Lexikai modellezés a közlés tervezettségének függvényében magyar nyelvű beszédfelismerésnél

Tarján Balázs[1], Fegyó Tibor[1,2], Mihajlik Péter[1,3]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
[2] AITIA International Zrt.
[3] THINKTech Kutatási Központ Nonprofit Kft.
{tarjanb, mihajlik, fegyo}@tmit.bme.hu

Kivonat: A morfémákban gazdag nyelvek nagyszótáras, gépi beszédfelismerésénél gyakran használnak szónál kisebb elemekre, ún. morfokra épülő nyelvi modelleket. Ezek alkalmazása azonban többletmunkát, magasabb rendszerkomplexitást igényel, ugyanakkor a javulás mértéke változó. Cikkünkben a morfalapú nyelvi modellezéssel elérhető hibacsökkenés előrejelzésére teszünk kísérletet. Ehhez először azonosítjuk a hibacsökkenést befolyásoló tényezőket, majd kísérleti úton megvizsgáljuk pontos hatásukat. Eredményeink alapján elmondható, hogy a morfalapú modellek alkalmazása kisméretű tanítószövegek, illetve korlátozott szótárméret mellett járhat jelentős előnnyel. Előnyös még a kevésbé spontán, tervezettebb beszédet tartalmazó adatbázisok esetén, míg a jel-zaj viszony romlása csökkenti a hibacsökkenés mértékét, csakúgy, mint az abszolút hibát. Az utolsó fejezetben bemutatunk egy mérőszámot, mely erős összefüggést mutat a kísérleti adatbázisainkon mérhető morfalapú hibacsökkenéssel. Ez a mérőszám nem csak a feladat tervezettségét, hanem a tanítószöveg mennyiségét is figyelembe veszi.

1 Bevezetés

Gépi beszédfelismeréssel, vagyis az automatikus beszéd-szöveg átalakítást lehetővé tevő megoldásokkal korábban csak a modern technológiák iránt elszántan érdeklődő kevesek találkozhattak. Túlzás lenne azt állítani, hogy azóta mindennapjaink része, tény viszont, hogy az okostelefonok terjedésével immáron **rengeteg felhasználó számára vált elérhetővé** egy-egy a technológia élvonalába tartozó megoldás, akár magyar nyelven is. Adja magát a következtetés, hogy ezek szerint a technológia beérett, és innentől kezdve csak apró finomításokra van szükség. A szakirodalmat tanulmányozva, vagy akár csak egy korszerű rendszert huzamosabb ideig tesztelve azonban láthatjuk, hogy ez távolról sincs így. A legelterjedtebb rejtett Markov-modell (Hidden Markov Modell – **HMM**) alapú statisztikai felismerők teljesítménye továbbra is durván leromlik zajos környezetben a humán észlelőkkel szemben, illetve akusztikus és nyelvi modelljeink továbbra is csak egyelőre meghatározott felismerési feladatra működnek optimálisan.

A fent vázolt problémák okainak jobb megértését tűzte ki célul az ún. OUCH (Outing Unfortunate Characteristics of HMMs) projekt. A kutatást összefoglaló ta-

nulmány [1] szerzői egyrészt rámutatnak a jelenleg használt technológiák hiányosságaira, majd bemutatják a technológia legfontosabb szereplőivel készített interjúik eredményét is. Ez alapján szakmai konszenzus van azzal kapcsolatban, hogy se az akusztikus és nyelvi modellek, se a beszédjellemzők kinyerését célzó technikák **nem tekinthetőek érettnek**, ugyanis működésük nem elég robosztus, még akkor sem, ha nagyon sok pénzt fektetnek a fejlesztésükbe. A tanulmány egyik fontos végkövetkeztetése, hogy beszédfelismerés területén dolgozó kutatóknak több energiát kellene fordítaniuk a felismerési hibák okainak mélyebb megértésére.

Cikkünk célja ezzel összhangban, hogy jobban megismerjük a szó- és morfalapú nyelvi modellezés hibaarányai közötti összefüggéseket. Számos vizsgálat bizonyítja [2]–[4], hogy morfémákban gazdag nyelveken a folyamatos, nagyszótáras gépi beszédfelismerő rendszerek **hibája csökkenthető**, ha szavak helyett statisztikai úton nyert morfémákat (ún. morfokat) [5] alkalmazunk a nyelvi modellben. Semmi nem garantálja azonban, hogy ez a hibacsökkenés jelentős mértékű lesz, sőt azt sem, hogy nem növekszik a hiba [6]. Figyelembe véve a többletmunkát és komplexitás növekedést, amivel a morfalapú rendszerek tanítása jár, felmerül az igény a **várható hibacsökkenés előrejelzésére**.

