Spontán, nagyszótáras, folyamatos beszéd gépi felismerési pontosságának növelése beszélőadaptációval a MALACH projektben

Tüske Zoltán[1], Mihajlik Péter[1], Fegyó Tibor[1,2]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
{tuske, mihajlik}@tmit.bme.hu

[2]Aitia International
tfegyo@aitia.ai

Kivonat: Cikkünkben bemutatjuk, hogy az MLLR (Maximum Likelihood Linear Regression) alapú beszélőadaptálás során a beszédfelismerési hatékonyság az adott spontán magyar nyelvű adatbázison jelentősen növekszik. Többféle módszert kipróbáltunk mind a felügyelt mind a felügyeletlen adaptálódás esetén is. A globális megoldás mellett regressziós osztályokon alapuló transzformációt is alkalmaztunk; felügyeletlen modellillesztés esetén a többszörös adaptálást is megvizsgáltuk. Továbbá folyamatos, nagyszótáras és spontán automatikus beszédfelismerővel kapott eredményekkel támasztjuk alá, hogy ha a szó alapú nyelvi modell helyett a magyar nyelvet pontosabban leíró morféma alapú modellezést alkalmazzuk, akkor a beszélőadaptálás által okozott javulás még szignifikánsabban jelentkezhet a felismerési hibaarányban.

1 Bevezetés

A nagyszótáras folyamatos beszédfelismerő rendszereket (LVCSR: Large Vocabulary Continuous Speech Recognition) nagyszámú beszélőtől összegyűjtött adatokon tanítják, hogy a tanító halmaz lefedje a különböző dialektusokat és beszédstílusokat. Az így nyert - általánosan használt rejtett Markov-modell (HMM: Hidden Markov-Modell) alapú - beszélőfüggetlen rendszer átlagosan jól teljesít minden beszélő esetén. Hátránya, hogy az átlagos beszédet modellezi, és az egyes beszélőkre nem optimális ez a modell. Kézenfekvő megoldás lehet, hogy a jobb felismerési pontosság elérése érdekében, adott beszélőre tanítsunk be egy teljesen új felismerőt. A baj csak az, hogy adott beszélőtől ehhez szükséges mennyiségű adatot szerezni meglehetősen nehéz feladat. A megoldás, hogy beszélőfüggetlen felismerőt az adott beszélőtől származó, a tanítóhalmazhoz képest kevés adatokkal nem újratanítjuk, hanem adaptáljuk, az egyes beszélőhöz „igazítjuk” az akusztikus modelleket. Ezzel egy köztes felismerőt kapunk, ami jobban teljesít, mint egy beszélőfüggetlen felismerő, de a felismerési hatásfok elmarad attól, amit egy elegendően sok tanítóadattal kapott egy beszélős tanítással érhetnénk el. Az adaptáló adatok mennyiségétől függően az adaptált rendszer valahol a két felismerő között helyezkedik el. Ennek a módszernek egyik

legnagyobb előnye a vele elérhető jobb szóhibaarány mellett, hogy attól a beszélőtől, amelyikhez illeszteni akarjuk a rendszert, nagyságrenddel kevesebb adat kell ahhoz képest, mintha egy teljesen új felismerőt tanítanánk be.

A szakirodalomban különböző módszereket találhatunk a rejtett Markov-modellek adaptálására. A beszédfelismerés területén publikációk sora bizonyítja, hogy a modellparaméterek adaptálóadatokra nézve ML (Maximum Likelihood) értelemben optimális lineáris transzformációja (MLLR) igen hatékony megoldás a modellillesztés megvalósítására.

A módszer áttekintése után, a kísérletekhez alkalmazott adatbázis, majd az összeállított felismerők ismertetése, végül az elért eredmények bemutatása következik.

2 Adaptálás MLLR-val

Lényege, hogy az adaptáló adatok alapján az akusztikus modellek minden egyes Gauss összetevőjének kovariancia mátrixát és várhatóérték vektorát lineáris transzformációval - utóbbit eltolással is - módosítja, így az adaptáló adatokhoz jobban illeszkedő akusztikus modelleket kaphatunk. A transzformált Markov-modell a kiinduló modellhez képest nagyobb valószínűséggel állítja elő az adaptáló adatokat. Minden állapot, minden komponensének várhatóértékére:

$$\underline{\hat{\mu}} = \underline{\underline{W}}\,\underline{\xi}$$

Ahol $\underline{\hat{\mu}}$ az új várhatóértékvektor, $\underline{\xi} = \begin{bmatrix} 1 & \underline{\mu}^T \end{bmatrix}^T$, $\underline{\mu}$ a régi várhatóértékvektor és $\underline{\underline{W}}$ *n*(n+1)* méretű egy állapothoz és annak egy komponenséhez tartozó transzformációs mátrix, *n* a jellemzővektorok dimenziója. Hasonlóan a kovariancia mátrixot is transzformálni kell a jobb illeszkedés érdekében.

