Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
Villamosmérnöki Doktori Iskola

Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül

PhD tézisfüzet

Mihajlik Péter, MSc

Témavezető:
Gordos Géza, DSc
Konzulens:
Tatai Péter, MSc
Távközlési és Médiainformatikai Tanszék

Budapest, 2010.

1. Bevezetés

A gépi beszédfelismerés kutatása több évtizedes múltra tekint vissza nemzetközi és hazai viszonylatban is. Az első, gyakorlati feladatokra is használható módszer a dinamikus idővetemítés (Dynamic Time Warping) volt [Vintsjuk 68], [Myers & Rabiner 81], [Gordos & Takács, 83]. Ez a dinamikus programozáson [Bellman 57] alapuló eljárás elsősorban nyelvtől független kis szótáras, személyfüggő beszédfelismerésre használható. Lényege, hogy tárolt referenciamintákhoz hasonlítja a bejövő beszédjel lényegkiemelt változatát, és a legjobban illeszkedő referenciamintára, mint felismerési eredményre dönt. Jelentősen korlátozza a megközelítés gyakorlati alkalmazhatóságát, hogy a felhasználónak kell betanítania a rendszert a referencia felvételek egyenkénti bemondásával.

Számottevő előrelépést a rejtett Markov-modellek (HMM: Hidden Markov-Model) bevezetése hozott [Baker 75], [Jelinek & Bahl+ 75]. A gépi beszédfelismerés folyamata lényegét tekintve nem sokat változott – *lényegkiemelés* és *mintaillesztés*, azaz tárolt modellekkel történő összehasonlítás dinamikus programozással. A modell-struktúrák ugyanakkor jóval összetettebbé váltak, illetve a paraméterek jórészt statisztikai úton kerültek meghatározásra. A több száz-, ill. ezerbeszélős tanító-beszédadatbázisok révén lehetővé vált a személyfüggetlen beszédfelismerés.

A folyamatos gépi beszédfelismeréshez a következő fontos lépés az ún. nyelvi modellek alkalmazása, azok rejtett Markov-modellekbe való integrálása volt [Jelinek & Mercer 80]. Az elemi beszédegységeket, pl. szavakat, szótagokat, hangokat egyszerű rejtett Markov-modellel modellezik és ezeket az elemi modelleket egyetlen (szintén rejtett Markov-modell) felismerési hálózattá kapcsolják össze. A rejtett Markov-modellek állapotaihoz tartozó emittálási valószínűségek jól feleltethetők meg az akusztikai megfigyelési valószínűségeknek, az átmeneti valószínűségek pedig jól használhatók az egyes szókapcsolatok valószínűségeinek reprezentálására is, vagyis a nyelvi modellezésre. A felismerés eredménye itt a felismerési HMM hálózat kezdő és végpontja(i) közti legjobb illeszkedésű útvonal [Ney 84], melynek meghatározására a Viterbi-algoritmus [Bellman 57] jól használható.

A (folyamatos) gépi beszédfelismerés kiinduló egyenlete:

$$\hat{W} = \arg\max_{W} P(W \mid O) \tag{1}$$

ahol $W = w_1,...,w_K$, $K \in N$ egy megengedett (modellezett) szósorozatot jelöl, és $O = o_1,...,o_T$ a bejövő beszédjel T elemű lényegkiemelt vektorsorozatát jelöli. $\hat{W} = \hat{w}_1,...,\hat{w}_{\hat{K}}$, $\hat{K} \in N$ pedig a felismert szósorozatot jelenti.

Rejtett Markov-modellek esetén a fenti egyenlet a Bayes-szabály segítségével a következő alakra hozható:

$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W) \tag{2}$$

A (2) képletet a beszédfelismerés alapegyenletének nevezzük, mely szemléletesen választja szét az adott szósorozatnak a nyelv által megszabott valószínűségét, $P(W)$-t az akusztikai megfigyelés valószínűségétől, $P(O|W)$-től. Az akusztikus modell részének tekintjük a (magasszintű) kiejtési modellt, mely környezetfüggő vagy környezetfüggetlen beszédhangmodellre képzi le a lexikai egységeket (hagyományosan a szavakat).

Adott nyelvű beszédfelismerő rendszer elkészítése tehát alapvetően az akusztikus és nyelvi modellek meghatározásából áll. Noha a klasszikus, elterjedten alkalmazott módszerek jelentős mértékben statisztikaiak, számos nyelvspecifikus szabály, nyelvi szakértelem is szükséges az alkalmazásukhoz.

Különös esetet jelent a magyar nyelv, részint a ragozás, toldalékozás folytán adódó nagy szóalaki változatossága és relatíve kötetlen szórendje miatt, másrészről pedig a hangtani sajátosságai miatt (például diftongusok hiánya).

A spontán beszéd elsősorban akusztikai szempontból (laza artikuláció) másodsorban nyelvi szempontból (pl. agrammatikus mondatok) állítja komoly kihívás elé a beszédfelismerő rendszereket.

A kutatócsoportunktól független publikációk máig szinte csak környezetfüggetlen beszédhangmodellezést használnak magyar nyelvre (kivétel: [Czap 05]) ugyanakkor a fonológiai koartikuláció (hasonulási jelenségek) modellezését általában fontosnak tartják és szó lexikai modellekkel dolgoznak [Tóth 2009], [Szaszák 2008], [Bánhalmi & Paczolay+ 08], [Zsigri & Tóth+ 04], [Vicsi & Szaszák 04]. Spontán magyar nyelvű nagyszótáras folyamatos gépi beszédfelismerési témában más kutatócsoportok publikációi nem érhetők el. A magyar nyelvtől elvonatkoztatva, a morfológiai gazdagság kezelése számos toldalékoló, ragozó nyelvben (finn, török, észt, arab) morféma alapon sikeresen történik [Kurimo & Creutz+ 06], [Afify & Sarikaya+ 06], azonban spontán beszéd esetén a szó helyett morfémaszerű lexikai egységek használata negatív eredménnyel járt [Creutz & Hirsimäki+ 07]. A nyelvi szabályoktól mentes ún. grafféma alapú akusztikus modellek [Kanthak & Ney 02], [Killer & Stüker+ 03] számos nyelv esetén bizonyultak versenyképesnek, de a (statisztikai) morfémákkal együtt történő alkalmazásuk általában ad-hoc, a szó-fonéma alapú klasszikus rendszerekkel készült összehasonlító elemzésről nem találtunk referenciát.

A következőkben a spontán, személyfüggetlen, nagyszótáras magyar nyelvű gépi beszédfelismerés irányában kitűzött céljaimat, az alkalmazott módszertant és az új kutatási eredményeimet mutatom be. A kutatásom kezdetén felállított hipotézis ellenkezőjére jutottam: mélyebb nyelvspecifikus tudás, szabályok nélkül is elérhető az előzőeket alkalmazóval szemben versenyképes beszédfelismerési technológia.

2. Kutatási célkitűzések

Általános célom a magyar nyelvű beszéd minél pontosabb, de kézben tartható számításigényű gépi felismerése[1]. Értekezésemben az elsődleges cél a *magyar nyelvi jellegzetességekkel kapcsolatos modellezési kérdések* megválaszolása.

[1] Gépi beszédfelismerés alatt az általános beszéd-szöveg átalakítást értjük, amely a nagyszótáras folyamatos spontán nyelvű beszéd szöveggé alakítását is magában foglalja.

Konkrét, *tézisekben is megjelenő célkitűzéseim* a magyar nyelvű gépi beszédfelismerés témakörében a következők voltak:

I. A magyar nyelvű *fonetikai koartikuláció* modellezésének vizsgálata az általános gépi beszédfelismerés szempontjából. Azaz, a tárgy a beszédhangok egymásra hatásának vizsgálata, a modellezés miként.
II. A magyar nyelvre jellemző *fonológiai koartikuláció* modellezésének vizsgálata általános gépi beszédfelismeréshez. Másképpen fogalmazva a hasonulási, egybeolvadási és egyes hangkiejtési szabályok alkalmazásának módja, szükségessége merül fel kérdésként.
III. A magyar nyelv *lexikai modellezésének* vizsgálata spontán nyelvű beszéd felismeréshez. Itt a nyelvünk morfológiai gazdagsága okozta kihívásokra (nagyszámú szóalak, ritka szóalakok nagy mennyisége, szótáron kívüli szavak magas aránya) adható válasz keresése a feladat alkalmas lexikai egységek megválasztásával (nyelvi, statisztikai morfémaszerű egységek).
IV. A magyar nyelv *kiejtés-modellezésének,* a kiejtett alak automatikus előállításának vizsgálata spontán beszéd felismeréséhez. Azaz, hogy az esetlegesen többféle ejtésmódú vagy kivételes ejtésű szavak miként modellezhetők, a fonologikus átiratkészítés hogyan automatizálható.

Ezeken felül természetesen a nyelvi modellezés is fontos – látszólag erősen nyelvfüggő – feladat, azonban a megfelelő lexikai egységek kiválasztása után a standard N-gram modelleken túlmutató megközelítés kidolgozása nem látszott feltétlenül szükségesnek. Hasonlóan, a fizikai szintű akusztikai modellezést sem tartom (a tonális nyelvektől eltekintve) nyelvspecifikusnak, ezért annak részleteivel az értekezés keretei között nem foglalkozom.

3. Módszertan

A dolgozatomban bemutatott kutatások során a beszédtechnológiában és a kapcsolódó tudományágakban elterjedt módszerekkel dolgoztam.

A következőkben röviden ismertetem a felhasznált beszédadatbázisokat, a felismerési feladatokat, körülményeket és az eredmények értékelésének módjait.

3.1. Beszédadatbázisok

A célkitűzéseknél említett vizsgálatokat a kutatások időpontjában elérhető legnagyobb és legismertebb magyar nyelvű beszédadatbázisokon végeztem.

Általános (fonetikai és fonológiai koartikulációs) vizsgálatokra az MTBA [Vicsi & Tóth 02], a Besztel, a SpeechDat és a Tesztel [Vicsi et al.] összességéből alakítottam ki tanító- és teszthalmazokat, illetve definiáltam rajtuk beszédfelismerési feladatokat. Ezek az adatbázisok elsősorban olvasott beszédet, úgymint szépirodalomból kiemelt fonetikailag változatos mondatokat, szavakat, valamint kisebb arányban spontán bemondásokat is tartalmaznak, tipikusan rövid vezérlő szavakat. Az anyag telefonon (mobil, vezetékes vegyesen) lett rögzítve mintegy 1600 beszélőtől, több mint 50 óra hosszan.