Korábbi munkáinkban megvizsgáltuk a szöveges tanítóadat mennyiségének, az akusztikus modell illeszkedésének és a felismerési feladat tervezettségének kapcsolatát az elérhető hibacsökkenéssel [7], [8]. Mostani munkákban egyrészt szeretnénk korábbi megállapításainkat új adatbázisokon is tesztelni, valamint kiterjeszteni az ún. **követőmorfos** [4] nyelvi modellekre is. Ezenkívül új szempontként megvizsgáljuk a feldolgozandó hanganyag jel-zaj viszonyának illetve a felismerő rendszer szótárméretének hatását is. Morfalapú nyelvi modellezés esetén sajnos elkerülhetetlen, hogy valamilyen típusú speciális jelölést vezessünk be a szóhatár későbbi visszaállíthatósága érdekében. Érdekes kérdés, hogy mennyivel lehetne pontosabb egy olyan morfalapú rendszer, melyben **eltekintünk a szóhatár-visszaállítástól**. Cikkünkben ennek a meghatározására is kísérletet teszünk.

A következőkben először a televíziós híradók felvételeit tartalmazó tanító- és tesztadatbázist ismertetjük, majd kitérünk a modellek tanításnál és kísérleteinknél alkalmazott módszerekre. A felismerési feladat és módszertan bemutatása után ismertetjük a híradó adatbázison kapott eredményeket, majd az utolsó előtti fejezetben a hibacsökkenés előrejelzésére teszünk kísérletet. Végül összefoglalását adjuk vizsgálataink legfontosabb eredményeinek.

2 Tanító és tesztadatbázis

Kísérleteink döntő többségét egy **televíziós híradófelvételeket** tartalmazó adatbázison végeztük. Ilyen típusú – az angol terminológia szerint *broadcast speech*nek nevezett – adatbázison már korábban is kísérleteztünk [6], [7], [9], melyek tapasztalatait cikkünkben a vonatkozó részeknél felidézzük majd. Hasonló magyar nyelvű felismerési feladaton két további munkát fontos megemlíteni. Az első, mély neuronhálók tanítási módszereit veti össze, mely technika segítségével meg is javítja a HMM alapú akusztikus modell eredményét híradós felvételek felhasználásával [10]. Míg egy másik a témában született cikkben elsősorban a kézi leiratok felhasználása nélkül törté-

nő, felügyelet nélküli tanítási módszereken van a hangsúly [11]. A fenti két cikk egyike sem alkalmaz azonban morfalapú lexikai modelleket.

2.1 Akusztikus tanító- és tesztanyagok

Összesen 50 órányi televíziós híradó kézi leiratát készítettük el és használtuk fel a felismerő akusztikus tanításához. Tesztelési célokra 6 teljes híradó felvételét, összesen 155 perc hanganyagot különítettünk el. A 6 híradó mindegyike 2012 januárjában került adásba a TV2, a Duna TV és az MTV műsorán. A tesztanyagot két részre bontottuk: 2 híradót a fejlesztés során szükség paraméterek hangolására (**Dev**), míg a maradék 4-et a rendszer kiértékeléséhez (**Eval**) használtunk fel. A tesztanyag kézi átírásakor az egyes szegmenseket akusztikai tulajdonságaik szerint különböző csoportokba soroltuk. Az egyes csoportok jelentését és méretének eloszlását az **1. táblázatban** foglaljuk össze. Fontos megjegyezni, abban az esetben, ha egy szegmensre többféle kategória leírása is illett, akkor mindig a nagyobb sorszámú kategóriába soroltuk. Ebből következik például, hogy az F4 kategóriájú szegmensekben a zaj nem biztos, hogy a szegmens teljes hosszára kiterjed. A jel-zaj viszony (Signal-to-Noise Ratio – **SNR**) hozzávetőleges meghatározásához a NIST STNR[1] és WADA SNR [12] algoritmusokat alkalmaztuk. Ahol kevés adat állt rendelkezésre ott nem adtuk meg az SNR-t, mivel az algoritmusok nem szolgáltak megbízható értékkel.

1. táblázat: A tesztadatbázis eloszlása akusztikai kategóriák szerint

Jelölés	Jelentés	Hossz [perc]	SNR [dB]
F0	Tervezett beszéd csendes környezetben	38	20-25
F1	Spontán beszélgetés csendes környezetben	18	20-25
F2	Telefonos beszélgetés	2	-
F3	Beszéd háttérzenével	10	8-10
F4	Beszéd zajos környezetben	84	10-15
F5	Nem anyanyelvi beszélő	3	-

2.2 Szöveges tanítóanyagok

A felismerő nyelvi modelljének tanításához szükséges szövegkorpuszokat több forrásból gyűjtöttük össze. Egyrészt felhasználtuk az akusztikus modell tanításához használt 50 órányi hanganyag kézi leiratát (**TRS**). Ez önmagában túl kevés lett volna egy hatékony felismerő tanításához, így különböző webes híroldalakról is gyűjtöttünk további adatokat. A webes szövegek gyűjtésével, tárolásával és feldolgozásával kapcsolatos részletek [9]-ben találhatóak meg. Az ott bemutatott rendszerhez hasonlóan most is két részre bontottuk a webes tanítószöveget. Az első a tesztanyag előtti 30 napban (2011. december 1-31.) közölt híreket tartalmazza (**WEB 30D+**), míg a második minden anyagot, ami korábban keletkezett (**WEB 30D-**). Ennek a szétválasztásnak az a célja, hogy a nyelvi modellek interpolációja során a tesztelés időpontjához közelebb eső hírek nagyobb súlyt kaphassanak, mint a régebbiek. Részletes adatok a