$$\underline{\underline{\hat{\Sigma}}} = \underline{\underline{H}}\,\underline{\underline{\Sigma}}\,\underline{\underline{H}}^T$$

Ahol $\underline{\underline{\hat{\Sigma}}}$ az új kovariancia mátrix és $\underline{\underline{H}}$ a transzformációs mátrix.

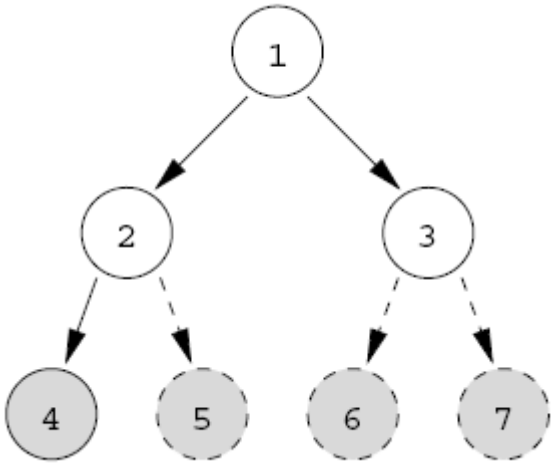
Az ideális transzformációs mátrixok, iteratív úton, EM (Expectation Maximalization) algoritmus segítségével határozódnak meg: először csak a várhatóérték transzformáció rögzített szórásmátrix mellett, majd csak a szórásmátrix transzformáció rögzített várhatóérték mellett. Az eljárás többszöri alkalmazásával egyre nagyobb valószínűséggel, pontosabban likelihood értékkel, fedi le az adaptáló adatokat a transzformált modell.

$$\ell\left(O|\tilde{\lambda}\right) \geq \ell\left(O|\hat{\lambda}\right) \geq \ell\left(O|\lambda\right)$$

Ahol O az adaptáló adatokból származó megfigyelési vektorok, ℓ a teljes likelihood érték adott λ modell paraméter mellett, O megfigyelés esetén. λ a transzformálatlan, $\hat{\lambda}$ a várhatóértékben transzformált, $\tilde{\lambda}$ pedig a két lépésben transzformált modellparamétereket jeleni adott Markov-modell esetén. A transzformációs paraméterek meghatározására vonatkozóan [1, 7] tartalmaz további részleteket.

Az egyes gauss komponensek transzformációs mátrixának meghatározására általában nem áll elegendő adat a rendelkezésre, ezért valamilyen távolság definíció alkalmazásával az egymáshoz közel eső komponensekből csoportokat vagy osztályokat képeznek, és erre a közös halmazra számoljuk a várhatóérték- és a szórás-transzformációs mátrixot. Az akusztikus modell gauss komponenseinek csoportosítására általánosan használt módszer a bináris regressziós fa. Ez a felülről lefele haladó módszer - kiindulásként az összes komponens egy csoportban van - azon komponenseket igyekszik egy osztályba gyűjteni, amelyek az akusztikus térben Euklédeszi távolság értelemben közel vannak egymáshoz, hasonló térrészt jellemeznek. A bemenő adat a beszélőfüggetlen felismerő akusztikus modellje, eredményként a komponensek általunk előírt számú csoportokra bontását adja vissza. A bináris fa továbbépítésének művelete a következő lépésekből áll:

- A felosztandó csoport szórását és várhatóértékét a benne szereplő komponensek alapján kiszámoljuk
- Két gyermek csoportot hozunk létre, a szülő várhatóértékétől ellentétes irányban eltávolodva.
- A szülő osztályt alkotó komponenseket a várhatóértékük alapján a közelebbi gyermek osztályhoz rendeljük.
- Újra számoljuk a gyermek csoportok várhatóértékeit
- A szülő komponenseket újra és újra hozzárendeljük a gyermek osztályokhoz, az újabb és újabb gyermek csoport várhatóértékek alapján, míg nincs már változás a hozzárendelésket illetően.