Spontán beszédet közvetlenül célzó (lexikai és kiejtés-modellezési) vizsgálatok esetén a magyar MALACH (Multilingual Access to Large Spoken Archives) [J1], [B1] adatbázis már lejegyzett részét használtam, mely nagy sávszélességű, mikrofonnal, otthoni környezetben rögzített beszélgetéseket tartalmaz. Tematika: második világháborús visszaemlékezések – beszélgetés a kérdező és a visszaemlékező között. A beszélők döntően idős, nem anyanyelvi környezetben élő emberek így a beszédjük gyakran megakadásokkal, idegen szavakkal tarkított, ugyanakkor esetenként kifejezetten szépen formált. Az adatbázis 34 órányi beszédet tartalmaz 114 beszélőtől.

3.2. Beszédfelismerési feladatok

Alapvetően folyamatos, beszélőfüggetlen, nagyszótáras beszédfelismerési teszteket végeztem, mivel az általános tapasztalatok szerint ezek jelentik a legnagyobb kihívást a beszédfelismerő rendszerek számára. Az alkalmazásorientált tesztelés érdekében azonban a telefonos adatbázison – ellenőrző jelleggel – izolált szavas, beszélőfüggetlen felismerési teszteket is végrehajtottam.

3.3. Tanító- és teszthalmazok

A *tanítóhalmazok* a beszélőfüggetlenség érdekében az adatbázisok nagyobb részéből állnak. A telefonos adatbázisok esetén 500-900-1300, a szélessávú adatbázis esetén 104 beszélő hanganyagát használtam. A telefonos adatoknál a tanító adatbázis méretének hatását is vizsgálandó négyféle tanítóhalmazt alakítottam ki. A legkisebb az MTBA-nak csak a kézzel szegmentált részét tartalmazza (~3 óra, 6000 felvétel). A következő az MTBA szinte teljes egészét tartalmazza (19 000 felvétel). A harmadik halmaz az előzőn felül a Besztel adatbázis hasonló struktúrájú adataiból az első 400 beszélőéit foglalja magában (39 000 felvétel). Végül, a legnagyobb tanítóhalmaz az előzőn felül a SpeechDat adatbázis általunk elérhető első 400 beszélőjének felvételeit tartalmazza, összesen mintegy 44 000 felvétel, 30 óra terjedelemben. A spontán adatbázisnál egy tanítóhalmaz volt definiálva, mely 26 órányi kézi erővel átírt hanganyagot jelentett.

A *teszthalmazok* kialakításánál a beszélőfüggetlenség alapvető feltétel volt, azaz csak olyan beszélőtől származó hanganyagok vannak bennük, melyek a tanító halmazokban nem szerepeltek (220 beszélő, 2400 bemondás a telefonos adatbázisoknál, 10 beszélő, 8 óra felvétel a MALACH adatbázis esetén). Továbbá arra is törekedtem, hogy a teszthalmazokat szövegtartalom szempontjából a tanítóhalmazhoz jobban, illetve kevésbé illeszkedő, diszjunkt részhalmazokra bontsam. A telefonos adatok esetén az illeszkedés mértéke a jobban *illeszkedő teszthalmaznál* jelentős (a tesztmondatok, szavak legtöbbje szövegszerűen a tanítóhalmazban is szerepel), a *nem illeszkedőnél* elhanyagolható. A szélessávú spontán adatbázisnál az *illeszkedő* és *kevésbé illeszkedő* halmazok közti különbség csak annyiban áll, hogy az utóbbiak a beszélgetések azon kezdeti szakaszából származnak (első negyedóra), mely szakaszt a tanító adatbázisból kihagytuk. A tézisfüzetben az utóbbi teszthalmaz-megkülönböztetéseket a könnyebb áttekinthetőség kedvéért nem tüntettem fel (a disszertációban igen).

3.4. Kísérleti konfigurációk

Kísérletek a telefonos beszédadatokon: A telefonos adatoknál minden tanítóhalmazzal külön-külön tanítottam az elemi akusztikus modelleket. Ezeket használtam mind folyamatos mind izolált szavas tesztelésnél. Akár a folyamatos, akár az izolált szavas

teszteknél a teszthalmazt két részre bontva (illeszkedő, nem illeszkedő) értékeltem ki a felismerési eredményeket.

Kísérletek a spontán, szélessávú beszédadatokon: A MALACH felismerési feladatnál a tanító adatbázis adott volt, és csak folyamatos nagyszótáras felismerést végeztem, egyéb tekintetben a fentiekhez mindenben hasonló beszélőfüggetlen kísérleteket folytattam.

3.5. Beszédfelismerési paraméterek, beállítások

A beszédfelismerési kísérletekben a következőkben részletezett alapbeállításokat használtam (a későbbiekben az ezektől való eltérést minden esetben közlöm).

Lényegkiemelés: Lényegkiemelési paraméterekként a bemenő beszédjelből a telefonos adatbázisnál 12 MFCC (Mel Frequency Cepstral Coefficients) [Mermelstein 76], a szélessávú adatbázisnál 17 PLP (Perceptual Linear Prediction) [Hermansky 90] elemű lényegvektorokat képeztem, melyekhez *logE* (keretenkénti logaritmikus energia) paramétert is csatoltam, majd dinamikus Delta és Delta-Delta értékeket számítottam (+-2 keretes időablakban számított lineáris regresszióval). A statikus energiát az előbbi esetben kicsatolva összesen 38 ill. 54 dimenziós jellemzővektorok keletkeztek. A telefonos adatoknál mind a tanítás, mind a tesztelés során alkalmaztam a vak csatornakiegyenlítés (Blind Equalization) módszerét [Mauuary 98], [C15], a szélessávú adatoknál pedig a kepsztrumátlag-kivonást.

Elemi akusztikus modellek: Az atomi modellek rejtett Markov-modell állapotok voltak rögzített hurok és továbblépési valószínűségekkel. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket [Titterington & Smith+ 85] használtam. A tanításnál a paraméterek becslése ML (Maximum Likelihood) alapú volt [Dempster & Laird+ 77]. (Inicializálás, Viterbi tanítás, majd Baum-Welch iteratív újrabecslés, ill. Gauss komponens növelés „mixture splitting” eljárással [Young 06].)

Fonetikai koartikulációs modellek: Mind a monofón mind a trifón modelleknél (utóbbi az alapértelmezés) a beszédhangokat 3 elemi akusztikus modellre képeztem le. Az előbbi esetben a környezettől függetlenül, az utóbbi esetben ML döntési fa alapján a fonetikus környezettől függően [Young 06]. A döntési fa építéshez mintegy 50-féle, alapvetően nyelvészeti ihletésű fonetikai kategóriát használtam. A trifón állapotcsoportosításokat tanítóhalmazonként külön-külön végeztem.

Környezetfüggőségi modell: szóhatárokon átívelő („cross-word”) trifón modellezés.

Fonológiai koartikulációs modell: explicit modellt nem alkalmaztam, azaz a fonológiai koartikulációkat nem jelölő implicit modellt használtam.

Fonológiai kiejtési modellek: A telefonos adatok esetén a kiejtési modellek lexikai, azaz fonológiai koartikulációkat nem tartalmazó fonemikus átiratait automatikusan, de kézi graféma-fonéma szabályok segítségével állítottam elő [J6]. A spontán beszédadatbázisnál a kézi lejegyzés során kivételes kiejtésűnek jelölt szavak ilyenkor megadott fonológiai kiejtési változatát használtam, gyakoriság szerint súlyozva az esetleges ejtési variációkat. A többi, nem kivételes kiejtésű szó fonológiai átiratát automatikus módszerrel generáltam. Allofónikus változatokat sehol nem jelöltem,

továbbá a hosszú és rövid mássalhangzókat sem különböztettem meg. Így – a szünetmodelleket nem számítva – összesen 39 fonológiai kategóriát használtam. A szünetmodell háromállapotú környezetfüggetlen modell volt.

Szótár (lexikai modell): Alapértelmezésben szavak szerepeltek a szótárban (a telefonos adatbázisnál kiejtési variációk nélkül, a MALACH adatbázisban kézi kivételekkel és kiejtési variációkkal). Az előbbi esetben az izolált szavas felismeréseknél ugyanazt az 1334 elemű szótárat (lexikont) használtam az illeszkedő és nem illeszkedő felvételek esetén is. Hasonlóan, a telefonos folyamatos felismeréseknél is ugyanazt az 5561 elemű szótárat és természetesen ugyanazt a nyelvi modellt alkalmaztam mindkét teszthalmaz esetén. Mind a folyamatos mind az izolált szavas telefonos tesztek esetén a teljes teszthalmazt lefedő szótárakat alkalmaztam, így szótáron kívüli elemek kezelésére nem volt szükség. A MALACH adatbázisnál a szótárméret 20 000 volt, és az OOV (Out Of Vocabulary) arány mindkét teszthalmaz esetén 15% körülinek adódott.

Nyelvi modell: N-gram nyelvi modelleket alkalmaztam. Telefonos adatoknál 3-gram modelleket Katz-féle visszametszéssel [Katz 87] és Good-Turing valószínűség-újraelosztással [Good 53]. A tanítószöveg az illeszkedő tesztmondatok szövege alapján készült úgy, hogy minden különböző mondatot csak *egyszer* szerepeltettem. Így az illeszkedő mondatokon PP=40-es perplexitást [Bahl & Jelinek+ 83], a nem illeszkedő tesztmondatokon PP=6230-as (nagyon magas, azaz igen kedvezőtlen) perplexitás értéket kaptam. A MALACH adatbázisnál az akusztikus modelltanításnál használt felvételek szövegátirataival tanítottam, módosított, interpolált Kneser-Ney simítási eljárást alkalmazva [Chen & Goodman 98]. A szó vagy szótöredék (morf) N-gram modellek fokszáma kísérletenként volt optimalizálva (szónál N=3, morfnál N=4). A teljes teszthalmazon szóalapon így PP=336 értéket mértem. A nyelvi modellezésre az SRILM eszközt alkalmaztam [Stolcke 02]

Felismerési hálózatépítés: A felismerési hálózatok építése (az előzőekben felsorolt tudásforrások integrációja) és optimalizációja a WFST (Weighted Finite State Transducer) keretrendszerben történt az AT&T FSM Toolkit segítségével [Mohri & Pereira+ 02]. Az optimalizáció a fonémaszintű integrált felismerési hálózat determinizációja réven jött létre.

