedig adott számú legjobb hipotézist. A gyakorlatban a kettő kombinációját is szokták használni. Keresési mélységnek, vagy beam szélességnek nevezzük a megtartott felismerési hipotézisek számát/mértékét.

BBS: A beam-alapú szelekció **(**Beam Based Selection). Ennek használatával a felügyelet nélküli beszélőadaptáció során az adaptációs beszédkorpusz a beam iteratív módon történő módosításával automatikusan áll elő.

Beszélőadaptált tanítás: A HMM-TTS-t először több beszélőtől (legalább 4-5) származó hosszabb (minimum 1-2 óra) beszédkorpuszokkal tanítjuk. Ennek hatására létrejön az átlaghang majd ezt egy célbeszélő hangkarakteréhez és beszédstílusához igazítjuk egy rövidebb (10-15 perces) adaptációs beszédkorpusszal.

Beszélőadaptáció: A beszélőadaptált tanítás azon része, amikor az átlaghangból előállítjuk a célbeszélőre jellemző gépi beszédhangot.

Beszélőfüggő tanítás: Egy beszélőtől származó, minimum 1-2 órás hanganyag alapján történő HMM-TTS gépi beszédhang létrehozása.

BF-FF*x*: Beszélőfüggő tanítás, *x*-edik férfi beszélő hangjával.

CD-HMM: Folytonos eloszlású rejtett Markov-modell (Continuous Density Hidden Markov Model).

CMOS teszt: Összehasonlításon alapuló szubjektív meghallgatásos teszt (Comparison Mean Opinion Score).

Döntési fák: A HMM-TTS rendszerekben a környezetfüggő címkék alapján döntési fák segítségével osztályokba soroljuk a paraméterfolyamokat.

Felügyelet nélküli beszélőadaptáció: Teljesen automatikusan, emberi beavatkozás nélkül történő beszélőadaptáció (Unsupervised Speaker Adaptation).

Felügyelt beszélőadaptáció: A beszélőadaptáció egyes lépései (pl. adaptációhoz felhasználandó beszédkorpusz előállítása) emberi segítséggel történnek (Supervised Speaker Adaptation).

Félspontán beszéd: Félspontánnak, vagy fél-reproduktívnak nevezzük azt a beszédtevékenységet, mely az élőszó igényével lép fel, de rendszerint az előadó által egy korábban megfogalmazott, elmondásra szánt írott szövegen alapszik.

FF*x*: *x*-edik férfi beszélő hangja.

FN: Felügyelet nélküli beszélőadaptáció.

FÜ: Felügyelt beszélőadaptáció.

HMM: Rejtett Markov-modell (Hidden Markov Model).

HMM-TTS: Rejtett Markov-modell alapú szövegfelolvasó (Hidden Markov Model based Text-To-Speech Synthesis).

Impulzus-zaj gerjesztés: A HMM-TTS tanítási és beszédelőállítási fázisában használt beszédkódoló eljárás gerjesztési modelljének egy fajtája, amely két típusú gerjesztési jelet használ: periodikus jelsorozatot, illetve zajt. Az impulzus-zaj gerjesztésű beszédkódoló eljárás az alábbi paraméterekkel működik: gerjesztés típusa, alapfrekvencia (zöngés gerjesztés esetén), spektrális paraméterek, időzítési paraméterek.

Kényszerített illesztés: Kényszerített illesztésnek (forced alignment) nevezzük azt a gépi eljárást, mely során a beszéd hullámformája alapján a fonetikus átirat egyes elemeihez (hangjaihoz) automatikus módon időzítési paramétereket (hanghatárokat) rendelünk. A magyar irodalomban kényszerített felismerésként is szoktak rá hivatkozni.

Kevert gerjesztés: A HMM-TTS tanítási és beszédelőállítási fázisában használt beszédkódoló eljárás gerjesztési modelljére vonatkozik. Ezzel már előállítható kevert gerjesztésű jel is (zönge és zaj egy időben és megfelelő arányossággal keverve). A kevert gerjesztésű beszédkódoló eljárás az alábbi paraméterekkel működik: alapfrekvencia, Fourier magnitúdók, zöngeerősségek sávonként, spektrális paraméterek, időzítési paraméterek.

Környezetfüggő címkék: Egy beszédhang leírására szolgáló, HMM-TTS rendszerekben alkalmazott jellemzés. Általában figyelembe veszi a hang környezetét (akár az egész mondat szintjéig is), továbbá számos más jellemzőt (hangsúlyok, szófajok, stb.).

LSP: Spektrumvonal páros (Line Spectral Pairs).

LVCSR: Nagy szótáras folyamatos beszédfelismerés / beszédfelismerő **(**Large-Vocabulary Continuous Speech Recognition**).**

Meghallgatásos teszt: A gépi szövegfelolvasók szubjektív minősítésére használt módszer. Kutatásom során MOS és CMOS alapú meghallgatásos teszteket végeztem.

Megkülönböztető jegyek: A megkülönböztető jegyek (Distinctive Features) segítségével minden egyes beszédhangot nyelvfüggetlenül, bináris és unáris értékek halmazával tudunk jellemezni. Fonetikai alapokra épül.

MGC: Mel-skála alapján általánosított kepsztrum (Mel Generalized Cepstrum). A kepsztrális és lineáris predikciós eljárás általánosított változata Mel-skála alapján**.**

MOS teszt: Szubjektív meghallgatásos teszt (Mean Opinion Score), ideális esetben a mintákat a tesztalanyok egymástól függetlenül értékelik. A teszt számos paraméterre vonatkozhat (minőség, természetesség, érthetőség, stb.).

MSD-HMM: Többterű eloszlású rejtett Markov-modell (Multi-Space probability Distribution Hidden Markov Model**).**

NO*x*: *x*-edik női beszélő hangja**.**

PER: Phone Error Rate, fonéma(tévesztési) hibaarány.

RND: Véletlen (Random) szelekció. A felügyelet nélküli beszélőadaptáció során az adaptációs beszédkorpuszt véletlenszerűen választottam ki (ehhez viszonyítva tudtam később a BBS eljárás hatékonyságát mérni**).**

Spektrumvonal páros: A spektrumvonal párost (Line Spectral Pairs, LSP) a lineáris predikciós együtthatók (Linear Prediction Coefficients, LPC) reprezentációjára használják. Az LPC-kkel szemben több előnyös tulajdonságuk miatt használatosak a beszédkódolásban (pl. kevésbé érzékenyek a kvantálási zajra).

Szegmentálás: A szegmentálás több jelentéssel is bír. Dolgozatomban a félspontán beszéd virtuális mondatokra való bontását jelöltem így.

Szelekció: Az adaptációhoz használt beszédkorpusz egy részhalmazának kiválasztását jelöli.

Szintézis: A gépi beszéd előállításának a folyamata a HMM adatbázisokból.

Tanítás: A HMM-ek tanítása, melynek eredményeként előáll a szintézis számára a HMM adatbázis.

WER: Word Error Rate, szó(tévesztési) hibaarány.

ZAJ: A maximális kivezérlés mondatonként 0 dB-re lett normalizálva, majd a teljes kivezérléshez képest -50 dB fehér zajt kevertem a jelhez.

ZAJ2: A maximális kivezérlés mondatonként 0 dB-re lett normalizálva, majd a teljes kivezérléshez képest -25 dB fehér zajt kevertem a jelhez.