Szegedi Tudományegyetem
Mesterséges Intelligencia Kutatócsoport

Posterior valószínűségekre épülő beszédmodellek és alkalmazásuk a magyar nyelvű beszédfelismerésben

PhD-értekezés tézisei

Tóth László

Témavezető:

Dr. Csirik János

Szeged
2006

Bevezetés

A jelenleg használatos beszédfelismerési technológia matematikai alapelvekre épül és csak minimálisan veszi figyelembe az emberi beszédpercepcióra vonatkozó eredményeket. Vannak ugyan próbálkozások ezen ismeretek felhasználására, de ezek a általában csak az előfeldolgozási lépést finomítják, a statisztikai keretrendszert érintetlenül hagyva. A háromállapotú rejtett Markov-modellekre (HMM) épülő, beszédhang-alapú modellezési technológia célszerűségét nagyon kevesen kérdőjelezték meg az elmúlt évtizedben. A fejlesztők az eredmények javítása céljából inkább az adatbázisok méretét növelték és a nyelvi modelleket finomították. Egyre inkább úgy tűnik azonban, hogy a jelenlegi metodológia elérte képessége határait. Talán itt az ideje, hogy az akusztikai modelleken is javítsunk, közelítsük tulajdonságaikat az emberi hallás működéséhez – mindeközben a statisztikai megközelítést is megtartva.

Jelen disszertáció indításaként összegyűjtjük a rejtett Markov-modellek azon jellemzőit, amelyek az emberi beszédpercepció tulajdonságainak ellentmondanak, majd felvázoljuk, hogy ezek helyett mit várnánk el egy fejlettebb megoldástól. Rávilágítunk a beszédjelek statisztikai modellezésének kulcskérdésére, miszerint a mondatokhoz rendelt valószínűségi értékeket valamely kisebb egységek valószínűségéből vagyunk kénytelenek összerakni. A valószínűségszámítás sajnos elég kevés eszközt nyújt egy sokváltozós valószínűségi eloszlás felbontására. Mégis, ezeket a szokványostól eltérő módon alkalmazva különféle, a HMM-től eltérő modelleket vezethetünk le. A disszertációban vizsgált modellek közös jellemvonása, hogy az építőelemek posterior valószínűségét közelítik és azokból építkeznek, míg a HMM az adatok osztályonkénti eloszlását igyekszik leírni. A posterior valószínűségek közelítésére mesterséges neuronhálókat (ANN) alkalmazunk. Ennek különféle előnyei vannak a szokványos – az osztályonkénti eloszlások Gauss-komponensekkel történő – modellezésével szemben.

A rejtett Markov-modelles technológia leginkább vitatható tulajdonsága, hogy a valószínűségek felbontásával egészen a beszédkeretek szintjéig megy le. Ezekről aztán feltételezi, hogy az egyes szegmentumokon belül függetlenek (ez az ún. naív-Bayes feltételezés), s ezért a hozzájuk rendelt valószínűségi értékeket összeszorozza. A disszertációban két olyan modellezési megoldást vizsgálunk meg, amelyek elkerülik ezt a felbontási lépést. Az egyik esetben egyáltalán fel sem bontjuk a szegmentumokat keretekre, hanem egy egységként parametrizáljuk őket. Így jutunk el az ún. szegmentum-alapú modellek családjához, és a disszertáció egy jelentős része ezekkel foglalkozik. Mint kiderül, ennek a reprezentációnak számos jó tulajdonsága van, például sokkal jobbnak bizonyul a fonetikai szegmentumok osztályozásában. Nehézségei vannak viszont a felismerés során a helyes szegmentálás megtalálásával, amely probléma kiküszöbölésére többféle megoldást javasolunk és tesztelünk le.

A szegmentum-alapú módszercsalád gyenge pontjaival szembesülve a másik vizsgált modellben visszatérünk a keretalapú feldolgozáshoz. Azonban, váratlan módon, a keretekhez rendelt valószínűségi értékeket szorzás helyett átlagoljuk. Ez első pillanatra ésszerűtlennek tűnhet, de több érvet is hozunk a használhatóságára. A beszédhang-osztályozási kísérletek is azt igazolják, hogy az értékek átlagolása nem rosszabb megoldás, mint a szorzás. A teljes felismerési feladat megoldására azonban az átlagolós módszer csak akkor képes, ha kiegészítjük egy, a szegmentálás valószínűségét megadó komponenssel – a szegmentum-alapú rendszerrel szerzett tapasztalatokkal összhangban. Az ily módon előálló modellt 'átlagolós HMM/ANN hibridnek' nevezzük el. Ez az új konstrukció mind a beszédhang-felismerési, mint az izolált szavas felismerési feladatokban valamivel jobb teljesítményt nyújt, mint a hagyományos HMM. Az utolsó fejezetben egy picivel még tovább fokozzuk a hibrid modellek képességeit, amikor is kiegészítjük őket egy explicit hosszmodellel, illetve egy újramintavételezés-alapú tanítási sémával.

A disszertáció egyik különlegessége, hogy az összes előforduló modellt szegmentum-alapú nézőpontból vizsgálja. Ezt a látásmódot eredetileg a szegmentális modellekkel való kísérletezés tette szükségessé, de a keretalapú rendszerek elemzésében is nagyon hasznosnak bizonyult. Ami a legfontosabb, ráirányította figyelmünket a szegmentálási részfeladat megoldásának fontosságára. Mind a hagyományos, mind a hibrid HMM modellt megvizsgáljuk ebből a szemszögből, és mindkét esetben arra jutunk, hogy lényegében a keretek valószínűségének szorzással való kombinálása teszi képessé ezeket a rendszereket a beszédjel fonetikai szegmentumainak megtalálására. Maga ez a szemléletmód és az általa nyert betekintés a keretalapú rendszerek működésébe fontosabb eredménynek tűnik, mint a két javasolt újszerű modell.