[1] http://labrosa.ee.columbia.edu/~dpwe/tmp/nist/doc/stnr.txt

2. táblázatban találhatóak. A korábbi rendszerhez képest lényeges eltérés azonban, hogy a tanítószöveg normalizálása során a kivételes kiejtéssel rendelkező szavakat (tipikusan nevezett entitásokat) a kiejtett alakjuknak megfelelően írtuk át. Ennek a változtatásnak a célja az volt, hogy a szó- és morfalapú eredmények összehasonlíthatóságát ne befolyásolják a kivételes írásmódú szavak szegmentálási nehézségei.

2. táblázat: A tanítószövegek részletes adatai

Tanítószöveg	Összes szó [millió szó]	Szótárméret [ezer szó]	Eval PPL [-]	Eval OOV [%]
TRS	0,53	74	598	10.0
WEB 30D+	1,2	115	761	8.3
WEB 30D-	50,1	955	551	1.5

3 Tanítási és kísérleti módszerek

Ebben a fejezetben a tanítás során alkalmazott modellezési lépéseket és a kísérleti körülményeket ismertetjük.

3.1 Akusztikus modell

A híradó felismerési feladathoz tartozó akusztikus modell tanításához az erre a célra elkülönített 50 óra hanganyagot használtuk fel. Az annotált felvételek segítségével háromállapotú, balról-jobbra struktúrájú, környezetfüggő rejtett Markov-modelleket tanítottunk a Hidden Markov Model Toolkit [13] eszközeinek segítségével. A létrejött akusztikus modell 4630 egyenként 15 Gauss-függvényből álló állapotot tartalmaz. Híranyag felismerési kísérleteink során minden esetben ezt az akusztikus modellt használtuk. A lexikai elemek fonetikus átírását a magyar nyelv hasonulási tulajdonságait figyelembe vevő, automatikus eljárással készítettük.

3.2 Nyelvi modellek

A felismerési tesztjeinkben használt összes nyelvi modell módosított Kneser-Ney simítás használatával készült az SRI Language Modeling Toolkit (SRILM) [14] segítségével. Az interpolált nyelvi modellek készítéséhez és optimalizálásához az SRILM beépített lineáris interpolációs és perplexitás számító eljárásait használtuk. A nyelvi modellek fokszámát minden modell esetén egyedileg optimalizáltuk.

3.2.1 Morfszegmentálás

A morfémákban gazdag nyelvek esetén (pl. magyar, finn, török, észt, stb.) visszatérő probléma szóalapú nyelvi modellezés (**WORD**) esetén a nagy szótárméret, az ebből fakadó adatelégtelenségi problémák, illetve az a tény, hogy még sok tanítóadat esetén is nagy lehet a szótáron kívüli szavak (Out of Vocabulary – **OOV**) aránya a felisme-

résí feladatban. Gyakori megközelítés a probléma enyhítésére, hogy a szótárat kisebb, ám gyakrabban előforduló elemekre darabolják fel, így csökkentve a szótárméretet és növelve a tanítóminták számát. Kísérleteinkben a szótári elemeket egy elterjedten alkalmazott felügyelet nélküli szegmentáló eljárással [5] ún. **morfokra** bontjuk:

„utalt az okra az uniós alapelv ekre amely eket a tagállam oknak tisztelet ben kell tartani uk”

3.2.2 Szóhatár-visszaállítás

A gépi beszédfelismerő kimenetén szóhatárok mentén szegmentált szöveget várunk, ezért a morfalapú rendszer tanítószövegében valamilyen módon jelölnünk kell azokat. Erre kétféle technikát alkalmazunk. Az első ún. **szóhatár-jelölő morfos** (word boundary tag – **WB**) megközelítésnél egy dedikált szimbólumot használunk a különböző szavakhoz tartozó morfok elválasztásához:

„utalt <w> az okra <w> az <w> uniós <w> alapelv ekre <w> és <w> jogszabály okra <w> amely eket <w> a <w> tagállam oknak <w> tisztelet ben <w> kell <w> tartani uk”

A módszer előnye, hogy mindössze egyetlen plusz szótári elem bevételével kezelni tudjuk a szóhatár-visszaállítás problémáját. Hátrány viszont, hogy ez a szimbólum nagyon gyakori elemé válik, és így rontja az n-gram modell predikciós képességét. A másik megoldás az ún. **követőmorfos** (Non-initial morph – **NI**) jelölés, azaz amikor minden olyan morfot megjelölünk a szövegben, mely nem az első tagja egy szónak:

„utalt az -okra az uniós alapelv -ekre és jogszabály -okra amely -eket a tagállam -oknak tisztelet -ben kell tartani –uk”

Előnyös tulajdonsága a módszernek, hogy kevesebb jelölésre van szükség, mivel a szavak többsége egyetlen morfból áll. Hátrány azonban, hogy jelentős mértékben megnőhet a szótárméret a szóhatár-jelöléses módszerhez képest, hiszen az egyes morfok követőmorfos és szó eleji alakjait meg kell különböztetni egymástól. Igaz, a szóalapú rendszerhez képest még így is jelentős lehet a szótárméret csökkenése.