1. Ábra: Csoportosítás regressziós bináris fával. Adaptáló adatok mennyiségétől függően más-más szinten számoljuk transzformációs mátrixokat.

A regressziós fa segítségével az adaptáló adatok mennyiségétől függő adaptálást alkalmazhatunk (*1. ábra*). Kevés adat esetén csak globális transzformációra van lehetőség, ilyenkor egy mátrixszal transzformálunk minden várhatóértéket, illetve egy másikkal az összes szórást. Ha elegendően sok adatunk van, akkor külön transzformációt határozhatunk meg az egyes csoportokra (az ábrán csak a 4-es). Ha azonban egy csoportba kevés adaptáló adat jut (5, 6, 7), a fán feljebb lépve összevonhatunk több osztályt és így érhetjük el a szükséges mennyiségű adatok meglétét (2, 3).

Az adaptációhoz szükséges szöveges átiratok, amik segítségével az adaptáló adatokat az egyes akusztikus modellekhez rendelhetjük, kétféle forrásból származhatnak. Az adaptálás felügyelt, ha a pontos szöveges tartalommal rendelkezünk. Felügyeletlen adaptálás mellett szöveges átirathoz csak a beszélőfüggetlen felismerőtől kapott felismerési eredményt használhatjuk fel, nem tökéletes az átiratunk, így a szegmentálás sem lesz az. Ilyenkor azonban általában sokkal több adattal adaptálódhatunk, mert az összes felismertetendő felvételt bevonhatjuk a transzformációs mátrixok becslésébe.

3 Az adatbázis

A MALACH (Multilingual Access to Large Spoken Archives) projekt célja, hogy adott esetben holokauszt-túlélők beszámolóihoz könnyebben lehessen hozzáférni, és az adatbázis szöveges tartalom alapján kereshetővé váljon. A projekt 32 nyelvű, amiből a magyar nyelvű adatbázis több mint 2000 órányi felvételt tartalmaz. Eddig mintegy 31 órányi anyagnak készült el a szöveges átirata. A kísérleteket ezen az adatbázisrészleten végeztük. A felvételek 44,1 kHz-es mintavételezés mellett, általános

környezetben (általában a beszélő lakásán) készültek. A beszélők idősek, és esetenként erős akcentussal beszélnek.

A beszélőfüggetlen felismerőt 104 beszélőtől, beszélőnként tizenöt percnyi felvétellel tanítottuk (összesen kb. 26 óra). Tesztelésre további 10 beszélőtől származó, váltakozó hosszúságú, összesen 5 órányi hanganyagot használtunk. Ezt a teszthalmazt további részhalmazokra bontottuk. A beszélőfüggetlen (SI) részhalmaz egyik felét – a 15. perc utáni bemondások – a tanuló halmaz szempontjából ILLESZTETTnek neveztük, míg az első tizenöt perc felvételei az ILLESZTETLEN részhalmazba kerültek. A beszélőfüggetlen eredményeket az *1. táblázat*ban foglaltuk össze. Egy másik részhalmaza a teszthalmaznak a beszélőfüggő (SD) részhalmaz, ami egy férfi és egy női alany 1-1 órányi hanganyagát tartalmazza.

1. táblázat: Beszélőfüggetlen felismerési eredmények az SI teszthalmazokon morféma és szó nyelvi modellekkel

Nyelvi model	Illesztetlen		Illesztett	
	WER	LER	WER	LER
Morféma	55.94	28.17	51.07	24.53
Szó	56.01	28.26	50.90	24.39

4 A felismerési tesztek során alkalmazott akusztikus és nyelvi modellek

4.1 Nyelvi modellezés

A hagyományos szó alapú nyelvi modell mellett, a morféma alapú nyelvi modellel is lefuttattuk az adaptálási teszteket. Utóbbi jobban illeszkedik a morfémákban gazdag nyelvekhez, így a magyarhoz is. Az utóbbi esetben statisztikus morfológia elemző segítségével [6] bontottuk a szavakat morfémákra, a felmerülő problémát és azok egy lehetséges megoldását [2, 3] részletezi. Mindkét esetben trigram nyelvi modellt alkalmaztunk, melyeket a SRILM programmal [4] számoltunk.

4.2.a Beszélőfüggetlen akusztikus modellezés.