A disszertációban előforduló felismerési teszteket magyar beszédkorpuszokon hajtottuk végre. Mivel néhány kivételtől eltekintve más még nem végzett hasonló vizsgálatokat ezeken az adatbázisokon, ezért összehasonlítás céljából a HTK nevű rendszerrel állítottunk elő rejtett Markov-modelles eredményeket.

A felbontás problematikája a beszédfelismerésben

A gépi tanulás elméleti megközelítései közül a gyakorlatban a statisztikai alakfelismerés a legsikeresebb. Az ennek alapját képező Bayes döntési szabály szerint a sikeres osztályozáshoz a $P(W|X)$ posterior valószínűség ismerete szükséges a W osztálycímkék és X mérési vektorok összes lehetséges kombinációjára. Sokféle módszer ismert a fenti valószínűség becslésére, és ezek számos gyakorlati területen beváltak. A beszédfelismerés azonban kilóg a szokványos felismerési feladatok közül olyan értelemben, hogy a lehetséges mérési adatok (szünet nélkül kimondható beszédjelek) és osztálycímkék (beszédhang- vagy szósorozatok) száma túl nagy, s ezért mindkettőt kisebb egységekre kényszerülünk bontani. A valószínűségi eloszlások becslésére szolgáló technikákat csak ezekre a kisebb egységekre alkalmazzuk, és a teljes jelhez tartozó valószínűséget a kisebb egységekhez rendelt értékekből rakjuk össze. Alapvető kérdés tehát, hogy hogyan választjuk meg ezeket az egységeket, hogyan becsüljük a hozzájuk tartozó valószínűségeket, és hogy ezeket az értékeket hogyan integráljuk egyetlen globális értékké.

A valószínűségszámítás viszonylag kevés (egyszerű) módszert kínál a többváltozós valószínűségek manipulálására és felbontására. Az egyik ezek közül a Bayes szabály, amelynek segítségével megcserélhetjük a változók szerepét egy feltételes valószínűségben. Ennek köszönhetően lehetőségünk van választani, hogy a $P(W|X)$ vagy a $p(X|W)$ értékeket szeretnénk-e modellezni. Mivel az előbbi diszkrét, utóbbi pedig folytonos eloszlás, ezért modellezésük eltérő technikát igényel. Az osztályok $P(W|X)$ posterior valószínűségének közelítésére a mesterséges neuronháló a legnépszerűbb eszköz, míg a $p(X|W)$ osztályonkénti eloszlásokat leginkább Gauss-görbék súlyozott összegével szokás közelíteni. A beszédfelismerésben az utóbbi, ún. generatív modellezés az elterjedtebb.

Egy másik fontos felbontási módszer a rejtett változók bevezetése. A beszédfelismerésben a beszédjel S fonetikai szegmentálását szokás ilyen rejtett változóként kezelni. Feltételezve, hogy a jelnek egyetlen helyes szegmentálása van, alkalmazhatjuk a teljes eseményrendszerekre vonatkozó szabályt, majd ez emiatt bejövő összegzést lecserélhetjük maximalizálásra:

$$p(X|W) = \sum_S p(X,S|W) = \sum_S p(X|S,W)P(S|W) \approx \max_S p(X|S,W)P(S|W), \qquad (1)$$

ahol a második lépésben a láncszabályt alkalmaztuk. Ily módon a maximalizálás révén a felismerési feladat egy W és S fölötti keresési problémává válik.

Egy további felbontási trükk a komponensek függetlenségének feltételezése (más néven a naív Bayes-feltevés). Ezt két szinten is alkalmazhatjuk. Első lépésben a teljes jelet fonetikai szegmentumokra bontjuk, és ezek függetlenségét feltételezve az kapjuk, hogy

$$p(X|S,W) \approx \prod_i p(X_i|u_i) \quad \text{és} \quad P(S|W) \approx \prod_i P(S_i|u_i), \tag{2}$$

ahol X_i az $S_i = (s_{i-1}, s_i)$ szegmentumhatárok által kijelölt adatblokkot jelöli, u_i pedig a szegmentum fonetikai címkéje.

A rejtett Markov-modell a szegmentumokat még tovább bontja. Ehhez az X_i szegmentumot adatkeretek egy $(x_{s_{i-1}}, ..., x_{s_i-1})$ sorozataként reprezentálja, a szegmentumhoz tartozó valószínűséget pedig a naív Bayes-feltevést ismét alkalmazva az alábbi módon közelíti:

$$p(X_i|u_i) \approx \prod_{j=s_{i-1}}^{s_i-1} p(x_j|u_i), \tag{3}$$

míg $P(S_i|u_i)$ becslésére pedig geometriai eloszlással leírt hosszmodellt alkalmaz (az állapotátmenet-valószínűségek segítségével).

Habár a fenti levezetés matematikai szemszögből vonzóan egyszerű, beszédpercepciós szempontból jóformán minden lépése megkérdőjelezhető. A legtöbb kritika a legutolsó lépést ((3)) éri. Már tudjuk például, hogy az emberi agy 'feldolgozási egysége' időben hosszabb és frekvenciában keskenyebb, mint a beszédfeldolgozás adatkeretei [9]. A keretek feltételes függetlenségének feltételezése is sok szempontból vitatható [13]. Harmadrészt, pusztán technikai szempontból az osztályonkénti valószínűségekre való visszavezetés nem feltétlenül a legjobb választás. Sokak szerint a posterior valószínűségekből való építkezés, valamint azok neuronhálókkal való közelítése a gyakorlatban sokkal előnyösebb [4].