Dekódolás: A beszédfelismerési kísérletek mindegyikét ugyanazon a 3GHz Pentium IV, 2GB operatív memóriájú személyi számítógépen végeztem. Az optimalizált, dinamikus programozáson alapuló dekódolás a VOXerver nevű eszközzel történt. Minden kísérletet olyan keresési mélység mellett végeztem, ahol a felismerési pontosság már erősen telítési szakaszban volt. Az I. és II. téziscsoport esetén a keresési mélység fix, míg a további kísérletekben közel azonos futási idő mellett vetettem össze a felismerési eredményeket.

3.6. A felismerési eredmények kiértékelése

A gépi felismerés pontosságának mérése mindig kézi referencia átiratokhoz viszonyítva történt. A következő metrikákat és szignifikancia-vizsgálati módszereket alkalmaztam.

Metrikák: A felismerési eredményt – mely felismerési egységek (pl. szavak, betűk) sorozata – a referencia átirathoz dinamikus programozás módszerével hasonlítjuk, ahol a következő súlyokat rendeljük az egyes lehetőségekhez:

C (helyes, „korrekt” felismerés): 0
S (helyettesítés, „szubsztitúció”): 10
D (törlés, „deletálás”): 7
I (beszúrás, „inzerció”): 7

A kiértékelés alapja a legkisebb összsúlyú összerendelés. A fenti betűjelekkel az adott jelenségek darabszámát jelölve, az alábbi felismerési mérőszámok definiálhatók:

$$Felismerési\ pontosság\ (Accuracy:"Acc") = \frac{N-S-D-I}{N} \times 100\%, \qquad (3)$$

ahol *N* az összes felismerési egység (pl. szó) száma a referencia átiratban.

A felismerési hiba definíciója:

$$Felismerési\ hiba\ (Error\ Rate:"ER") = \frac{S+D+I}{N} \times 100\% \qquad (4)$$

A két alapvető metrika, amely alapján az egyes eredményeket összehasonlítom, a következő:

- WER (Word Error Rate): szófelismerési hiba, a felismerési egységek a szavak (összetett szó is egy egység). A legáltalánosabban elterjedt mérték.
- LER (Letter Error Rate): betűfelismerési hiba. A morfológiailag gazdag nyelvek esetén megbízhatóbb mérőszámként szolgál, mint a WER. Továbbá a kézi javítás „költségével” az előzőnél jobban korreláló mennyiség. A szóközt is betű értékűnek definiáljuk, egyébként ugyanúgy számoljuk ki karakter egységenként, mint a szóhibaarányt.

A gyakorlatban azonban általában a javulás relatív mértéke az érdeklődés tárgya. Ezt az alábbiak szerint definiáljuk mind WER, mind LER esetén.

$$Relatív\ javulás\ (-\Delta ER_{rel}) = \frac{ER_{referencia} - ER_{új}}{ER_{referencia}} \times 100\% \qquad (5)$$

Végül, gyakorlati szempontból igen lényeges metrika lehet a felismerés időigényének az alakulása is. Erre az RTF (Real Time Factor) a szokásos mérték.

$$RTF = \frac{felismerésre\ fordított\ idő}{felismert\ beszéd\ hossza} \qquad (6)$$

Tehát az alacsonyabb értékek a jobbak.

Szignifikancia-vizsgálatok: Statisztikai hipotézis vizsgálatokkal ellenőriztem a javulások megbízhatóságát. A NIST ajánlás szerinti nem parametrikus Wilcoxon-féle előjeles rangtesztet [Kanji 94], [Daniel 78] alkalmaztam.

Független eseményeknek a telefonos adatbázisnál az egyes bemondásokat tekintettem, míg az MALACH adatbázisnál az annotátorok által megjelölt – olykor több száz szót tartalmazó – nagyobb közlési egységeket. Az eseményekhez tartozó valószínűségi változóknak pedig a lokális – adott adatbázisrészletre számolt – WER és LER értékeket.

Az kiértékeléseknél a $p=0.05$ szignifikancia szintet (0.95-ös konfidencia szintet) használtam. A szignifikáns *javulások* dőlt betűvel jelennek meg a táblázatokban.

4. Új eredmények

4.1. I. téziscsoport: A fonetikai koartikuláció modellezése magyar nyelvű beszéd gépi felismeréséhez

Az egymást követő beszédhangok egymásra hatása, egymás képzésének befolyásolása a beszéd alapvető jellegzetessége. A gépi beszédfelismerés ezt a jelenséget a statisztikai elvű akusztikai modellezés révén – alapszinten – figyelembe is veszi. Fontos kérdés azonban, hogy a magyar nyelvre elterjedt *implicit* modellezéshez képest – melynél a beszédhangokat fonetikai környezetüktől függetlenül jellemezzük [Tóth 09], [Szaszák 08], [Bánhalmi & Paczolay+ 07], [Tóth 06], [Vicsi & Velkei+ 05] – milyen mértékben javíthat az explicitebb, környezetfüggő beszédhangmodellezés. A magyar nyelvet illetőn korábban nem volt ismeretes olyan átfogó tanulmány, mely a fonetikai koartikuláció modellezését e tekintetben vizsgálta volna, inkább kezdeti eredmények születtek kisebb, egyedi adatbázisokon, szignifikancia-vizsgálat nélkül [Szarvas 03], [Czap 05].

Kutatásaimban arra kerestem a választ, hogy a vajon a magyar nyelvi környezetnek van-e olyan megkülönböztető sajátossága (pl. diftongusok hiánya, megfelelő méretű adatbázisok hiánya), ami miatt a környezetfüggetlen, azaz implicit koartikuláció modellezés adekvát, avagy a környezetfüggő beszédhangmodellezés szignifikáns felismerési hibacsökkenéssel is járhat.

A környezetfüggő beszédhangmodellezés alapproblémája, hogy egyrészről szükséges lenne megkülönböztetni minden olyan környezetet, mely különböző képzési módot, hangátmenetet eredményez, másrészről viszont korlátozni kell a modell komplexitást, hogy minden elemi modellre elég tanítóadat jusson.

Az első tézisben egy, a kutatócsoportunk által kidolgozott környezetfüggő beszédhangmodellezési módszer eredményei alapján vonok le következtetéseket. A módszer lényege, hogy a környezeteket, azok struktúráját szakértői szabályok alapján alakítja ki, ám ha egy adott környezetfüggő beszédhangmodell részletre (HMM állapotra) kevés tanítóminta jutna, akkor addig csökkenti a környezetfüggés mértékét (jobb, bal vagy mindkét oldali környezet elhagyása), míg elegendő tanítóminta nem áll rendelkezésre. A technikát visszametszéses („backoff”) trifón állapotcsoportosítási technikának neveztük el, mivel hasonlóan elven működik, mint a „backoff” nyelvi modell simítás.

I.1. tézis: [B2, B3, C7, C8] *Kísérleti úton megmutattam, hogy visszametszéses trifón állapotcsoportosítású környezetfüggő beszédhangmodellekkel elérhető szignifikáns beszédfelismerési pontosságjavulás magyar nyelven a környezetfüggetlen beszédhangmodellezéssel elért eredményekhez képest.*

Az 1. táblázat mutatja az összesített beszédfelismerési eredményeket, melyeket több mint 1500 beszélős telefonos beszédadatbázisok felhasználásával mértem.

1. Táblázat
Telefonos, környezetfüggetlen és visszametszéses trifón állapot-csoportosítású folyamatos beszédfelismerési eredmények összefoglalása

Teszthalmaz	Átlagos szófelismerési pontosság [%]		Átlagos relatív hibacsökkenés [%]
	Referencia: környezetfüggetlen beszédhangmodell	Visszametszéses trifón állapotcsoportosítású beszédhangmodell	
Izolált szavas felismerés			
Illeszkedő	85.7	95.0	*65*
Nem illeszkedő	82.7	91.5	*51*
Folyamatos felismerés			
Illeszkedő	80.3	90.8	*53*
Nem illeszkedő	20.5	41.6	*26*

Látható, hogy egy egyszerűen implementálható környezetfüggő beszédhangmodellezési technika is drasztikus javulásokat hozott. Izolált szavas felismerés és illeszkedő folyamatos beszédfelismerési tesztek esetén felezte, harmadolta a szóhibaarányt, míg a nem illeszkedő folyamatos tesztnél megkétszerezte a felismerési pontosságot. A javulás minden esetben, kis és nagyobb méretű tanító adatbázisok mellett is[2] szignifikáns volt.

A felismerési idők ugyan különböztek a fenti teszteknél[3], azonban egyrészt – ahogy korábban említettem – a keresési mélység bőven a telítési szakaszra állította a felismerési pontosságokat, másrészt minden tesztnél a valós időnél gyorsabb volt a rendszer (RTF<1).

A következőkben két környezetfüggő beszédhangmodellezési eljárást vetek össze az előzőek szerinti tesztekben.

I.2. tézis: [B2]. *Kísérleti úton megmutattam, hogy az alapvetően nyelvi szabályok által vezérelt visszametszéses trifón állapotcsoportosítású beszédhangmodellekkel elért beszédfelismerési pontosságoknál elérhető szignifikánsan jobb eredmény a jelentősebb mértékben statisztikai elvű, ún. maximum likelihood fonetikus döntési fa alapú trifón állapotcsoportosítási módszer [Young & Odell+ 94] alkalmazásával.*

A felismerési eredmények a 2. táblázatban találhatóak. A tézis jelentősége nem csupán gyakorlati, hanem elvi is. Noha az első rendszernek részletesebb nyelvi, szakértői tudást adtunk bemenetül mint a másodiknak, az utóbbi, a statisztikával

[2] Az eredményeket a disszertáció 4.4. alfejezetében részletezem, lásd 4.2., 4.3., 4.5. és 4.6. táblázatok.
[3] A részletes RTF eredményeket a disszertáció 4.4. és 4.7. táblázata foglalja össze.

közvetlenül támogatott és nemzetközi szinten széles körben alkalmazott ML döntési fa alapú megoldás jóval hatékonyabbnak bizonyult annak révén, hogy a környezeti struktúrát közvetlenül az adatok által vezérelten alakította ki.

2. Táblázat
Telefonos, környezetfüggő állapotcsoportosítású folyamatos beszédfelismerési eredmények összefoglalása

Teszthalmaz	Átlagos szófelismerési pontosság [%]		Átlagos relatív hibacsökkenés [%]
	Referencia: Visszametszéses trifón beszédhangmodell	ML döntési fa alapú trifón beszédhangmodell	
Izolált szavas felismerés			
Illeszkedő	95.0	96.3	*26*
Nem illeszkedő	91.5	93.6	*24*
Folyamatos felismerés			
Illeszkedő	90.8	92.5	*19*
Nem illeszkedő	41.6	50.0	*14*

Noha a 2. táblázat is csak összesített eredményeket tartalmaz, minden kísérleti beállításnál szignifikáns javulás volt tapasztalható[4]. A futási idők itt sem voltak kiegyenlítettek, ugyanakkor a jobb módszer már gyorsabbnak is bizonyult a hatékonyabb keresésnek köszönhetően[5].