3.3 Hálózatépítés és dekódolás

A 16 kHz-en mintavételezett felvételek lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre, és ún. vak csatornakiegyenlítő eljárást is alkalmaztunk. A súlyozott véges állapotú átalakítókra (Weighted Finite State Transducer – **WFST**) épülő felismerő hálózatok generálását és optimalizálását az Mtool keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztéshez a VOXerver [6] nevű WFST dekódert használtuk. A cikkünkben összehasonlított szó- és morfalapú rendszerek futásidejében keletkező különbségeket minden esetben kiegyenlítettük a keresési szélesség hangolásával. A felismerő rendszerek teljesítményének értékeléséhez szóhiba-arányt (Word Error Rate – **WER**) illetve néhány esetben betűhiba-arányt (Letter Error Rate – **LER**) számoltunk.

4 Kísérleti eredmények a híradó adatbázison

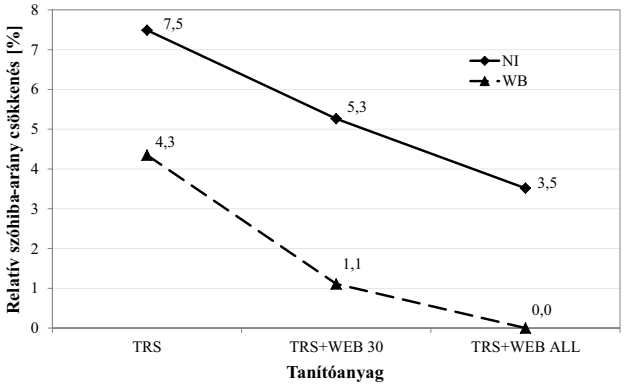
4.1 Morfalapú hibacsökkenés a tanítószöveg méretének függvényében

Első kísérletünk célja a morfalapú nyelvi modellekkel elérhető felismerési hibaarány-csökkenés és a tanítóadat-mennyiség közötti összefüggés vizsgálata volt. Korábbi munkáinkban [7], [8] arra jutottunk, hogy a morfalapú módszerek előnye a tanítószöveg méretének növekedésével egyre csökken, sőt bizonyos méret fölött teljesen el is tűnik [6]. Fontos megjegyezni ugyanakkor, hogy a fent idézett cikkeinkben csak a szóhatár-jelöléses (WB) megközelítést vizsgáltuk. Mostani összevetésünkben három mérési pontot alkalmazunk. Az első esetben csak a kézi leiratok (**TRS**) alapján tanítjuk a modelleket. Második esetben a TRS és WEB 30+ korpuszok alapján (**TRS+WEB 30**), míg a harmadik esetben a TRS, a WEB 30+ és WEB 30- korpuszokat is felhasználjuk a tanításhoz (**TRS+WEB ALL**). A nyelvi modellek interpolációs súlyát a tesztanyag Dev halmazán optimalizáltuk.

3. táblázat: Felismerési eredmények különböző méretű tanítószövegekkel

Tanító-anyag	Lexikai modell	N-gram fokszám	Szótár-méret [ezer szó]	Dev WER [%]	Eval WER [%]	Rel. WER csök. [%]
TRS	WORD	3	74	38,2	41,4	
	WB	4	12	35,2	39,6	*-4.3*
	NI	4	16	34,8	38,3	*-7.5*
TRS+ WEB 30	WORD	3	151	32,4	36,1	
	WB	4	24	31,1	35,7	*-1.1*
	NI	4	31	30,9	34,2	*-5.3*
TRS+ WEB ALL	WORD	4	978	23,0	25,6	
	WB	5	175	23,1	25,6	*0.0*
	NI	4	204	22,3	24,7	*-3.5*

A felismerési eredményeket a **3. táblázatban** foglaltuk össze. A morfalapú rend-

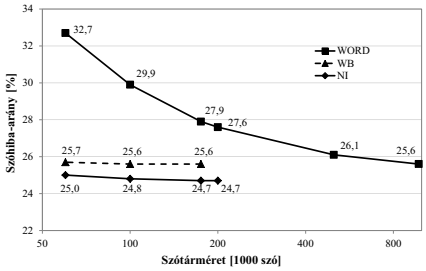


1. ábra. A különböző méretű tanítószövegek alapján tanított morf nyelvi modellekkel elérhető relatív szóhiba-arány csökkenés

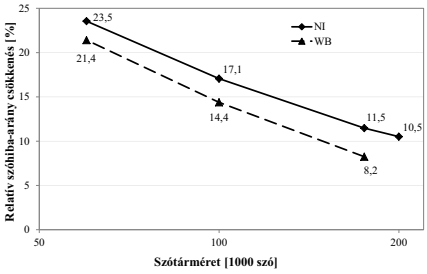
szerekkel a szóalapú rendszerhez képest elérhető relatív szóhiba-arány csökkenéseket az **1. ábrán** mutatjuk be. A WB típusú morfmodell – korábbi eredményeinkkel összhangban – a tanítószöveg méretének növekedésével elveszti az előnyét a szóalapú rendszerhez képest. Ellenben az NI típusú modell még a legnagyobb modellméret mellett is 3,5%-kal jobban teljesít. Igaz a tendencia itt is arra utal, hogy idővel elvész az előny. Érdemes lehet a jövőben még az eddigieknél is nagyobb tanítószöveget bevonni a vizsgálatainkba.