A továbbiakban két olyan felbontást vizsgálunk meg, amelyek posterior-alapú felírásokat eredményeznek. Az egyik a szegmentumokat egy egységként kezeli, így a keretekre való szétbontást teljes mértékben elkerüli. A másik keretekből építkezik ugyan, de azok posterior valószínűségét szorzás helyett átlagolással kombinálja.

A posterior-alapú szegmentális beszédmodell

A szegmentum-alapú modellek bevezetését a (3) alatti felbontási lépés kikerülésének célja motiválta. Az irodalomban előforduló szegmentális modellek többsége generatív [13], posterior-alapú megoldások viszonylag ritkák [1][12][21]. A kettő közötti lényeges különbség, hogy míg az előbbiekben a felbontás az (1) egyenlettel indul, addig az utóbbiban az első lépés ez:

$$P(W|X) = \sum_S P(W,S|X) = \sum_S P(W|S,X)P(S|X) \approx \max_S P(W|S,X)P(S|X). \tag{4}$$

Az előálló komponenseket ezután szegmentumokra bontjuk:

$$P(W|X) \approx \prod_i P(u_i|X_i) \quad \text{és} \quad P(S|X) \approx \prod_i P(S_i|X_i). \tag{5}$$

Osztályozási hibaarány	
Csak energiajellemzők	Energiajellemzők + hossz
47.72%	42.15%

Off-line normalizálási módszerek	CER%
Átlag és szórás normalizálása (teljes spektrum)	40.27%
Átlag és szórás normalizálása (sávonként)	37.75%
On-line normalizálási módszerek	
RASTA szűrés	43.86%
Átlag és szórás norm. (sávonként, $\tau = 250ms$)	41.12%
Átlag és szórás norm. (sávonként, $\tau = 1sec$)	40.36%
Nemlineáris AGC (sávonként, $\tau = 250ms$)	39.64%
Nemlineáris AGC (sávonként, $\tau = 1sec$)	38.49%

	τ méretű megfigyelési környezettel		
Normalizálás	$\tau = 150msec$	$\tau = 250msec$	$\tau = 1sec$
Off-line átlag és szórás norm.	33.18%	34.49%	36.12%
Nemlineáris AGC (*1sec*)	33.51%	34.85%	36.25%

onset/offset jellemzőkkel (off-line norm., 150ms környezet)	32.17%

1. táblázat. Beszédhang-osztályozási hibaarányok az MTBA korpuszon

$P(u_i|X_i)$ feladata a szegmentumokhoz beszédhang-címkék valószínűségét rendelni, ezért ezt a komponenst *beszédhang-osztályozónak* fogjuk nevezni. $P(S_i|X_i)$ neve *a szegmentum valószínűsége* és azt adja meg, hogy az adott jelrészlet milyen valószínűséggel tartozik beszédhanghoz. Az ezen értékekből előálló $P(S|X)$ komponenst a *szegmentálás valószínűségének* fogjuk nevezni.

Ha a beszédhangok osztályozására neuronhálót szeretnénk használni, akkor minden szegmentumot a hosszától függetlenül ugyanannyi jellemzővel kell leírnunk. A legegyszerűbb megoldás ilyen jellemzők előállítására, ha a hagyományos keretalapú jellemzők átlagával reprezentáljuk a szegmentumot. Ezt a technikát mi a jelek Bark-sávokban mért energiáira alkamaztuk, a szegmentumot az időtengely mentén három harmadra osztva és az átlagolást mindhárom darabon elvégezve. Az így kapott energiajellemzőket további olyan jellemzőkkel egészítettük ki, mint a szegmentum hossza vagy az átlagos sávenergiák a szegmentum valamely környezetében. Kísérleteztünk továbbá 'onset/offset' jellemzőkkel, amelyek a változás erősségét mérik a szegmentum elején/végén. A beszédjel energiájának normalizálása céljából egy automatikus erősítésszabályozó (AGC) algoritmussal, valamint az átlag és szórás normalizálásával is próbálkoztunk. A jellemzők paramétereit az MTBA telefonos adatbázison lőttük be. A 1. táblázatban megfigyelhetjük, hogy az egyes jellemzők fokozatos hozzáadássa hogyan javította a beszédhang-osztályozási eredményeket. Habár hasonló energiajellemzőkkel mások is kísérleteztek [6], az általunk bevezetett többi jellemző az általuk használtaktól eltér.

A kapott jellemzőkészlettel a számbemondásokat tartalmazó OASIS-Numbers adabázison is végeztünk beszédhang-osztályozási teszteket. Ez esetben HMM-es és SVM-es összehasonlító eredményeink is vannak (ld. 2. táblázat), amelyek egyértelműen a a szegmentális jellemzőkészletünket használó posterior-alapú modellek (ANN és SVM) javára billennek.

HMM	ANN	SVM
9.34%	7.78%	5.81%

2. táblázat. Beszédhang-osztályozási hibaarányok az OASIS-Numbers korpuszon

A szegmentumok osztályozása sokkal egyszerűbb feladat, mint jó becslést adni a $P(S|X)$ szegmentálási valószínűségre. Az irodalomban találhatunk olyan példákat, ahol egy HMM-felismerővel állítják elő az N legvalószínűbb szegmentálást, s így a $P(S|X)$ komponenst egyszerűen elhagyják [1]. Egy másik lehetőség keretről keretre megbecsülni annak valószínűségét, hogy az adott pozíció szegmentumhatár. Ezután ezek értékek szorzataként becsülhetjük $P(S|X)$-et [12]. Mi a számítási hatékonyság érdekében olyan megoldást igyekeztünk keresni, amely a beszédhang-osztályozót teszi képessé arra, hogy egyúttal a $P(S_i|X_i)$ értékeket is közelíteni tudja.