4.2. II. téziscsoport: A fonológiai koartikuláció (hasonulási jelenségek) modellezése magyar nyelvű beszéd gépi felismeréséhez

A fonológiai koartikuláció az egymást követő beszédhangok egymásra hatásának olyan típusát jelenti, ahol legalább egy résztvevő beszédhang fonémaértéke megváltozik. Egyrészről magától értetődőnek látszik e jelenségek explicit modellezésének szüksége, hiszen a fonémaszint majd' minden nagyszótáras beszédfelismerő rendszerben expliciten megjelenik, valamint elvi okokból is, tudni illik, a fonéma értékű változás a szó értelmét is megváltoztathatja. Ugyanakkor az a kérdés is felvethető, hogy szükséges-e egyáltalán az általános fonetikai koartikuláció modellezését szétválasztani a fonológiai koartikuláció modellezésétől.

Korábban a fonológiai koartikulációs jelenségek gépi beszédfelismerésnél történő explicit modellezésének nagy jelenséget tulajdonított a nemzetközi beszédkutató közösség. [Cohen 89] mind a magán- mind a mássalhangzók ejtésvariációit alternatív allofón realizációkkal javasolta modellezni. Részben ennek nyomán igen elterjedt a fonológiai szintű alternatív kiejtési változatok alkalmazása, melynél a fonológiai koartikulációs és az egyéb (pl. nyelvjárási) eredetű kiejtési variációkat általában nem választották szét. [Kaplan & Kay 94], [Mohri & Sproat 96], [Hazen & Hetherington+ 02] súlyozott FST alapú fonológiai szabályreprezentáció mellett 4 – 8 % (relatív) felismerési pontosság javulásról számol be angol nyelvű telefonbeszéd-felismerés esetén.

[4] A részeredmények a disszertáció 4.2., 4.3., 4.5. és 4.6. táblázatában találhatók meg.
[5] A futási idők a disszertáció 4.4. és 4.7. táblázatából olvashatók ki.

A többféle kiejtési változat azonban nem feltétlenül javítja a felismerési pontosságot. Amint [Lamel & Adda 96] rámutat, a túl sok alternatíva konfúzzá teheti a felismerési hálózatot, így a felismerési pontosság jelentősen romolhat. Majd [Jurafsky & Ward+ 01] meggyőző kísérletekkel támasztja alá, hogy a szótagszintű kiejtésbeli megváltozásoknál kisebbeket – a fonológiai koartikulációk döntően ilyenek – a trifón modellezés önmagában implicite jól kezeli. Ezután a nemzetközi kutatási trendek mind inkább az implicit kiejtés-modellezés felé irányulnak [Hain 02], [Kanthak & Ney 02], [Killer & Stüker+ 03], ami megkérdőjelezi a fonológiai koartikuláció explicit kezelésének szükségességét a statisztikai alapú gépi beszédfelismerésben.

A magyar nyelvi fonológiai koartikulációs jelenségek, ejtésvariációk tanulmányozásával több munka is foglalkozik, pl. [Gósy 98] [Vicsi & Szaszák 04] [Zsigri & Tóth+ 04], [Tóth 09] de konkrét, beszédfelismerési alkalmazásokban elért és valamely referenciával összehasonlított eredményekről általában nem szólnak. Kivételt jelent [Szarvas 03], ahol szóhatárokon is átívelő fonológiai koartikuláció modellezés hatására a folyamatos magyar nyelvű beszédfelismerés pontosságának javulásáról számol be a szerző. Az általános következtetések levonása azonban itt is nehéz (a kísérlet körülményei nincsenek pontosan megadva, az adatbázisméretek kicsik) és szignifikancia-vizsgálat sem történt.

A következőkben bemutatom a magyar nyelvű fonológiai koartikuláció modellezés terén elért eredményeimet. Az első tézis megmutatja, milyen körülmények között javította a felismerést az explicit modell, míg a másik tézis azt mutatja meg, hogy általánosabb körülmények között az implicit fonológiai koartikulációs modell versenyképes lehet az explicittel szemben.

II.1. tézis: [J2, B3, B4, C6, C9, C10] *Kísérleti úton megmutattam, hogy – amennyiben az akusztikus modellek tanításakor a tanító adatbázisban a fonológiai koartikuláció figyelembe lett véve (például kézi fonetikus átirat révén) – a felismerési tesztekben egyes tipikus fonológiai koartikulációs jelenségek explicit (szóhatárokon is átívelő) modellezésével elérhető szignifikánsan magasabb beszédfelismerési pontosság, mint a jelenség tekintetbe vétele nélkül.*

3. Táblázat
Folyamatos, telefonos, fonológia koartikulációs beszédfelismerési eredmények kézi fonológiai átírású tanítóhalmaz mellett

Teszthalmaz	Szófelismerési pontosság [%]		Relatív hibacsökkenés [%]
	Referencia: nincs fonológiai modell a teszthálózatban	Explicit fonológiai koartikuláció modellezés	
Illeszkedő	91.4	92.6	*14*
Nem illeszkedő	49.5	51.1	*3.2*

A fenti kísérletek a tézisfüzet 3. fejezetében már ismertetett telefonos magyar nyelvű beszédadatbázisokon történtek. A modellezett hasonulási jelenségek köre és módja a következő volt:

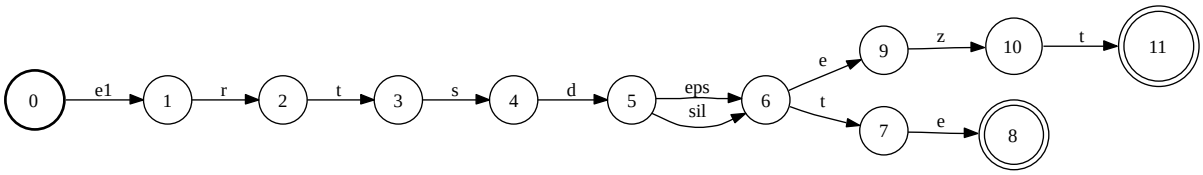
P_1: Zöngésségi hasonulás /kötelező/
P_2: Összeolvadás + Rövidülés /kötelező/
P_3: Képzés helye, módja szerinti részleges hasonulások /opcionális/
P_4: Képzés helye, módja szerinti teljes hasonulások /opcionális/

Az explicit fonológiai koartikulációs modell, P, az alábbi véges állapotú átalakító (FST: Finite State Transducer) kompozíciósorozattal adódik:

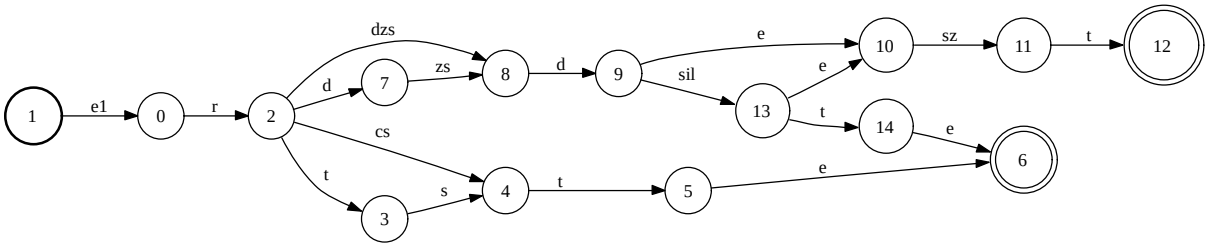
P = P2 o P4 o P3 o P2 o P1 (7)

Ez a modell a fonológiai koartikulációs jelenségeket explicit módon, *szóhatárokon átívelve* (is) kezeli.

A P transzducer fonológiai koartikulációs modellezési képessége a következő példákkal szemléltethető:



1. ábra. Kapcsolt szavas fonémaszintű felismerési hálózat (F), jelöletlen fonológiai koartikulációval. (Az „értsd te” és „értsd ezt” szókapcsolatok nyers fonológiai szintű FST reprezentációja.)



2. ábra. Kapcsolt szavas fonémaszintű felismerési hálózat explicit fonológiai koartikuláció modellezésével (P o F). (Az „értsd te” és „értsd ezt” szókapcsolatok felszíni fonológiai szintű FST reprezentációja.)

Ahogy tehát a 3. táblázat mutatja, *kézi* fonológiai átirat mellett tanított akusztikus modellek esetén szignifikánsan javítható volt a felismerési pontosság a felismerési hálózat explicit fonológiai modellekkel való kiterjesztésével. Azonban a kézi fonológiai átirat tipikusan nem áll rendelkezésre, ilyenkor a gépi módszerek egymáshoz vetése lehet hasznos.

II.2. tézis: [C6], *Kísérleti úton megmutattam, hogy implicit fonológiai koartikuláció modellezéssel – amikor is mind az akusztikus modellek tanításakor, mind a felismerési tesztek során eltekintünk a fonológiai koartikuláció jelenségétől – elérhető kompetitív (nem szignifikánsan alacsonyabb) beszédfelismerési pontosság ahhoz képest, mint amikor a fonológiai koartikulációs jelenségek jelentős részét expliciten modellezzük mind tanítás, mind tesztelés során.*

Következmény: *A tipikus fonológiai koartikulációs jelenségek explicit modellezése nem nélkülözhetetlen a magyar nyelvű gépi beszédfelismerésben, hiszen az explicit modellek körülményes integrációja elhagyható anélkül, hogy a felismerési pontosság feltétlenül szignifikánsan csökkenne.*

4. Táblázat
Folyamatos, telefonos, fonológia koartikulációs beszédfelismerési eredmények összefoglalása gépi fonológiai átírású tanítóhalmazok mellett

Teszthalmaz	Átlagos szófelismerési pontosság [%]		Átlagos relatív hibacsökkenés [%]
	Referencia: Explicit fonológiai koart. modellezés	Implicit fonológiai koartikuláció modellezés	
Illeszkedő	92.5	92.5	0.2
Nem illeszkedő	50.0	50.8	-1.6

A 4. táblázat kísérletei négyféle tanító adatbázisméret mellett történtek és szignifikáns különbség egyetlen esetben sem adódott az implicit és explicit fonológiai modellezési megközelítések között[6]. Ellenben a futási idők az explicit modell esetében szignifikánsan nagyobbak voltak[7].