4.2 Morfalapú hibacsökkenés a szótárméret függvényében

Minden korábbi vizsgálatunkban teljes szótárméret mellett hasonlítottuk össze a szó- és morfalapú nyelvi modelleket. Döntésünknek az volt az oka, hogy úgy éreztük, aránytalanul nagy előnyt élveznének a morfalapú megközelítések, ha szóalapú rendszer szótárméretét velük megegyező szintre csökkentenénk. Ezzel szemben számos tanulmányban [15], [16] kiegyenlített szótárméret mellett mérik a hibaarány csökkenést. Itt két szemlélet ütközik. Az egyik szerint a rendszereket nem csak azonos számításigény, de azonos memóriaigény mellett kell vizsgálni. A memóriaigényt azonban csupán a szótárérettel nem lehet kézben tartani, így adott tanítószöveg esetén érdemesebb megpróbálni kihozni a maximumot a vizsgált modellezési technikából. Elfogadva mindkét szemlélet létjogosultságát célunk az volt, hogy kimutassuk a kettő



2. ábra. A különböző szótárméretű nyelvi modellekkel mért felismerési hibák



3. ábra. Relatív szóhiba-arány csökkenés a szótárméret függvényében

közötti különbséget. Méréseinket 60, 100, 175, 200, 500 és 978 ezres szótárméret mellett végeztük a TRS+WEB ALL nyelvi modellek felhasználásával.

A **2. ábrán** jól kivehető, hogy míg a szóalapú nyelvi modellek felismerési hibája erősen függ a szótárérettől, addig a morfalapú modellek az általunk vizsgált 60 ezres határig nagyjából érzéketlenek rá. Ebből következik az is, hogy a szótárméret csökkentésével szignifikánsan nagyobb morfalapú hibaarány-csökkenést mérhetünk, mintha az egyes modelleket teljes szótárméret mellett hasonlítanánk össze (**3. ábra**). Nem meglepő módon a követőmorfos technika ebben a kísérletben is őrizte az előnyét a szóhatár-jelöléses megközelítéshez képest.

4.3 Morfalapú hibacsökkenés a tervezettség és a jel-zaj viszony függvényében

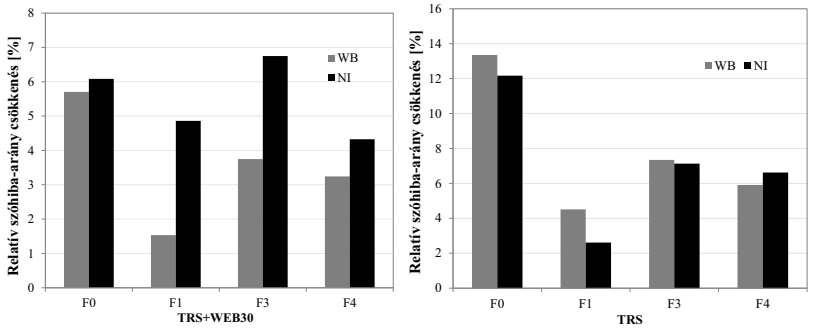
Mint a 2.2-es fejezetben ismertettük, rendelkezésünkre áll a tesztanyag akusztikai viszonyok és tervezettség szerinti felbontása is, melynek köszönhetően a morfalapú

hibacsökkenés mértékét vizsgálhatjuk e paraméterek tekintetében is (**4. táblázat**). Az F0, F2, F3 és F4 kategóriák összehasonlításával képet kaphatunk arról a jelenségről, melyre már a bevezetőben is utaltunk. A tiszta, tervezett beszéd (F0) felismerése még kevés szöveges tanítóadat esetén is tolerálható hibával jár (~30%), növelve a tanítószöveg mennyiségét viszont nagyon alacsony szinte is csökkenthető (~16%). Ez a hibaarány például már lehetővé teszi egy jól érhető felirat készítését a televíziós anyagokhoz. Látható azonban, hogy a jel-zaj viszony csökkenésével (F3, F4) jelentősen megnő a hibák száma. 10-15 dB-es átlagos jel-zaj viszony mellett már 10-15%-kal magasabb hibát kapunk. Telefonos beszéd esetén (F2) nehézséget jelent az alacsony spektrális sávszélesség, így nem véletlen, hogy erre a feladatra külön akusztikus modellt szokás tanítani [17]. Nem csak az akusztikus körülmények játszanak azonban fontos szerepet. A hibaarány még magas jel-zaj viszony esetén is megnőhet, ha tervezett helyett spontán vagy félig spontán (F1) beszédet írunk át, melynek oka a szavak nehezebb predikálhatóságában és a lazább artikulációban keresendő [7].