Hogy megértsük $P(S_i|X_i)$ szerepét, gondoljunk arra, hogy a dekódolóalgoritmus a felismerés során olyan szegmentumokkal is találkozik, amelyek nem felelnek meg beszédhangoknak – Glass és társai terminológiájával 'antifónok' [6]. A beszédhang-osztályozó ezeket sem felismerni, sem jelezni nem képes, hiszen ilyen szegmentumokon nem tanítottuk és speciálisan ezekhez rendelt kimenete sincs. Ha azt akarjuk, hogy az osztályozónk ezeket a szegmentumokat is kezelni tudja, akkor ki kell egészítenünk egy további osztállyal, továbbá a tanítás során ebbe az osztályba tartozó tanítópéldákat is kell generálnunk. Utóbbit úgy oldottuk meg, hogy a (kézileg szegmentált) tanítóadatbázisból vett fonetikai szegmentumok mindkét határát eltoltuk mindkét irányban 30-30 ezredmásodperccel, és ezen 'hibás' határok segítségével beszédhangonként 6-6 antifón tanítópéldát állítottunk elő.

Ha már rendelkezünk a $P(S_i|X_i)$ becslésekkel, akkor $P(S|X)$-et az alábbi módon közelíthetjük:

$$P(S|X) \approx \prod_i P(S_i|X_i). \tag{6}$$

Sajnos a gyakorlatban azt találtuk, hogy ez a módszer nem normalizálja megfelelően a jel különféle lehetséges szegmentálásait. Glass et al. szerint kiegyensúlyozottabb értékeket nyújt egy olyan közelítés, amely nem csak az adott S szegmentálás szegmentumait veszi figyelembe, hanem az összes szegmentumot, amellyel a felismerő találkozik a dekódolás során. Ez alapján a következő formulához juthatunk:

$$P(S|X) \approx \prod_i P(S_i|X_i) \prod_{s \in \overline{S}} (1 - P(s|X(s))), \tag{7}$$

ahol $\overline{S}$ az S-be eső S_i szegmentumok halmazának komplementerét jelöli. Ez a képlet mindig felhasználja az összes szegmentális becslést – ezek mindegyike vagy az első, vagy a második tényezőbe esik attól függően, hogy az éppen kiértékelt szegmentálásnak az adott szegmentum része-e vagy sem.

A (7) képlettel az a gond, hogy a második tényezője túl sok komponenst tartalmaz, így a gyakorlatban nem tudjuk hatékonyan kiértékelni. Ezért azt a közelítést alkalmaztuk, hogy $\overline{S}$-nek csak azokat az elemeit vettük figyelembe, amelyek S elemeihez bizonyos definíció szerint közel esnek. A módszer alapötletét a SUMMIT rendszerből vettük [6], de az 'antifón' elemeket eltérő módon használjuk, hiszen a miénkkel szemben a SUMMIT generatív felbontást alkalmaz.

Rendszerünk beszédhang-felismerési képességét az MTBA adatbázison mértük le, nyelvi modell támogatása nélkül. A felismert beszédhangok arányát (10%-on tartott beszúrási hiba mellett) az 3. táblázat összesíti. Izolált szavas felismerési kísérleteket is végeztünk, mégpedig az OASIS-Numbers adatbázison; az eredményeket a 4. táblázat mutatja. Mindkét táblázatból az olvashatjuk ki, hogy a

Mondatszintű felismerési eredmények			
antifónok nélkül	antifón-kombinálás (6) szerint	antifón-kombinálás (7) szerint	HMM
53.44%	58.74%	61.34%	61.60%

3. táblázat. Beszédhang-felismerési hatékonyság (% korrekt) az MTBA adatbázison

Szegmentum-alapú modell	HMM
0.95%	0.80%

4. táblázat. Szófelismerési hibaarányok az OASIS-Numbers adatbázison

	Jellemzőkészlet	
Antifón-modell	I.	II.
Antifón-modell nélkül	67.17%	68.58%
Antifón-modell (7) szerint	72.28%	77.28%
RNN	72.39%	75.21%

5. táblázat. Szószintű felismerési pontosságok RNN-t használva, a BeMe-Children adatbázison

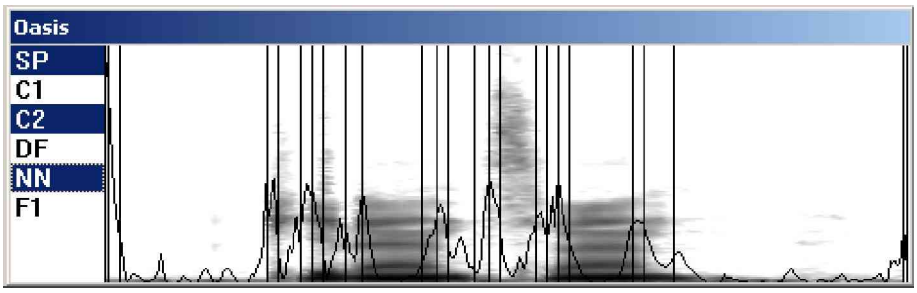
szegmentum-alapú rendszer épp csak utolérte a HMM-et, de felülmúlni nem tudta azt. Mivel a szegmentumok osztályozásában a szegmentális rendszer sokkal jobb volt, az eredményekből arra következtettünk, hogy a szegmentálási valószínűségek becslése még mindig nem elég pontos. A javítás érdekében két további módszerrel kísérleteztünk.