Összevetve az I. és II. téziscsoport eredményeit, látható, hogy míg a fonetikai koartikuláció explicit modellezése úgymond elengedhetetlen, a fonológiai koartikulációé nem az – következetes tanítási-tesztelési körülmények esetén. A mérnöki szempontokon túl az eredmények felvetik azt az elvi kérdést is, hogy lényegét tekintve különböző jelenségekről van-e szó, avagy a fonológiai koartikulációs jelenségek csupán az általános koartikuláció egyes megnyilvánulásai.

4.3. III. téziscsoport: Spontán magyar nyelvű beszéd lexikai modellezése gépi beszédfelismeréshez

Magától értetődő, hogy a beszéd akusztikai modellezésénél a szavakat hangokra bontjuk. Nem egyértelmű viszont, hogy a (nagyszótáras) folyamatos beszéd felismeréséhez alkalmazott nyelvi modellezésnél is szükséges-e a szavakat kisebb egységekre bontani, és ha igen, hogyan. Ezzel a kérdéssel a lexikai modellezés foglalkozik. A lexikai modellezés célja olyan lexikon és technika előállítása, mely a morfológiailag változatos nyelveknél (is) lehetővé teszi a szótárméret, illetve a nyelvi modellezés adatelégtelenségi („data sparsity”) problémáinak kezelését, kézben tartását.

Formailag a szónál rövidebb („subword”) alapú beszédfelismerés a szó alapúnak egy egyszerű általánosításával érhető el:

$$\hat{M} = \arg\max_{M} P(M)P(O \mid M) \tag{8}$$

$$\hat{W} = f(\hat{M}) \tag{9}$$

ahol W szósorozatot, M a szónál kisebb lexikai egységekből alkotott sorozatot, O akusztikus megfigyelés (jellemzővektor) sorozatot jelöl és f pedig egyszerű szöveg-összefűzési és törlési műveleteket a becsült (felismert) lexikai egységekből álló sorozaton.

[6] A részletes beszédfelismerési eredmények a disszertáció 5.2. alfejezetében találhatók meg.
[7] A futási időket a disszertáció 5.3. táblázata foglalja össze.

A korábban leginkább kutatott nyelvek (angol, francia, német, spanyol, stb.) esetén szinte kizárólagos a szó lexikai egységek használata. A morfológialag változatosabb nyelvek esetén – mint pl. a finn, észt, magyar, török, arab – a toldalékolás, tő- és toldalékváltás miatt kézenfekvőbbnek látszik morfémaszerű egységek használata. Ugyanis, a nagy szóalaki változatosság miatt szó alapon óriási szótárméretek, a szótáron kívüli szavak nagy aránya és rosszul becsült nyelvi modell paraméterek adódnának. Egyes esetekben jelentős [Hirsimäki & Creutz+ 06], más nyelveknél, feladatoknál kisebb [Arisoy & Can+ 09], és esetenként negatív javulást is [Creutz & Hirsimäki+ 07] hozott a szónál kisebb lexikai egységek (továbbiakban: morf-ok) alkalmazása. A pozitív eredmények minden esetben tervezett vagy nem pontosan definiált beszédstílus [Afify & Sarikaya+ 06] mellett keletkeztek, spontán beszéd felismerésénél pedig romlást jelentettek [Creutz & Hirsimäki+ 07].

A magyar nyelv tekintetében nem ismeretes korábbról olyan tanulmány, mely a szó alapú lexikai modellezést valamely morf alapúval összevetve javulást mért volna a felismerési pontosságban az utóbbi javára. [Szarvas 03] morf alapú statisztikai modellezést javít morfo-szintaktikai szabályok hozzáadásával, [Vicsi & Velkei+ 05] pedig kötött témakörű folyamatos diktálásnál alkalmaz grammatikai morfokat, de a szóalakokat nem állítja vissza.

A harmadik téziscsoport tehát kettős előrelépésről számol be: egyrészt magyar nyelven mutat be szignifikáns javulást a szó helyett morf lexikai modellek hatására, másrészt a magyar nyelvtől elvonatkoztatva úttörő eredmény, hogy ezt spontán beszéd felismerésével teszi.

III.1. tézis: [J1, B1, C1, C2, C3, C4, C5] *Kísérleti úton megmutattam, hogy spontán beszéd gépi felismerése esetén szó helyett kisebb, morfémaszerű (továbbiakban: morf) lexikai egységek megfelelő alkalmazásával elérhetők szignifikánsan magasabb felismerési pontosságok.*

A tézist alátámasztó kísérleteket a MALACH magyar nyelvű beszédadatbázison végeztem. A teljes teszthalmazon mért felismerési eredmények az 5. táblázatban találhatók.

Az előző tézis finomítása a III.2. tézis, mely megmutatja, hogy a szó alapúnál jobb eredményt adó lexikai modellezéshez nincs szükség morfo-szintaktikai ismeretek explicit alkalmazására.

III.2. tézis: [J1, B1, C1, C2, C5] *Kísérleti úton megmutattam, hogy spontán beszéd gépi felismerése esetén felügyelet nélküli statisztikai módszerrel [Creutz & Lagus 05b] származtatott morf lexikai egységek alkalmazásával elérhetők a szó alapú megközelítés eredményeitől szignifikánsan magasabb, a nyelvi szabály alapú ill. kombinált (statisztika + nyelvi szabályok alapján előállított) morf megközelítések eredményeitől pedig nem szignifikánsan alacsonyabb felismerési pontosságok.*

5. Táblázat
Spontán folyamatos nagyszótáras beszédfelismerési eredmények (MALACH) összefoglalása különféle lexikai modellek mellett

Lexikai modell	Szótár-méret	Felismerési pontosság [%]		Relatív hibacsökkenés [%]	
		Szó	Betű	Szó	Betű
Szó – referencia	20k	45.5	72.9	-	-
Stat. morf (MB)	4.6k	46.4	73.4	1.7	*1.8*
Stat. morf (MC-MAP)	5.5k	46.8	73.7	*2.4*	*3.0*
Gramm. morf (HSF)	8k	46.5	73.7	1.8	*3.0*
Gramm. morf (HCG)	6.7k	46.8	73.8	*2.4*	*3.3*
Komb. morf (CHM)	6.7k	47.0	73.9	*2.8*	*3.7*

A kísérletekben a szónál kisebb lexikai egységek kialakítása az alábbi módszerek szerint történt:

1.) Statisztikai úton – felügyelet nélküli tanítással – származtatott morf lexikai egységek:

- ***MB (Morfessor Baseline)*** alapú megközelítés: a [Risannen 78]-szerinti MDL (Minimum Description Length) alapelven nyugvó módszerre épít. A szógyakoriságokat nem, csak a szóalakokat vesszük figyelembe. A szóalakokon kívül semmilyen más információt nem használ az MDL értelemben optimális szó-morf leképezés kialakításához [Creutz & Lagus 05a].
- ***MC-MAP (Morfessor Categories – MAP*)** alapú módszer: Az MB továbbfejlesztése, automatikus prefix, tő, és suffix kategorizálással finomítja az MB által létrehozott morf-okat [Creutz & Lagus 05b].

2.) Nyelvi tudás alapján – nyelvfüggő morfoszintaktikai szabályok és tő-toldalék adatbázisok alapján [Trón & Németh+ 05], [Trón & Halácsy+ 06] származtatott morf lexikai egységek:

- ***HSF(Hunmorph Strict Fallback)*** alapú megközelítés: első körben reguláris, nem összetett szóként próbálja elemezni és tő, toldalék morf-okra bontani az adott szót. Ha ez sikertelen, akkor összetett szóként próbálkozik, és ha így sincs eredmény, heurisztikák alapján szegmentálja a bemeneti egységet.
- ***HCG(Hunmorph Compound Guessing)*** alapú módszer: egy nekifutásra történik az elemzés, ahol összetett szó feltételezése, és heurisztikák alkalmazása is megengedett. Sokkal többféle alternatív szegmentációt eredményez, mint a HSF eljárás.

Ha többféle elemzés – és így kimenet – adódik a nyelvi szabályok alapján, a legtöbb morfot eredményező első felbontást választjuk.

3.) Nyelvi és statisztikai tudás kombinálásával:

- ***CHM (Combined Hunmorph Morfessor):*** az MB és a HCG módszer kombinációja, mely a HCG többszörös elemzési kimeneteiből az MB technikára alapozva választja ki a végleges morf készletet. A technika részleteiben [C5]-ben található, az értekezésben felső referenciaként használatos.

Amint az 5. táblázatban is látható, mind felügyelet nélküli statisztikai MC-MAP módszerrel, mind a nyelvi szabályokon alapuló HCG módszerrel, mind a kombinált CHM módszerrel szignifikánsan sikerült meghaladni a hagyományos szó alapú lexikai modellezés által elért beszédfelismerési eredményeket (szó- és betűhibaarány értelemben is). Ugyanakkor az előbb említett három morf alapú megközelítés eredményei között szignifikáns eltérés nem volt. A pontos összehasonlítás érdekében az RTF=4.2-4.3 tartományon lett tartva.

4.4. IV. téziscsoport: Spontán magyar nyelvű beszéd akusztikai és kiejtés-modellezése gépi beszédfelismeréshez

A hagyományos beszédfelismerő rendszerekben a leírt (ortografikus) szavak absztrakt fonémasorozattá képződnek le, majd a fonémák környezetfüggő beszédhangrészletek sorozatává. Ahhoz, hogy az ortografikus szóalakoktól a fizikai beszédhangrészletekig eljussunk, számos nyelvspecifikus szabály, tudásforrás alkalmazása szokásos. Ilyenek például a graféma-fonéma átalakítási szabályok, a kiejtési kivételszabályok, (ide értve az alternatív kiejtések kezelését is), fonológiai koartikulációs jelenségek és fonetikai csoportosítások. Láttuk, hogy a korábbi kísérletekben a fonológiai koartikuláció explicit modellezése nem tette érdemben hatékonyabbá a gépi beszédfelismerést, ugyanakkor meglepő lehet a feltevés, hogy a többi nyelvfüggő szabály sem nélkülözhetetlen, azaz teljes hiányuk sem csökkenti jelentősen a felismerési pontosságot.