4. táblázat: Felismerési eredmények F-kategóriák szerinti felbontásban

Tanító-anyag	Lexikai modell	Dev+Eval WER [%] F0	F1	F2	F3	F4	F5
TRS	WORD	33.7	42.1	76.0	46.3	42.3	55.4
	WB	29.2	40.2	67.4	42.9	39.8	54.7
	NI	29.6	41.0	67.1	43.0	39.5	52.2
TRS+ WEB30	WORD	26.3	39.1	67.1	40.0	37.0	49.5
	WB	24.8	38.5	60.5	38.5	35.8	48.7
	NI	24.7	37.2	64.7	37.3	35.4	47.3
TRS+ WEB ALL	WORD	15.6	31.5	56.2	28.7	27.5	38.2
	WB	15.9	29.7	53.5	30.7	27.4	35.8
	NI	15.6	29.5	56.6	28.1	26.6	37.6

A számszerű felismerési hibák mellett érdemes megvizsgálni a morfalapú módszerekkel nyerhető hibacsökkenést is. Ennél a vizsgálatnál az F2 és F5 kategóriákat figyelmen kívül hagytuk, ugyanis a tesztanyag csak nagyon kis része tartalmaz ilyen mintákat (**1. táblázat**). Érdemi következtetéseket csak az F0 és F4 kategória eredményeiből érdemes levonni, mivel ezek a kategóriák képezték a tesztanyag legnagyobb

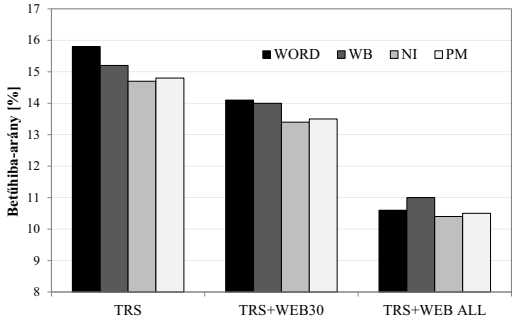


4. ábra. A morfalapú módszerekkel nyerhető relatív szóhiba-arány csökkenés a tervezettség és a jel-zaj viszony függvényében

részét. A **4. ábrán** megfigyelhető, hogy a tiszta, tervezett felvételeken (F0) jelentősen nagyobb a morfalapú hibacsökkenés, mint az alacsony jel-zaj viszonnyal rendelkezőkön (F4). Hasonló eredményre jutottunk korábban, amikor az akusztikus modell illeszkedésének hatását vizsgáltuk [8]. A jelenség oka abban keresendő, hogy morfok jellemzően a szavaknál kevesebb fonémából állnak, így akusztikailag könnyebben összetéveszthetőek. A jobb jel-zaj viszony és akusztikus modell segíti kiemelni a morfalapú nyelvi modell előnyeit. Részben ugyanerre vezethető vissza az F0 és F1 kategóriák közötti különbség is. Itt a jel-zaj viszony megegyezik, azonban a tervezett beszédhez jobban illeszkedik az akusztikus modell, illetve a morfalapú nyelvi modell.

4.4 Szóhatár-jelölés hatása a felismerési hibára

A híradó felismerő rendszerünk kiértékelésének utolsó lépésében azt vizsgáltuk, hogy a morfalapú nyelvi modellekben használt szóhatár-jelölésnek milyen hatása van a felismerés pontosságára. Hipotézisünk szerint ezek csökkentik a morfalapú modellezés hatékonyságát, ám használatuk elkerülhetetlen a szóhatár-visszaállítás miatt. A rendszerek összehasonlításához itt természetesen szóhiba-arány mérése nem jöhetett szóba, ezért betűhiba-arányokat mértünk (**5. ábra**). Meglepő módon előzetes feltevé-



5. ábra. Betűhiba-arányok a híradó Eval tesztadatbázison szóalapú, morfalapú és szóhatár-jelölés nélküli morfalapú nyelvi modellekkel

sünkkel ellentétes eredményt kaptunk. Bár a szóhatár-jelölés nélküli morfalapú rendszer (Pure morph – **PM**) minden tanítókorpusz méret mellett jobban teljesített, mint a szóalapú és WB morf modellezés, az NI morf megközelítést azonban nem tudta túlszárnyalni. Ebből azt a következtetést vonhatjuk le, hogy a szóhatár-jelölés csupán a WB technika esetén tekinthető szükséges rossz megoldásnak, az NI esetén még javítja is a modellezés pontosságát. Ez magyarázhatja tehát, hogy az NI megközelítés minden esetben alacsonyabb felismerési hibát eredményez, mint a WB. Meg kell jegyeznünk azonban, hogy a morfalapú modellek összevethetőségét valamelyest nehezíti, hogy a szóhatár-jelölés nélküli morfmodellben minden morf elejére helyeztünk egy opcionális szünetmodellt a kiejtési modellben, míg a szóhatár-jelöléses modelleknél (WB, NI) csak a lehetséges szóhatárokra.