Neuronhálóink mesterségesen előállított 'antifón' tanítópéldákon való tanítása helyett olyan tanulóalgoritmusokkal is próbálkozhatunk, amelyek egyosztályos tanulásra képesek, azaz meg tudják tanulni valamely osztály határait a környezetéből származó ellenpéldák nélkül is. A replikátor neuronháló (RNN) egy olyan neurális modell, amely képes erre [7]. Az RNN alapötlete elég egyszerű: megtanulandó célértékekként is a bemenő adatokat használjuk. Ekkor a tanulás során a négyzetes hiba minimalizálása által a háló arra kényszerül, hogy minél pontosabban rekonstruálja az input-adatokat. Teszteléskor az ezektől távol eső minták általában nagy rekonstrukciós hibát eredményeznek, amely hiba ily módon felhasználható a megtanult osztályon kívül eső pontok detektálására.

Az RNN-re épülő antifón-modellt egy harmadik, gyerekek hangfelvételeit tartalmazó adatbázison teszteltük. A teszteket két különböző jellemzőkészlettel is lefuttattuk; az elért felismerési százalékokat az 5. táblázat összegzi. Az eredmények azt mutatják, hogy az RNN a korábbi módszerhez képest egy lehetséges alternatívát nyújt az antifónok kezelésére – tanítópéldák mesterséges generálása nélkül.

Az RNN mellett azzal is próbálkoztunk, hogy egy neuronhálót betanítunk (keretenként) a szegmentumhatárok felismerésére. A háló segítségével jelentősen csökkenteni lehet a felismerés során kiértékelt lehetséges szegmentumhatár-pozíciók számát. Ily módon a felismerés gyorsabbá és pontosabbá is tehető. A 1. ábra szemlélteti a háló kimenetét, illetve az abból előállított ritka szegmentálást. Amint az a 6. táblázatból kiolvasható, ezzel a módszerrel mind a futásidőt, mind a hibaarányt sikerült csökkenteni az Oasis-Numbers adatbázison (vö. 4. táblázat).

A szegmentális jellemzőkészletet TIMIT korpuszon mért beszédhang-osztályozási eredményekkel [10]-ben publikáltuk. A szegmentális felismerő szerkezetét beszédhang-osztályozási, -felismerési és szószintű eredményekkel [15]-ben és [16]-ban publikáltuk. Az RNN-es eredmények [17]-ben jelentek meg.



1. ábra. Ritka szegmentálás a határokat tanuló neuronháló segítségével

Szegmentálási módszer	szószintű hiba	kiértékelés időtartama
egységesen 5 keretenként	0.95%	549,250 msec
ANN-alapú ritka szegmentálás	0.78%	204,734 msec

6. táblázat. Szószintű hibaaarány az OASIS-Numbers korpuszon ANN-alapú előszegmentálással

A naív Bayes-feltételezés szerepe a beszédfelismerésben

A szegmentális modellt azért vezettük be, hogy kiküszöböljük a szegmentumok keretekre bontását s ezáltal a hibásnak vélt naív Bayes-feltevést. Természetesen azt reméltük ettől, hogy az ily módon előálló új rendszer jelentősen felülmúlja a régit. Ezzel szemben azt tapasztaltuk, hogy a szegmentum-alapú rendszer épp csak utolérte a HMM-et. Ez nem cáfolja, hogy a HMM nagyfokú modellezési torzítással rendelkezik, de úgy tűnik, hogy ez a torzítás valamiért nem rontja le a felismerési képességeit. Az alábbiakban megpróbáljuk megmagyarázni, hogy ez hogyan lehetséges. Az elvi okfejtés mellett néhány kísérletet is elvégzünk, amelyekben egy leegyszerűsített HMM-et és generatív modellt mérünk össze. Ezek a kísérletek is segíthetnek megérteni a kétfajta modell előnyeit és hátrányait egymáshoz képest.

Elég könnyű érveket hozni arra, hogy a keretek HMM által feltételezett feltételes függetlensége miért nem teljesül. A szomszédos adatkeretek eleve korreláltak az artikuláció folytonossága miatt, és az olyan jelfeldolgozási lépések, mint a trajektóriák simítása vagy a delta-jellemzők kinyerése a korrelációt tovább növelik. Habár ezen érveket számos szerzőnél megtalálhatjuk [13], úgy tűnik, egyetlen beszédfelismeréssel foglalkozó kutatóban sem fogalmazódott meg a kérdés – hogy akkor mégis miért működik olyan jól a HMM a gyakorlatban? A gépi tanulás irodalmában azonban több cikk is foglalkozik a naív Bayes-feltevés sokszor indokolatlanul jónak tűnő osztályozási teljesítményével. Ezek rámutatnak arra, hogy a naív Bayes-feltevés sok olyan esetben is optimális osztályozáshoz vezet, amikor az általa adott valószínűségi becslések nagyon pontatlanok [5]. Ilyen eset például az, amikor a jellemzők között teljes funkcionális függés áll fenn. De a naív Bayes-osztályozás olyankor is közel optimális, amikor a függőség csak közelítőleg teljes. Könnyen megérthetjük, hogy miért: ilyen esetekben az erősen függő jellemzők (esetünkben a szomszédos beszédkeretek) közel azonos valószínűségi becsléseket adnak, így amikor összeszorozzuk őket, az kb. olyan, mintha egyetlen becslést hatványoznánk. Az így kapott valószínűségi értékek jelentősen alul fogják becsülni a valódi valószínűségeket – de mivel a hatványozás a nagyságrendi viszonyokat megőrzi, ezért ennek ellenére helyes osztályozást fognak eredményezni.