Az előző fejezetben bemutatott adatbázisból 26 órányi anyaghoz tartozó szöveges adatbázisból automatikus, szabályalapú konverzióval [5] készült el a szavak valamint a morfémák fonetikus átirata, az idegen eredetű és a hagyományos írásmód körébe tartozó szavak megfelelő fonéma sorozatra való átalakítása kivételszótár alapján történt. A mintegy 3000 általánosított, 3 állapotú balról jobbra haladó HMM-lel modellezett akusztikus trifón modelleket PLP jellemzőkből előállított vektorokkal tanítottuk.

4.2.b Beszélőadaptált akusztikus modellezés

Egy beszélő teljes adathalmazának 1/5-ét használtuk felügyelt adaptálásra, 4/5-öd részével teszteltünk. Felügyeletlen adaptálás esetén a felismertetendő 4/5-öd részen kapott beszédfelismerési eredményeket használtuk fel. Mindkét esetben globális adaptálással - ekkor minden gauss-összetevőn azonos a transzformáció -, és 32 levelű bináris regressziós fa építésével kapott csoportokra bontással is mértük az adaptáció utáni szóhibaarányt. Tapasztalataink szerint a gauss-összetevők 32 regressziós csoportba foglalása hozta a nem globális adaptáláskor a legjobb eredményeket. A felügyeletlen modellillesztés esetén többszörös adaptálással, valamint a globális és regressziós fán alapuló adaptálások kombinálásával igyekeztünk még pontosabbá tenni az akusztikus modellt.

5 Eredmények

A *2.* és *3. táblázat*ban a női beszélő adatain a beszélőfüggetlen és a beszélőfüggő felismerővel elért felismerési eredmények láthatóak. Az adaptálási eredmények mellett a relatív %-os javulást is feltüntettük a beszélőfüggetlen esethez képest. A táblázatból egyértelműen látszik, hogy a morfémaalapú nyelvi modell mellett sokkal hatékonyabb a javulás mind relatív, mind abszolút értékben is. Már a kiindulási eredmények is jobbak a morfémaalapú megközelítésnél (abszolút 5%-os szóhibaarány-különbség), ennek ellenére abszolút értékben is sokkal többet használt az adaptáció, tovább növelve a különbséget a szó- és morfémaalapú modellezés között, adaptálás után közel 10%-os a WER különbség. Felügyelt és felügyelet nélküli adaptálás mellett is relatív 28%-ot javított az adaptálás a betűhibaarányon. Kétszeri felügyelet nélküli adaptálással –először globális majd regressziós fás – sikerült a felügyelten adaptált felismerő felismerési hatásfokát elérni.

A férfi beszélő adatain elért felismerési hatásfok javulása az adaptálás hatására a *4.* és *5. táblázat*ban látható. Ennek a beszélőnek az adatain gyengébbek a morféma alapú felismerési eredmények a beszélőfüggetlen felismerővel (az előző beszélőhöz képest kb. átlag abszolút 2%-kal). A morfémaalapú nyelvtan itt is jobban teljesít, mint a szóalapú, az adaptálásnak köszönhetően ennek a beszélőnek is kb. relatív 25%-kal sikerült javítani a betűhibaarányán. A kétszeres adaptációval elért legjobb eredmény picivel elmarad a felügyelt tanítás mögött. A pontatlanabb felismerésből eredően - a második beszélő szempontjából rosszabb az akusztikus modellezés - adaptáció után a két beszélő közötti hibaarány különbség 2%-ról 4%-ra nőtt.

2. Táblázat: Az adaptálás nélkül és az adaptálással elért szóhibaarányok (női beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		WER [%]	LER [%]	WER [%]	LER [%]
beszélő független	nincs	51,86	21,70	46,07	18,79
felügyelten adaptált	globális	47,18	18,33	38,93	14,90
	regressziós fás	45,84	16,87	36,87	13,48
1x felügyelet nélkül adaptált	globális	46,66	17,96	38,62	14,70
	regressziós fás	46,23	17,32	37,56	13,84
2x felügyelet nélkül adaptált	2x globális	46,56	18,05	38,88	14,62
	globális majd regressziós fás	45,99	17,06	36,58	13,48
	2x regressziós fás	46,41	17,23	37,44	13,86

3. Táblázat: Adaptálással a szóhibaarányban elért relatív %-os javulás a beszélőfüggetlen esethez képest (női beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		ΔWER_{rel} [%]	ΔLER_{rel} [%]	ΔWER_{rel} [%]	ΔLER_{rel} [%]
felügyelten adaptált	globális	9,02	15,53	15,50	20,70
	regressziós fás	11,61	22,26	19,97	28,26
1x felügyelet nélkül adaptált	globális	10,03	17,24	16,17	21,77
	regressziós fás	10,86	20,18	18,47	26,34
2x felügyelet nélkül adaptált	2x globális	10,22	16,82	15,61	22,19
	globális majd regressziós fás	11,32	21,38	20,60	28,26
	2x regressziós fás	10,51	20,60	18,73	26,24