5 Kísérletek a morfalapú hibacsökkenés előrejelzésére

Az előző fejezetben híradó felismerési feladaton mutattuk be a tanítókorpusz és a szótár méretének, valamint a jel-zaj viszonynak és a beszéd tervezettségének hatását a morfalapú hibacsökkenésre. Ebben a fejezetben az eddigi eredményeket kiegészítjük három különböző ügyfélszolgálati adatbázis eredményeivel, és kísérletet teszünk a hibacsökkenés mértékét előre jelezni. Egy korábbi cikkünkben [7] abból a feltételezésből indult ki, hogy szóalakok száma összefügg a feladat tervezettségével. Ezzel a megközelítéssel összefüggést találtunk egy adott méretű tanítószövegben előforduló szóalakok száma és a relatív szóhiba-arány csökkenés között. Láthattuk azonban, hogy a tervezettségen kívül más tényezők is szerepet játszanak. Jelenlegi vizsgálatunkban a tanítószöveg méretének hatását is megpróbáltuk figyelembe venni. Minden nyelvi modellben teljes szótárat használtunk, míg az akusztikai viszonyok hatását úgy igyekeztünk kiküszöbölni, hogy a kísérleteket beszélőfüggetlen akusztikus modellel végeztük. Az ügyfélszolgálati felismerések eredményei a Magyar Telefonos Ügyfélszolgálati Beszédadatbázisok (**MTUBA**) kiértékelésből származnak. Az MTUBA II. [18] biztosítási, az MTUBA III. távközlési, míg az MTUBA IV. [17] banki ügyfélszolgálati telefonbeszélgetések rögzítéséből származik. Az adatbázisokkal kapcsolatban bővebb információk a megjelölt hivatkozásokban találhatóak. A tanítószövegek adatait és a felismerési eredményeket az **5. táblázatban** ismertetjük.

5. táblázat: Ügyfélszolgálati felismerési eredmények

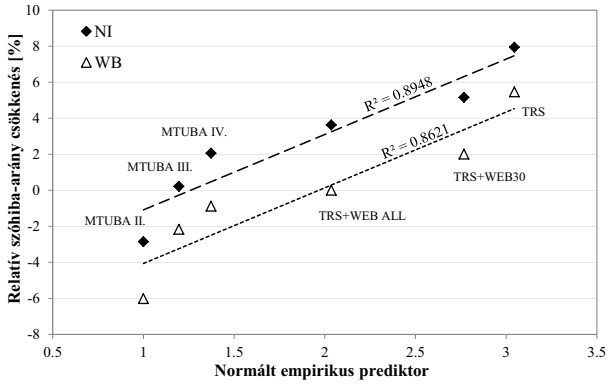
Tanító-anyag	Lexikai modell	$Tokens(T)$	$\frac{1}{n}\sum_{i=1}^{n} Types(T_i)$	WER [%]	Rel. WER csök. [%]
MTUBA II.	WORD	487	10938	31,6	
	WB			33,5	*-6.0*
	NI			32,5	*-2.8*
MTUBA III.	WORD	1313	14064	46,1	
	WB			47,1	*-2.2*
	NI			46,0	*0.2*
MTUBA IV.	WORD	794	15576	34,0	
	WB			34,3	*-0.9*
	NI			33,3	*2.1*

A morfalapú hibacsökkenés előrejelzéséhez először felbontjuk a tanítószövegeket *n* db, diszjunkt, egyenként *k* szót tartalmazó részkorpuszra (T_i). A részkorpuszok mérete kísérletünkben $k = 160000$.

$$\bigcap_{i=1}^{n} T_i = \emptyset \qquad Tokens(T_i, 0 \le i \le n) = k$$

Az empirikus prediktort ez alapján úgy számoljuk, hogy meghatározzuk az egyes részkorpuszokon mért szóalakszámok átlagát és elosztjuk a teljes tanítószöveg (T) szószámának logaritmusával.

$$Emprikus\ prediktor = \frac{\frac{1}{n}\sum_{i=1}^{n} Types(T_i)}{\log_{10} Tokens(T)}$$



6. ábra. Összefüggés a morfalapú rendszerekkel kapható szóhiba-arány csökkenés és a csökkenés mértékét előrejelző empirikus mérőszám között

A képletben új elemként a nevezőt azonosíthatjuk, melytől azt várjuk, hogy képes kompenzálni a hibacsökkenés függését a tanítószöveg méretétől. Az ügyfélszolgálati és híradós adatbázisokon mérhető összefüggést a prediktor és a morfalapú hibacsökkenés között a **6. ábrán** mutatjuk be. Mint az ábrán látható sikerült egy olyan immáron a tanítószöveg méretet is figyelembe vevő mértéket bevezetnünk, mely erősen korrelál a WB (R^2=0.86) és az NI morfalapú (R^2=0.89) megközelítésekkel kapható relatív szóhiba-arány csökkenéssel.