A fenti érvelés megmagyarázza ugyan, hogy a HMM miért képes helyesen osztályozni a szegmentumokat, de azt is meg kell fejtenünk, hogy a szegmentálás problémáját hogyan oldja meg. Elvileg az állapotátmenet-értékek határozzák meg, hogy a HMM milyen állapotsorozaton megy keresztül,

Beszédhang-modell	Osztályozási Pontosság	Felism. pontosság	
		Unigram	Szótár
Keretalapú, szorzással	92.33%	82.05%	96.87%
Keretalapú, átlagolással	78.04%	—	86.28%
Keretalapú, szorzással, n-edik gyök	92.33%	—	41.78%
Szegmentális	94.58%	46.25%	87.00%
Szegmentális, n-edik hatvány	94.58%	57.99%	88.29%

7. táblázat. Osztályozási és felismerési pontosságok az Oasis-Numbers adatbázison

így azt gondolhatnánk, hogy lényegében ezen értékeken múlik, hogy a dekódolás során mely szegmentálásokhoz rendelünk nagy valószínűséget. Meglepő módon azonban sok kutató arról számolt be, hogy az állapotátmenet-valószínűségeknek jóformán semmi hatásuk sincs a felismerés hatásfokára [4]! Ebből az következik, hogy lényegében a szegmentálást is a naív Bayes kombinálási szabály oldja meg!

Hogy megértsük ennek mikéntjét, vizsgáljuk meg, hogyan viselkedik a HMM, amikor a felismerés során minden lehetséges szegmentálást kiértékel. Nyilvánvaló, hogy a naív Bayes-szabály a rövid szegmentumokat preferálja, hiszen kicsi pozitív számokat szoroz össze. A gyakorlatban a nyelvi modell segít elkerülni, hogy a HMM a jelet apró darabokra vágja – nyelvi modell híján általában be kell vezetnünk egy, a beszúrásokat büntető tényezőt. Ha a HMM-et rákényszerítjük a szomszédos keretek összevonására, akkor nyilván azokat a kereteket fogja összevonni, amelyeknek valamely osztály szerinti valószínűsége konzisztensen nagy. Ha a keretenkénti valószínűségértékek viszonylag megbízhatóak, akkor az ilyen keret-részsorozatok többnyire egybe fognak esni a valódi fonetikai szegmentumokkal, míg a szomszédos beszédhangokon átívelő részsozoratokban a keretenkénti értékek nagysága erősen szórni fog, s így az ilyen 'antifónokhoz' a naív Bayes-szabály kis valószínűségi értéket fog rendelni.

A kísérletekben egy 1-állapotú HMM-et hasonlítottunk össze egy egyszerű szegmentális modellel. Utóbbi csak az energiajellemzőket használta, és ezek eloszlását generatív módon, Gauss-komponensekkel közelítette. Ily módon a két rendszer közt különbség egyedül a $p(X_i|u_i)$ komponensben volt. A naív Bayes-feltevés hatását megértendő egyrészt megpróbáltuk a belőle adódó alulbecslést kompenzálni n-edik gyök vonásával (ahol n a keretek száma a szegmentumban), másrészt megkíséreltünk hasonló torzítást előidézni a szegmentális modell valószínűségi becsléseiben n-edik hatványra emeléssel. Puszta kíváncsiságból a keretenkénti értékek átlagolását is kipróbáltuk.

A kísérleti eredményeket a 7. táblázat összegzi (az Oasis-Numbers adatbázist használtuk). Az osztályozási eredmények egybeesnek korábbi megfigyeléseinkkel – a szegmentális reprezentáció jobb ugyan a keretenkénti becslések szorzásánál, de csak kicsivel. A hatványozásos és gyökvonásos vizsgálatokból arra következtethetünk, hogy a valószínűségek naív Bayes-feltételezésből eredő alulbecslése egyáltalán nem ártalmas a szegmentumhatárok megtalálására nézve; sőt, a hasonló jellegű torzítás mesterséges előidézése még javított is a szegmentum-alapú modell dekódolási képességén. Az eredmények rávilágítanak arra, hogy a jobb szegmentum-osztályozási képesség (ld. szegmentális modell) nem hozza automatikusan magával a felismerési eredmények javulását – a szegmentálási probléma hatékony kezelése legalább olyan fontos. Míg a HMM – a naív Bayes-feltevésnek köszönhetően – ezt a feladatot automatikusan megoldja, addig a szegmentum-alapú rendszerekben erre külön oda kell figyelni.

A naív Bayes-feltételezésre vonatkozó okfejtésünket és kísérleteinket [18]-ban publikáltuk.

Az átlagolós HMM/ANN hibrid

A szegmentum-alapú modell gyengeségeivel szembesülve úgy döntöttünk, hogy visszatérünk a keretalapú megközelítéshez. Mivel továbbra is neuronhálókat szándékoztunk használni, természetes, hogy figyelmünk a HMM/ANN hibrid modellek felé fordult. Megőriztük viszont speciális érdeklődésünket a keretek naív Bayes-feltételezéstől eltérő módon való kombinálása iránt. Megtartottuk továbbá a dekódolási folyamat tényezőinek a szegmentális rendszerrel végzett kísérletek során elsajátított értelmezését. Egész pontosan, megkíséreltük beazonosítani, hogy a keretalapú HMM/ANN hibridben mely komponensek felelnek meg a szegmentális modell szegmentum-osztályozójának, illetve szegmentálási valószínűségének. A két rendszer párhuzamba állítása egészen érdekes következetésekhez, valamint egy újszerű modellhez vezetett, melyet 'átlagolós hibridnek' fogunk nevezni.

A HMM/ANN hibrid modellben [3] a HMM-ben használt (3) alatti képletet lecseréljük a

$$p(X_i|u_i) \propto \prod_{j=s_{i-1}}^{s_i-1} \frac{P(u_i|x_j)}{P(u_i)}, \tag{8}$$

formulára, ahol $P(u_i|x_j)$-et neuronhálókkal közelítjük. A bal oldalon a Bayes szabályt alkalmazva némi átrendezések után azt kapjuk, hogy $P(u_i|X_i)$ – azaz a szegmentális modell *beszédhang-osztályozója* (vö. (5)) – a hibridben a következő módon áll elő:

$$P(u_i|X_i) \approx \frac{\prod_{j=s_{i-1}}^{s_i-1} P(u_i|x_j)}{P(u_i)^{l(i)-1}}, \tag{9}$$

ahol $l(i)$ a szegmentum hossza. Az osztályozók kombinációjával foglalkozó irodalomban (9) úgy ismert, mint a posterior valószínűségi becslések kombinálására vonatkozó *szorzási szabály*.