4. Táblázat: Az adaptálás nélkül és az adaptálással elért szóhibaarányok (férfi beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		WER [%]	LER [%]	WER [%]	LER [%]
beszélő független	nincs	49,09	23,73	48,00	23,41
felügyelten adaptált	globális	46,31	21,21	43,96	19,78
	regressziós fás	44,44	19,11	40,76	17,54
1x felügyelet nélkül adaptált	globális	46,92	21,81	44,07	20,07
	regressziós fás	43,94	19,67	41,34	18,36
2x felügyelet nélkül adaptált	2x globális	46,65	21,70	43,82	19,99
	globális majd regressziós fás	44,57	20,13	41,55	18,48
	2x regressziós fás	43,69	19,60	40,73	18,16

5. Táblázat: Adaptálással a szóhibaarányban elért relatív %-os javulás a beszélőfüggetlen esethez képest (férfi beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		ΔWER_{rel} [%]	ΔLER_{rel} [%]	ΔWER_{rel} [%]	ΔLER_{rel} [%]
felügyelten adaptált	globális	5,66	10,62	8,42	15,51
	regressziós fás	9,47	19,47	15,08	25,07
1x felügyelet nélkül adaptált	globális	4,42	8,09	8,19	14,27
	regressziós fás	10,49	17,11	13,88	21,57
2x felügyelet nélkül adaptált	2x globális	4,97	8,55	8,71	14,61
	globális majd regressziós fás	9,21	15,17	13,44	21,06
	2x regressziós fás	11,00	17,40	15,15	22,43

A *6. táblázat*ban a két beszélőn elért átlagos legjobb felismerési eredményeket foglaltuk össze. A legjobb felismerési eredményeket morféma alapú modellel kaptuk. Elmondhatjuk, hogy ha nem ismert az adaptáló adatok pontos tartalma, akkor érdemesebb az akusztikus modelleket először csak globálisan adaptálni a félreismerésekkel zajosított szöveges átiraton, kiküszöbölendő a hibásan felismert fonémasorozatokból eredő félretranszformálást. Majd az így nyert pontosabb felismeréssel, most már regressziós fával továbbtranszformálva az akusztikus modelleket, tovább növelhető az illeszkedés.

6. Táblázat: Az átlagos szóhibaarány adaptálással és nélküle, relatív javulás a beszélőfüggetlen esethez képest

A felismerő fajtája	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
	WER [%]	ΔWER_{rel} [%]	WER [%]	ΔWER_{rel} [%]
beszélő független	50,47	-	47,03	-
felügyelten 32 levelű regressziós fával adaptált,	45,14	10,57	38,81	17,48
2x felügyelet nélkül adaptált globális, majd 32 Levelű regressziós fával	45,28	10,29	39,06	16,94

6 Összefoglalás

Elmondhatjuk, hogy beszélőadaptálással sokkal hatékonyabbá tettük a beszédfelismerő-rendszert. Pontosabb felismerési eredmények mellett az adaptálás jobban teljesít, a hagyományos szó nyelvi modell helyett morféma nyelvi modellt alkalmazva kb. másfélszer akkora relatív javulást értünk el a szó-hibaarányban.

Bibliográfia

1. M. J. F. Gales, Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, Computer Speech and Language, Vol. 12, pp. 75-98, 1998
2. Mihajlik, P., Fegyó T., Németh B., Tüske Z., Trón V., Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages – Hungarian ASR for the MALACH Project, TSD 2007, Pilsen, Czech Republik
3. Mihajlik, P., Fegyó, T., Tüske, Z., and Ircing, P., "A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian" – In Interspeech 2007. Antwerp, Belgium, August 27-31, (2007).
4. Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. On Spoken Language Processing, Denver (2002) 901–904
5. Szarvas M., Fegyó, T., Mihajlik, P., and Tatai, P., "Automatic Recognition of Hungarian: Theory an Practice", International Journal of Speech Technology, 3:277-287, December, 2000.
6. 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
7. Young S., Ollason, D., Valtchev, V., and Woodland, P., The HTK Book (for HTK version 3.2.1), Cambridge University Engineering Department, 2002.