Az ún. graféma alapú beszédfelismerésnél a szavakat alfabetikus karakterek sorozatára bontjuk, majd az akusztikai modelleket közvetlenül a *betűkre* építve az I. téziscsoportban is hivatkozott ML döntési fával alakítjuk ki. A korábban alkalmazott fonetikai osztályokat egyszerűen le lehet képezni graféma osztályokká, melyek révén részben hasznosítani tudjuk az alacsonyszintű fonetikai ismereteket. Ezzel a technikával – és így a fonéma szint, vagyis a graféma-fonéma átalakítási szabályok, kiejtési kivételek nélkülözésével – versenyképes eredményeket értek el német, spanyol nyelven [Kanthak & Ney 02]. [Killer & Stüker+ 03] még tovább ment, és a fonetikai ismeretek alkalmazását is elhagyva, teljesen adatvezérelt módon képezte le a környezetfüggő graféma modelleket fizikai beszédhangmodell részletekké („szingleton” technika).

A szónál kisebb lexikai egységeket használó nyelvi modellezést különösen jól egészíti ki a graféma alapú akusztikus és kiejtési modellezés. Ugyanakkor nem ismert korábbról olyan tanulmány, mely morf alapú lexikai modellek esetén veti össze a graféma és fonéma akusztikus modellek teljesítményét. Továbbá, magyar nyelvű beszédfelismerésnél sem ismeretes olyan tanulmány, mely a graféma és fonéma alapú akusztikus modellezés eredményességét összevetette volna, csupán kezdeti, referencia nélküli eredmények születtek parancsszó-felismerésre [Zgank & Kacic+ 2005].

A negyedik téziscsoport újdonsága tehát egyrészt az, hogy valós magyar nyelvű, folyamatos, nagyszótáras beszédfelismerési feladaton hasonlítja össze a fonéma és graféma alapú akusztikus modellezés hatását. Másrészt, hogy szónál kisebb lexikai egységek esetén történik az összehasonlítás.[8] További érdekesség, hogy a felismerési

[8] A tézisfüzetben csak egyféle morf alapú lexikai modellezés eredményeit tárgyaljuk. Átfogó összehasonlítás lexikai és akusztikai modellezési eredmények között a disszertáció 6.2. és 7.2. táblázataiban, illetve a [J1]-ben található.

feladat jellegéből következően jelentős a kivételes ejtésű szavak aránya (lásd 6. táblázat).

6. Táblázat
A magyar MALACH adatbázis esetén a szakértői kézi címkézés alapján számolt kivétel- és súlyozott kivételszótárak mérete és fedése a tanító adatbázison. A kivételszótár részét képezi a súlyozott kivételszótár.

Lexikai modell típus	Teljes szótár mérete	Kivételszótár		Súlyozott kivételek szótára	
		Méret	Fedés [%]	Méret	Fedés [%]
Szó	20k	1743	47.1	720	46.2
Morf (MC–MAP)	5.5k	492	27.3	163	26.9

IV. 1. tézis: [J1, C3] *Kísérleti úton megmutattam, hogy spontán, magyar nyelvű beszéd gépi felismerése esetén környezetfüggő graféma (alfabetikus karakter) alapú akusztikus modellezéssel elérhető nem szignifikánsan alacsonyabb felismerési pontosság, mint fonéma alapúval (morf lexikai modellezés mellett).*

Következmény: *Kézi kivételszótárak és graféma-fonéma átalakítási szabályok alkalmazásának hiánya nem feltétlenül okoz szignifikáns felismerési pontosságromlást magyar nyelvű gépi beszédfelismerésnél. Ezek a nyelvi tudásforrások tehát nem tekintendők nélkülözhetetlennek a magyar nyelvű gépi beszédfelismerésben.*

A környezetfüggő grafémák leképezése fizikai beszédhangrészletekké ugyanazzal a ML döntési fán alapuló trifón állapotcsoportosítási technikával történt, mint amit az I. téziscsoportnál is alkalmaztam. A döntési fa építéshez felhasznált fonéma osztályokat (nazális, labiális, dentális, zöngés, stb.) [Kanthak & Ney 02] szerinti módszerrel képeztem graféma osztályokká. A MALACH adatbázissal készült kapcsolódó kísérleti eredmények összefoglalása a 7. táblázaton látható.

Természetesen adódik a kérdés, hogy a döntési fa építésnél használt, a fonéma vagy graféma osztályok által reprezentált nyelvspecifikus tudás elhagyható-e, hasonlóan a többi nyelvspecifikus szabályhoz. A választ a következő tézis adja meg.

IV. 2. tézis: [J1] *Kísérleti úton megmutattam, hogy spontán magyar nyelvű beszéd gépi felismerése esetén környezetfüggő ún. graféma-szingleton alapú akusztikus modellezéssel – amikor is az alkalmazott ML döntési fa alapú trifón állapotcsoportosításnál csupán triviális, egyelemű graféma osztályokat definiálunk – elérhető nem szignifikánsan alacsonyabb felismerési pontosság, mint fonéma alapú akusztikus modellekkel (morf lexikai modellezés mellett).*

Következmény: *Nyelvspecifikus szabályok és szakértői nyelvi tudás explicit alkalmazásának hiánya nem feltétlenül okoz szignifikáns pontosságcsökkenést magyar nyelvű gépi beszédfelismerés esetén. A nyelvi szakértői tudás és a nyelvspecifikus szabályok explicit alakjukban (fonetikai osztálydefinícók, kiejtési és betű-hang átalakítási szabályok, szótárak, stb.) nem tekintendők tehát a magyar nyelvű gépi beszédfelismerés létfontosságú kellékeinek.*

7. Táblázat
Spontán folyamatos nagyszótáras MC-MAP (statisztikai) morf alapú (MALACH) beszédfelismerési eredmények összefoglalása különféle akusztikai modellek mellett

Akusztikai modell	Felismerési pontosság [%]		Relatív hibacsökkenés [%]	
	Szó	Betű	Szó	Betű
fonéma - referencia	46.8	73.7	-	-
graféma	46.2	73.5	-1.1	-0.7
graféma-szingleton	46.3	73.6	-0.9	-0.3

Az eredmények a 7. táblázaton láthatók. A szingleton osztályokba csak egyetlen tag tartozik, mely egy graféma (pl. „s” vagy „y”). Azaz, ilyenkor semmilyen ismeretünk nincs arról, hogy mely graféma mely más grafémákhoz hasonló akusztikai paraméterekkel realizálódik (pl. elöl vagy hátul képzett, nazális, bilabiális stb.). Elmondhatjuk tehát, hogy ekkor a döntési fa építéshez használt segédinformációban sem jelenik meg semmilyen nyelvspecifikus szakértői tudás.

A szemléletesség kedvéért összefoglaljuk, melyik megközelítés mely típusú nyelv-specifikus szabályokat alkalmazza.

- Fonéma alapú modell (referencia):
  - Valószínűségi súlyozású alternatív kiejtések, pl.
    miért 0.011 m é
    miért 0.426 m é r
    miért 0.269 m i é r
    miért 0.292 m i é r t
  - Idegen és hagyományos írású szavak kivételes kiejtései, pl.
    Churchill cs ö r cs i l
    Kossuth k o s ú t
  - Graféma-fonéma átalakítási szabályok, pl.
    cz c
    ch# cs
    ck# k
    ly j
    (# a szóhatár szimbólumot jelöli)
  - Fonetikai kategóriák, pl.
    NASAL: m, n, ny
    FRONT: e, é, i, í, ö, ő, ü, ű

- Graféma alapú modell:
  - (Gra)fonetikai kategóriák, pl.
    NASAL: m, n, n, y
    FRONT: e, é, i, í, ö, ő, ü, ű

- Graféma-szingleton alapú modell:
  –
  (nincs nyelvspecifikus szabály)

Ahogy a 7. táblázat mutatja, nemcsak hogy alig és nem szignifikánsan romlott a graféma alapú rendszerek felismerési pontossága, a mindenféle nyelvi szabályt nélkülöző megközelítés még valamivel jobban is teljesített, mint a köztes, fonetikai kategóriákat használó, de egyébként adatvezérelt módszer.

4.5. V. téziscsoport: Spontán magyar nyelvű beszéd felismerése explicit nyelvi szabályok nélkül (szintézis)

V. 1. tézis: [J1] *Kísérleti úton megmutattam, hogy spontán magyar nyelvű beszéd gépi felismerésénél explicit nyelvi ismeretek alkalmazása nélkül*[9] *is elérhető kompetitív felismerési pontosság a klasszikus szó-fonéma alapú megközelítéshez képest, mely számos nyelvspecifikus szakértői tudás*[10] *alkalmazását igényli.*

8. Táblázat
Klasszikus és nyelvspecifikus szabályoktól mentes spontán folyamatos nagyszótáras (MALACH) beszédfelismerési eredmények összefoglalása

Lexikai – akusztikai modell	Felismerési pontosság [%]		Relatív hibacsökkenés [%]	
	Szó	Betű	Szó	Betű
Szó – fonéma	45.5	72.9	-	-
Stat. morf (MC-MAP) – graféma-szingleton	46.3	73.6	1.5	*2.6*

Látható, hogy a széles körben elterjedt klasszikus szó-fonéma megközelítéshez képest a teljesen adatvezérelt, de a magyar nyelvű beszéd struktúráját a statisztikai morf lexikai modell révén figyelembe vevő technika nemcsak, hogy nem rosszabb, de a betűhibaarányt tekintve szignifikánsan jobb eredményt ért el.

A további, automatikus beszélőadaptációt alkalmazó kutatások a jelen és az előző két téziscsoport megállapításait – magasabb abszolút felismerési pontosságok mellett – megerősítették [J1, B1, C1, C2, C3, C4, C5].

[9] felügyelet nélküli statisztikai módszerrel meghatározott morf lexikai egységekkel, n-gram statisztikai nyelvi modellel, triviális morf-graféma leképezéssel, graféma-szingleton akusztikus modellel
[10] súlyozott ejtésvariációk, kivételszótárak, graféma-fonéma átalakítási szabályok, fonetikai-fonológiai kategóriák.

5. Az eredmények alkalmazhatósága

Az új tudományos eredmények gyakorlati alkalmazhatósága közel magától értetődő, hiszen a gépi beszédfelismerés pontosságának növelése, egyszerűbb, gyorsabb kialakítása kézzelfogható előnyök. A következőkben téziscsoport szerinti bontásban röviden részletezem az alkalmazhatóságot.

Az első téziscsoport (fonetikai koartikuláció-modellezés) eredményei várhatóan a beszédfelismerési hiba jelentős csökkentésére használhatók a magyar nyelvre klasszikusan használt környezetfüggetlen beszédhangmodellezéshez képes, kisméretű tanító adatbázisok esetén is.

A második téziscsoport (fonológiai koartikuláció-modellezés) eredményei az egyszerűbb és kisebb erőforrásigényű magyar nyelvű beszédfelismerést teszik lehetővé egy korábban fontosnak gondolt komponens nélkülözhetőségének megmutatásával.