A hibrid modell bevezetői a $P(u_i)^{l(i)-1}$ tényezővel való osztás szükségességét kétségesnek ítélték [3]. A szorzási szabálynak ezen osztás nélkül előálló verzióját *egyszerűsített szorzási szabálynak* fogjuk nevezni. Egy további lehetőség a keretenkénti becslések átlagolása, amelyet *átlagolási szabályként* fogunk emlegetni. Ez utóbbi két szabály képlettel:

$$P(u_i|X_i) \approx \prod_{j=s_{i-1}}^{s_i-1} P(u_i|x_j) \quad \text{ill.} \quad P(u_i|X_i) \approx \frac{\sum_{j=s_{i-1}}^{s_i-1} P(u_i|x_j)}{l(i)}. \tag{10}$$

További fontos megjegyzés, hogy az átlagolás garantálja, hogy a $P(u_i = c_k|X_i)$ becsléseket az összes $\{c_1, ..., c_K\}$ osztálycímkére összeadva 1-et kapunk, a két szorzási szabály viszont nem. Ezért bevezettünk két további kombinációs szabályt, a *normalizált szorzási szabályt* és a *normalizált egyszerűsített szorzási szabályt*, amelyek a két szorzási szabályt kiegészítik egy normalizálási lépéssel.

Kétféle tesztet végeztük annak megállapítására, hogy melyik szabály eredményezi a legjobb beszédhang-modellt. Az egyik ezek körül természetesen a beszédhang-osztályozási hatásfok lemérése volt. A naív Bayes-feltevés elemzésekor viszont már megtanultuk, hogy a jó osztályozási képesség nem feltétlenül jelent pontos valószínűségi becslést. Ezért egy másik módszert is kipróbáltunk, ami remélhetőleg jobban méri a valószínűségi becslések pontosságát. Ez utóbbi módszer a valószínűségek marginalizálásán, egész pontosan az $\int_x p(x)P(u|x)dx = P(u)$ összefüggésen nyugszik. E képlet bal oldalára kapunk egy közelítést, ha átlagoljuk a kombinácós szabályok kimenetét az adatbázison, míg a jobb

Kombinációs szabály	PhER	MSE
szorzási szabály	43.19%	$8.12 \cdot 10^{110}$
egyszerűsített szorzási szabály	42.44%	$7.16 \cdot 10^{-4}$
átlagolási szabály	43.29%	$5.77 \cdot 10^{-5}$
normalizált szorzási szabály	43.19%	$1.34 \cdot 10^{-4}$
normalizált egyszerűsített szorzási szabály	42.44%	$5.09 \cdot 10^{-5}$

8. táblázat. Beszédhang-osztályozási hibaarányok (PhER), valamint a $\hat{P}(u)$ és $\frac{1}{|x|}\sum_x \hat{P}(u|x)dx$ becslések négyzetes különbsége (MSE) az egyes kombinációs szabályok esetén

oldal egyszerű leszámolással becsülhető. A két érték eltéréséből tudunk következtetni a kombinációs szabályok pontosságára.

A 8. táblázat mutatja az osztályozási, illetve a marginalizációs kísérlet eredményét az MTBA korpuszon. Az osztályozási eredmények gyakorlatilag megegyeznek, míg a marginalizációs vizsgálat arra utal, hogy a normalizált szabályok pontosabb valószínűségi becsléseket nyújtanak.

A következő lépésként megvizsgáltuk, hogyan viselkednek kombinációs szabályaink, ha a feladat beszédhang-sorozatok felismerése. Ezt az akadályt csak a szorzási és az egyszerűsített szorzási szabály vette, a három normalizált szabály katasztrófális eredményeket adott. Mivel az osztályozásban utóbbiak semmivel sem voltak rosszabbak, arra kellett gondolnunk, hogy valamiért képtelenek a helyes szegmentálás megtalálására. Másszóval, ideje volt azonosítanunk a hibrid modellben a szegmentum-alapú modell *szegmentálási valószínűségének* megfelelőjét. Mivel a normalizálás tönkretette a szorzási szabályok dekódolási képességét, jó okunk volt megvizsgálni a normalizáló tényezőt.

Abból kiindulva, hogy a neuronháló egyes osztályokhoz tartozó kimeneteinek összege mindig 1, könnyen levezethetjük az egyszerűsített szorzási szabály által adott szegmentális posterior valószínűségek összegét. Erre az alábbit kapjuk:

$$\sum_{k=1}^{K} P(u_i = c_k | x_{s_{i-1}}^{s_i-1}) = 1 - \sum_{\substack{1 \le k_{s_{i-1}}, \cdots, k_{s_i-1} \le K \\ \exists p, r : k_p \neq k_r}} \left(\prod_{j=s_{i-1}}^{s_i-1} P(u_i = c_{k_j} | x_j) \right). \tag{11}$$

Szavakban kifejezve, a jobb oldalon szereplő szumma csupa olyan szorzatot tartalmaz, amelyekben a c_k osztálycímkék keverednek. Ez az összeg annál nagyobb, minél nagyobb az ellentmondás az egyes keretek alapján kapott valószínűségi becslések között. Következésképp, (11) értelmezhető a $P(S_i|X_i)$ szegmentum-valószínűség egy lehetséges közelítéseként. Vagyis elmondhatjuk, hogy megtaláltuk az egyszerűsített szorzási szabályra épülő hibrid modell szegmentálási valószínűséget megadó tényezőjét: ez inherens módon benne van magában a kombinációs szabályban annak köszönhetően, hogy az egyes osztálycímkékhez rendelt valószínűségek összege nem 1! Úgy is fogalmazhatunk, hogy a hibrid a *normalizált* egyszerűsített szorzási szabállyal közelíti $P(u_i|X_i)$-et és (11)-gyel $P(S_i|X_i)$-t.