6 Összefoglalás

Cikkünkben arra kerestük a választ, hogy milyen tényezőktől függ a morfalapú nyelvi modellezéssel, a szóalapú rendszerekkel szemben elérhető hibacsökkenés a nagyszótáras gépi beszédfelismerésben. Televíziós híradók adatbázisán végzett kísérleteink rámutattak, hogy a tanítószöveg méretének növekedésével csökken az adatelégtelenség, így a morfalapú rendszerekkel nyerhető előny is. Megállapítottuk, hogy a morfalapú nyelvi modellek kisebb, tömörebb szótáruk miatt keveset veszítenek pontosságukból még akkor is, ha erősen korlátozzuk a szótár méretét. Azaz a morfalapú rendszerek használata kevés és korlátozott szótárméretű tanítószöveg esetén lehet különösen előnyös. Számszerűsítettük továbbá a jel-zaj viszony romlásának hatását, illetve azt is megmutattuk, hogy a közlés tervezettségnek növekedése, hogyan növeli a várható hibacsökkenést.

Az utolsó fejezetben bevezettünk egy mérőszámot, mely erős összefüggést mutat három telefonos ügyfélszolgálati és a híradó adatbázison mérhető morfalapú hibacsökkenésekkel. Ez mérőszám jelenleg csak a feladat tervezettségét és tanítószöveg mennyiségét veszi figyelembe, így a jövőben szeretnénk továbbfejleszteni oly módon, hogy az akusztikus körülményeket is számításba vegye. Emellett, de ettől nem teljesen függetlenül elméleti magyarázatot is szeretnénk találni a morfalapú hibacsökkenést befolyásoló egyes tényezők közötti összefüggésekre.

Köszönetnyilvánítás

Köszönettel tartozunk Balog Andrásnak a híradó tesztanyag F-kategória szerinti válogatásáért, illetve Sárosi Gellértnek az ügyfélszolgálati rendszerek kiértékeléséért. Kutatásunkat a Mindroom (KMOP-1.1.3-08/A-2009-0006), WEBRA TIME SAVE (GOP-1.1.1-11-2012-0377), FuturICT.hu (TÁMOP-4.2.2.C-11/1/KONV-2012-0013) és DIANA (KMR_12-1-2012-0207) projektek támogatták.

Hivatkozások

1. N. Morgan, J. Cohen, S. H. Krishnan, S. Chang, and S. Wegmann, “Final Report : OUCH Project (Outing Unfortunate Characteristics of HMMs),” 2013.
2. P. Mihajlik, Z. Tuske, B. Tarján, B. Németh, and T. Fegyó, “Improved Recognition of Spontaneous Hungarian Speech—Morphological and Acoustic Modeling Techniques for a Less Resourced Task,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 18, no. 6, pp. 1588–1600, Aug. 2010.
3. M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, “Unlimited vocabulary speech recognition for agglutinative languages,” in HLT-NAACL 2006, 2006, pp. 487–494.
4. E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, “Turkish Broadcast News Transcription and Retrieval,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 17, no. 5, pp. 874–883, Jul. 2009.
5. M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0,” in Publications in Computer and Information Science, Report A81, 2005.
6. B. Tarján, P. Mihajlik, A. Balog, and T. Fegyó, “Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection,” in 2nd International Conference on Cognitive Infocommunications (CogInfoCom), 2011, pp. 1–5.
7. B. Tarján and P. Mihajlik, “On morph-based LVCSR improvements,” in Spoken Language Technologies for Under-Resourced Languages (SLTU-2010), 2010, pp. 10–16.
8. L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, “Speech Recognition Experiments with Audiobooks,” Acta Cybern., vol. 19, no. 4, pp. 695–713, 2010.
9. B. Tarjan, T. Mozsolics, A. Balog, D. Halmos, T. Fegyo, and P. Mihajlik, “Broadcast news transcription in Central-East European languages,” in IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), 2012, pp. 59–64.
10. L. Tóth and T. Grósz, “A Comparison of Deep Neural Network Training Methods for Large Vocabulary Speech Recognition,” pp. 1–8, 2012.
11. A. Roy, L. Lamel, T. Fraga, J. Gauvain, and I. Oparin, “Some Issues affecting the Transcription of Hungarian Broadcast Audio,” in 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), 2013, no. August, pp. 3102–3106.
12. C. Kim and R. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in INTERSPEECH, 2008, pp. 2598–2601.
13. S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, The {HTK} Book, version 3.4. Cambridge, UK: Cambridge University Engineering Department, 2006.
14. A. Stolcke, “SRILM – an extensible language modeling toolkit,” in Proceedings International Conference on Spoken Language Processing, 2002, pp. 901–904.

15. V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in Proc. Eurospeech’03, 2003, pp. 2293–2296.
16. M. A. Basha Shaik, A. El-Desoky Mousa, R. Schlüter, and H. Ney, “Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR,” in Interspeech, 2011, pp. 1441–1444.
17. B. Tarján, G. Sárosi, T. Fegyó, and P. Mihajlik, “Improved Recognition of Hungarian Call Center Conversations,” in The 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), 2013, pp. 65–70.
18. G. Sárosi, B. Tarján, T. Fegyó, and P. Mihajlik, “Automated Transcription of Conversational Call Center Speech – with Respect to Non-verbal Acoustic Events,” Intell. Decis. Technol.