A harmadik téziscsoport (lexikai modellezés) eredményei a magasabb felismerési pontosság elérésének érdekében használhatók spontán, magyar nyelvű, nagyszótáras, folyamatos beszédfelismerésnél. Ezzel egyidejűleg a morf lexikai modellezés az erőforrásigényeket is csökkentheti az által, hogy jóval kisebb szótárméretet igényel.

A negyedik téziscsoport (kiejtési modellezés) fő vívmánya, hogy gyors alkalmazásfejlesztést tesz kilátásba spontán magyar nyelvű, nagyszótáras, folyamatos beszédfelismerési és ezzel rokon területeken, ugyanis mentesíti az időigényes nyelvspecifikus szakértői szabályok alkalmazásától a fejlesztőket.

Végül az ötödik téziscsoport mintegy szintézisként foglalja össze az eredményeket. Alkalmazható lehet a gyors és redukált költségű, versenyképes beszédfelismerési alkalmazásfejlesztésre, mivel nyelvspecifikus szakértői szabályokat a javasolt megközelítés nem alkalmaz, ugyanakkor a magyar nyelv morfológiájának és írásmódjának jellegét figyelembe veszi.

A tézisek eredményeinek nagy része gyakorlati alkalmazásokban is hasznosul.

Irodalmi hivatkozások listája

[Afify & Sarikaya+ 06] Afify, Mohamed; Sarikaya, Ruhi; Kuo, Hong-Kwang Jeff; Besacier, Laurent; Gao, Yuqing (2006): "On the use of morphological analysis for dialectal Arabic speech recognition", In INTERSPEECH-2006, pp. 1444-1447

[Arisoy & Can+ 09] Ebru Arisoy, Dogan Can, Siddika Parlak, Hasim Sak and Murat Saraclar. Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio, Speech, and Language Processing, 17(5):874-883, July 2009

[Bahl & Jelinek+ 83] L. R. Bahl, F. Jelinek, R. L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179–190, March 1983.

[Baker 75] J. K. Baker. Stochastic modeling for automatic speech understanding. In Reddy, R., editor, Speech recognition, pp. 512–542, New York, USA, Academic Press, 1975.

[Baum & Eagon 67] L. E. Baum, J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. Amer. Math. Soc. Bull., Vol. 73, pp. 360–362, 1967.

[Bánhalmi & Kocsor+ 05] Banhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével, in Proc. of MSZNY 2005, pp. 337 – 347, Szeged, 2005.

[Bánhalmi & Paczolay+ 08] Banhalmi, A., Paczolay, D., Toth, L., Kocsor, A.: Investigating the robustness of a Hungarian medical dictation system under various conditions, International Journal of Speech Technology, VOLUME 9, ISSUE 3-4 (2008), PAGE 121-131.

[Bellegarda & Nahamoo 90] J. R. Bellegarda, D. Nahamoo. Tied mixture continous parameter modeling for speech recognition. IEEE Trans ASSP, Vol. 38, No. 12, pp. 2033–2045, December 1990.

[Bellman 57] R. E. Bellman. Dynamic Programming. Princeton University Press, Princeton, USA, 1957.

[Beulen & Ney 98] K. Beulen and H.Ney, Automatic Question Generation for Decision Tree Based State Tying, Proceedings of the ICASSP, pp- 805-808, Seattle, WA, 1998.

[Chen & Goodman 98] Stanley F. Chen and Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[Cohen 89] M. H. Cohen. Phonological structures for speech recognition. Ph.D. dissertation, University of California, Berkeley, USA, 1989.

[Creutz & Lagus 05a] Creutz, M. and Lagus, K., “Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.”, Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, (2005)

[Creutz & Lagus 05b] Creutz, M. and Lagus, K., “Inducing the Morphological Lexicon of a Natural Language from Unannotated Text”, In Proceedings of AKRR'05, Espoo, Finland, 15–17 June, (2005)

[Creutz & Hirsimäki+ 07] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, & A. Stolcke, Morph-based speech recognition and modeling of out-of-vocabulary words across languages, ACM Transactions on Speech and Language Processing 5(1), 2007.

[Czap 05] Czap L.: Audiovizuális beszédfelismerés és szintézis, PhD értekezés, BME, Budapet, 2005.

[Daniel 78] W. Daniel, Applied Nonparametric Statistics, Houghton Mifflin, 1978.

[Dempster & Laird+ 77] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1–38, 1977.

[Good 53] Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika, 40(3 and 4):237-264.

[Gordos & Takács 83] Gordos G., Takács Gy. (1983) *Digitális beszédfeldolgozás*, Műszaki Könyvkiadó, Budapest.

[Gósy 98] Gósy Mária. A zöngésségi hasonulás a (spontán) beszédben. Beszédkutatás 1998, Ed. Gósy Mária, Akadémiai kiadó, Budapest, pp. 1-20, 1998

[Hain 02] T. Hain. Implicit pronunciation modeling in ASR. Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, pp. 129–134, Estes Park, Colorado, USA, September 2002.

[Hazen & Hetherington+ 02] Timothy J. Hazen, I. Lee Hetherington, Han Shu and Karen Livescu, "Pronunciation modeling using a finite-state transducer representation," Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation, Estes Park, Colorado, September, 2002

[Hermansky 90] H. Hermansky. (1990) Perceptual linear predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738-1752.

[Jelinek & Bahl+ 75] F. Jelinek, F. Bahl, R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. Information Theory, 21(3), pp. 250–256, 1975.

[Jurafsky & Ward+ 01] Jurafsky, Dan – Ward, Wayne – Jianping, Zhang – Herold, Keith – Xiuyang, Yu – Sen, Zhang. “What kind of pronunciation variation is hard for triphones to model?”, in IEEE ICASSP-01, Salt Lake City, Utah, 2001, pp. I.577–580.

[Kanji 94] G. Kanji, 100 Statistical Tests, SAGE Publications, 1994

[Kanthak & Ney 02] S. Kanthak, H. Ney. "Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition". In IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1, pp. 845-848, Orlando, FL, May 2002. download PostScript

[Kaplan & Kay 94] Kaplan, R. M. & Kay, M. (1994). 'Regular Models of Phonological Rule Systems'. Computational Linguistics 20, nr 3, 332-387.

[Katz 87] S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 35, No. 3, pp. 400–401, March 1987.

[Kirchoff & Vergyri+ 06] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh, and A. Stolcke. 2006. Morphology-based language modeling for Arabic speech recognition. Computer Speech and Language, 20(4):589–608.

[Killer & Stüker+ 03] M. Killer, S. Stüker, and Tanja Schultz. Grapheme based Speech Recognition. Proc. Eurospeech, Geneva, Switzerland, September 2003

[Kwon & Park 03] O.-W. Kwon and J. Park. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. Speech Communication, 39(3–4):287–300.

[Leggetter & Woodland 95] C.J. Leggetter and P.C. Woodland. (1995) "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression, " Proc. ARPA Spoken Language Systems Technology Workshop, Austin, TX, pp. 110-115.

[Levinson & Rabiner+ 83] S. E. Levinson, L. R. Rabiner, M. M. Sondhi. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Techn. Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[López & Graña+ 03] López, K., Graña, M., Ezeiza, N., Hernández, M., Zulueta, E., Ezeiza, A. and Tovar, C., “Selection of Lexical Units for Continuous Speech Recognition of Basque”, Proc. of CIARP, Havana, Cuba (2003) 244–250

[MacQueen 67] J. B. MacQueen. (1967) "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[Mauuary 98] L. Mauuary. (1998) Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition, *Proc. EUSPICO’98*, Vol.1, pp. 359-363.

[Mermelstein 76] P. Mermelstein. (1976) Distance measures for speech recognition, psychological and instrumental, *Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed*., pp. 374–388. Academic, New York.

[Mohri & Sproat 96] Mehryar Mohri and Richard Sproat. An Efficient Compiler for Weighted Rewrite Rules.In 34th Meeting of the Association for Computational Linguistics (ACL '96), Proceedings of the Conference, Santa Cruz, California. Santa Cruz, California, 1996.

[Mohri 97] Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. Computational Linguistics, 23:2, 1997.

[Mohri & Pereira+ 02] Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language, 16(1):69-88, 2002.

[Myers & Rabiner, 81] C. S. Myers and L. R. Rabiner. (1981) A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389-1409, September

[Ney 84] H. Ney. The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 32, No. 2, pp. 263–271, April 1984.

[Ney & Mergel+ 87] H. Ney, D. Mergel, A. Noll, A. Paeseler. A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. I, pp. 833–836, Dallas, USA, April 1987.

[Risannen 78] Rissanen, J. (1978), 'Modeling By Shortest Data Description', Automatica, Vol. 14, pp 465-471

[Schillo & Fink+ 00] C. Schillo, G. A. Fink, and F. Kummert, Grapheme based speech recognition for large vocabularies, in Int. Conf on Spoken Language Processing, Beijing, China, Oct. 2000, pp. 129 132.

[Shafran & Hall 06] I. Shafran and K. Hall. 2006. Corrective models for speech recognition of inflected languages. In Proc. EMNLP, Sydney, Australia.

[Singh & Raj+ 99] Singh, R., Raj, B., Stern, R. M.: Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models. in Proc. Int. Conf. on Spoken Language Processing. Vol. 1 (1999) 117-120

[Stolcke 02] Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. on Spoken Language Processing, Denver (2002) 901–904

[Steinbiss & Tran+ 94] V. Steinbiss, B.-H. Tran, H. Ney. Improvements in Beam Search. Proc. Int. Conf. on Spoken Language Processing, Vol. IV, pp. 2143–2146, Yokohama, Japan, September 1994.

[Szarvas & Furui 02] Mate Szarvas and Sadaoki Furui "Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes" Proc. ICSLP2002, Denver, U.S.A., pp.1297-1300 (2002-9)

[Szarvas 03] Máté Szarvas. "Efficient large vocabulary continuous speech recognition using weighted finite-state transducers - The development of a Hungarian dictation system" Ph.D. dissertation, TITECH, Tokyo, Japan, 2003.

[Szaszák 2008] Szaszák György: Szupraszegmentális jellemzők szerepe és felhasználása a beszédfelismerésben, PhD disszertáció, BME, 2008.

[Titterington & Smith+ 85] Titterington, D., A. Smith, and U. Makov (1985) "Statistical Analysis of Finite Mixture Distributions," John Wiley & Sons.

[Tóth 06] Tóth, L.: Posterior-Based Speech Models and their Application to Hungarian Speech Recognition, Ph.D. Dissertation, University of Szeged, 2006.

[Tóth 09] Tóth, L.: Beszédfelismeresi kísérletek hangoskönyvekkel, Proc. MSZNY, pp. 206-216, 2009.

[Tóth & Kocsor+ 04] Tóth, L., Kocsor, A., Gosztolya, G.: Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, No. 4, 2004.

[Trón & Németh+ 05] Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. and Varga, D., “Hunmorph: open source word analysis”, In Proc. ACL 2005 Software Workshop, (2005) 77–85

[Trón & Halácsy+ 06] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon (2006), Morphdb.hu: Hungarian lexical database and morphological grammar, In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA, pages 1670--1673.

[Vicsi & Vig 98] Vicsi, K. - Vig, A.: Az elsõ magyarnyelvû beszédadatbázis, Beszédkutatás '98, MTA Nyelvtudományi Intézete, Budapest 1998, pp. 163-177

[Vicsi & Velkei+ 05] Vicsi K. Velkei Sz., Szaszák Gy., Borostyán G., Teleki Cs., Tóth Sz. L., Gordos G..: Középszótáras, folyamatos beszédfelismerőrendszer fejlesztési tapasztalatai, Proc. of MSZNY 2005, pp. 348 – 360.

[Vicsi & Tóth 02] Vicsi K., Tóth L. Kocsor A., Gordos G. Csirik J. (2002): MTBA - Magyar nyelvű telefonbeszéd adatbázis. Hiradástechnika 2002/8. sz. pp. 35-39.

[Vicsi & Szaszák 04] Klára Vicsi, György Szaszák: Examination of Pronunciation Variation from Hand-Labelled Corpora. TSD 2004: 473-480. Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings. Lecture Notes in Computer Science 3206 Springer 2004, ISBN 3-540-23049-1.

[Vicsi et al.] Vicsi Klára et al. http://alpha.tmit.bme.hu/speech/databases.php

[Vintsjuk 68] T. K. Vintsyuk, „Speech discrimination by dynamic programming”, Kibernetika, Vol. 4, pp. 81-88, Jan.-Feb. 1968

[Wilcoxon 45] Wilcoxon, F. "Individual Comparisons by Ranking Methods." Biometrics 1, 80-83, 1945.

[Young 06] S. J. Young. The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge, England, December, 2006.

[Young & Odell+ 94] Young, S. – Odell, J. – Woodland, P. Tree-based state tying for high accuracy acoustic modelling. DARPA Human Language Technology Workshop, pages 307–312, March 1994.

[Zgank & Kacic+ 05] Zgank, A. - Kacic, Z. - Diehl F. - Juhar, J. - Lihan, S. - Vicsi, K. - Szaszák, Gy.: Graphemes as basic units for crosslingual speech recognition,, COST 278 Workshop, 2005

[Zsigri & Tóth+ 04] Zsigri, Gy., Toth, L., Kocsor, A. Sejtes, Gy.: Az automata és kézi szegmentálás ejtésvariációk okozta problémai, Proc. MSZNY 2004.

A tézispontokhoz kapcsolódó tudományos közlemények

Folyóiratcikkek

[J1] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, T. Fegyó: Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task, *IEEE Transactions on Audio Speech and Language Processing*, Volume 18, Issue 6, pp. 1588-1600, 2010.

[J2] P. Mihajlik, T. Révész, P. Tatai: Phonetic Transcription in Automatic Speech Recognition, *Acta Linguistica Hungarica*, Volume 49, Issues 3-4, pp. 407-425, 2003.

Cikkek szerkesztett könyvekben

[B1] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske, V. Trón: Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages: Hungarian ASR for the MALACH Project, In: V. Matousek, P. Mautner (ed.): *Text, Speech and Dialogue*, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 2007, Proceedings, Lecture Notes in Computer Science, Volume 4629/2007, pp. 342-350.

[B2] Mihajlik P., Fegyó T., Tatai P.: Új eljárás a gépi beszédfelismerés környezetfüggő beszédhangmodelljeinek kialakítására. In: Gósy M (szerk.): *Beszédkutatás 2006*. MTA Nyelvtudományi Intézet, Budapest, 2006. pp. 218-230.

[B3] P. Mihajlik, P. Tatai, G. Gordos: Automatic Phonetic Transcription and Its Application in Speech Recogniser Training: A case study for Hungarian. In: P. Divenyi, S. Greenberg, G. Meyer (ed.): *Dynamics of Speech Production and Perception*, IOS Press, Amsterdam, NATO Science Series; I., 374., Life and Behavioural Sciences, 2006. pp. 245-262.

[B4] Mihajlik P., Tatai P.: Automatikus fonetikus átírás magyar nyelvű beszédfelismeréshez, In: Gósy M. (szerk.): *Beszédkutatás 2001*, MTA Nyelvtudományi Intézet, Budapest, 2001. pp. 172-185.

Konferenciacikkek[11]

[C1] B. Tarján and P. Mihajlik: On Morph-based LVCSR Improvements, *Proc. SLTU 2010*, May 3-5, 2010, Penang, Malaysia, pp. 10-16.

[C2] P. Mihajlik, B. Tarján, Z. Tüske, T. Fegyó: Investigation of Morph-based Speech Recognition Improvements across Speech Genres, *Proc. Interspeech 2009*, Sep. 6-10, 2009, Brighton, United Kingdom, pp. 2687-2690.

[C3] P. Mihajlik, T. Fegyó, Z. Tüske, P. Ircing: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian, *Proc. Interspeech 2007*, August 27-31, 2007, Antwerp, Belgium, pp. 1497-1500.

[11] Minden külföldi cikk teljes terjedelmében lektorált. Az MSZNY cikkek absztrakt alapján lektoráltak.

[C4] Tüske Z., Mihajlik P., Fegyó T.: Spontán, nagyszótáras, folyamatos beszéd gépi felismerési pontosságának növelése beszélőadaptációval a MALACH projektben, *V. Magyar Számítógépes Nyelvészeti Konferencia*. 2007. december 6-7, Szeged, pp. 47-55.

[C5] Németh B., Mihajlik P., Tikk D., Trón V.: Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz, *V. Magyar Számítógépes Nyelvészeti Konferencia*. 2007. december 6-7, Szeged, pp. 95-105.

[C6] Mihajlik P.: Koartikulációs modellek a magyar nyelvű gépi beszédfelismerésben, *IV. Magyar Számítógépes Nyelvészeti Konferencia*, 2006. december 7-8, Szeged, pp. 231-242.

[C7] T. Fegyó, P. Mihajlik, M. Szarvas, P. Tatai, G. Tatai: Voxenter™ – Intelligent Voice Enabled Call Center for Hungarian, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 1905-1908.

[C8] T. Fegyó, P. Mihajlik, P. Tatai: Comparative Study on Hungarian Acoustic Model Sets and Training Methods, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 829-832.

[C9] P. Mihajlik, T. Fegyó, P. Tatai, G. Gordos: Pronunciation Modeling in Continuous Number Recognition, *Proc. ECMCS 2001*, Sep. 11-13, 2001, Budapest, Hungary, pp. 330-333.

[C10] T. Fegyó, P. Mihajlik, P. Tatai, G. Gordos, Pronunciation Modeling in Hungarian Number Recognition, *Proc. Interspeech 2001*, Sep. 3-7, 2001, Aalborg, Denmark, pp. 1465-1468.

A szerző további tudományos közleményei (gépi beszédfeldolgozás témában)

Folyóiratcikkek

[J3] Németh G., Olaszy G., Bartalis M., Zainkó Cs., Fék M., Mihajlik P.: Beszédadatbázisok előkészítése kutatási és fejlesztési célok hatékonyabb támogatására. *Híradástechnika*, LXIII. évfolyam, 2008/5, pp. 18-24.

[J4] Tüske Z, Mihajlik P, Tobler Z, Fegyó T, Tatai P.: Beszéddetekciós módszerek vizsgálata és optimalizálása gépi beszédfelismerő rendszerekhez, *Híradástechnika*, LXI. évfolyam, 2006/3, pp. 59-67.

[J5] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Eredmények a magyar nyelvű nagyszótáras kapcsolt-szavas gépi beszédfelismerésben. *Híradástechnika*, LVI. évfolyam, 2001/6, pp. 31-36.

[J6] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Automatic Recognition of Hungarian: Theory and Practice, *International Journal of Speech Technology*, Volume 3, Numbers 3-4, pp. 237-251, 2000.

Cikkek szerkesztett könyvekben

[B4] Németh G., Olaszy G., Bartalis M., Kiss G., Zainkó Cs., Mihajlik P., Haraszti Cs.: Automated Drug Information System for Aged and Visually Impaired Persons, In: Miesenberger K, Klaus J, Zagler W, Karshmer A (ed.): *Computers Helping People with Special Needs*, Lecture Notes in Computer Science, Volume 5105/2008, Springer Berlin / Heidelberg, 2008. pp. 238-241.

[B5] Tüske Z., Simon M., Mihajlik P., Fegyó T.: Érzelmek automatikus felismerése a beszéd akusztikus jellemzői alapján. In: Gósy M. (szerk.): *Beszédkutatás 2007*. MTA Nyelvtudományi Intézet, Budapest, 2007. pp. 151-161.

[B6] Fegyó T., Mihajlik P., Tatai P.: Automatikus beszédfelismeréshez használt beszédhangmodellek betanítási módszereinek összehasonlító elemzése, In: Gósy M. (szerk.): *Beszédkutatás 2002*. MTA Nyelvtudományi Intézet, Budapest, 2002. pp. 185-196.

Konferenciacikkek

[C11] Tüske Z., Simon M., Mihajlik P., Gordos G.: A beszéd érzelemi töltetének számítógépes felismerése, *V. Magyar Számítógépes Nyelvészeti Konferencia*, 2007. december 6-7, Szeged, pp. 81-91.

[C12] Tarján B., Györki M., Mihajlik P., Gordos G.: Eredmények a magyar nyelvű beszédfelismerési konfidenciabecslésben. *IV. Magyar Számítógépes Nyelvészeti Konferencia*, 2006. december 7-8, Szeged, pp. 243-254.

[C13] Tüske Z., Mihajlik P., Tobler Z.: Új, zajbecsléssel kombinált beszéddetektálási eljárás a beszédfelismerési hatásfok javítására, *III. Magyar Számítógépes Nyelvészeti Konferencia*, 2005. december 8-9, Szeged, pp. 371-382.

[C14] P. Mihajlik, Z. Tobler, Z. Tüske, G. Gordos: Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 2677-2680.

[C15] Z. Tüske, P. Mihajlik, Z. Tobler, T. Fegyó: Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 245-248.