A fenti okoskodás igazolása céljából megvizsgáltuk, vajon (11)-et használva $P(S_i|X_i)$ becslésére képessé válik-e az átlagolós szabály beszédhang-, ill. szófelismerésre. Mint a 9. táblázat mutatja, az így előállt 'átlagolós hibrid' mindkét feladatban a legjobbnak bizonyult (ismétcsak az MTBA adatbázison).

Az átlagolós hibrid alapötletét a [19] publikációban vázoltuk fel.

Modell	HMM	Szorzás	Egysz. szorzás	Átlagolás
CORRECT	56.29%	56.82%	57.46%	57.01%
ACCURACY	46.38%	46.44%	47.71%	47.92%

Modell	WER
HMM	4.80%
hagyományos hibrid	4.57%
átlagolós hibrid	3.20%

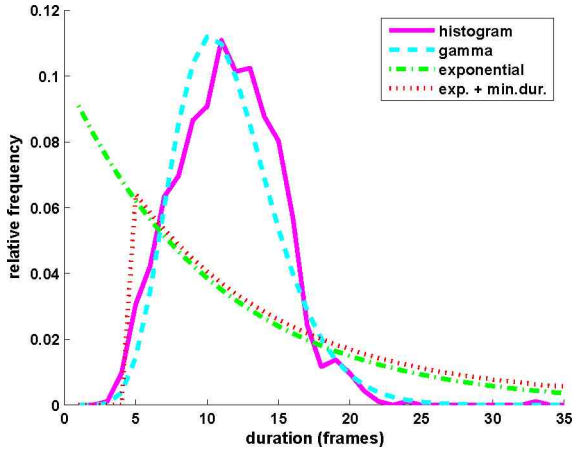
9. táblázat. A hibrid modellek hang- ill. szófelismerési teljesítménye az MTBA korpuszon

Explicit hosszmodellek alkalmazása HMM/ANN hibridekben

A magyarban a hosszúság fonológiai megkülönböztető jegy, ezért fontos lenne precízen modellezni. A HMM azonban elég pontatlanul írja le a hangok hosszát. Habár az állapotátmenet-valószínűségek szorzata értelmezhető úgy, mint egy geometriai (vagy exponenciális) eloszlásra épülő hosszmodell, az exponenciális eloszlás nagyon rossz közelítését adja a hanghosszoknak. Másrészt az is közismert, hogy az állapotátmenet-értékekenek jóformán semmi hatásuk sincs a felismerés pontosságára [4].

A fenti állítások letesztelése céljából többféle hosszmodellt is kipróbáltunk a HMM/ANN hibridekbe integrálva. Mind a hagyományos (szorzáson alapuló), mind az átlagolós hibridet megvizsgáltuk e szempontból. A hosszmodell által adott értékeket egy jól belőtt α_D hatványra emeltük, ezután szorzás útján kombináltuk a $P(u_i|X_i) \cdot P(S_i|X_i)$ komponenssel. A hosszmodell mellett egy I hangbeszúrást büntető konstanssal is kibővítettük a modellt. Az α_D és I konstansok optimális értékének megtalálására a SNOBFIT nevű globális optimalizáló algoritmust használtuk [8].

Hosszmodellként a következő függvényeket próbáltuk ki. Legelőször is leteszteltük a felismerőket *hosszmodell nélkül* – az így kapott felismerési eredmények összehasonlítási alapként szolgáltak. A következő kísérletben *exponenciális hosszmodelleket* használtunk, beszédhangonként egyet-egyet. Ez a beállítás megfelel a HMM-ek szokványos hosszmodellezési módjának. Megpróbáltunk továbbá egyetlen *közös exponenciális hosszmodellt* használni az összes beszédhanghoz, valamint kiterjeszteni az *exponenciális hosszmodellt egy (4 keretnyi) minimumhossz-megkötéssel*. Végezetül, kisérleteket végeztünk a *gamma-hosszmodellel*, amely gamma-eloszlást illeszt a hanghosszok hisztogramjára [14]. A 2. ábra szemlélteti, mennyire jól vagy rosszul közelítik az egyes modellek egy konkrét hang hosszeloszlását.



2. ábra. Beszédhang hossz-hisztogramjának közelítése különféle hosszmodellekkel

Hosszmodell	Hagyományos hibrid			Átlagolós hibrid		
	α_D	I	WER	α_D	I	WER
Hosszmodell nélkül	–	1.511	**5.80%**	–	0.254	**4.87%**
Közös exponenciális	0.266	2.036	**5.80%**	0.934	0.806	**4.87%**
Exponenciális	0.340	3.804	**5.80%**	0.560	1.098	**4.87%**
Gamma modell	0.382	3.311	**5.10%**	0.306	0.415	**3.94%**

10. táblázat. Szófelismerési hibaarány (WER) α_D és I optimalizálása után

Elég nyilvánvaló, hogy a gamma modell adja messze a legjobb közelítést. Az előzetes kísérletekben a hangok minimális hosszára adott korlátozó feltétel mindig javított a felismerési eredményeken. Mivel emellett jelentősen fel is gyorsítja a felismerést, ezért az 'éles' kísérletekben mindig alkalmaztuk.

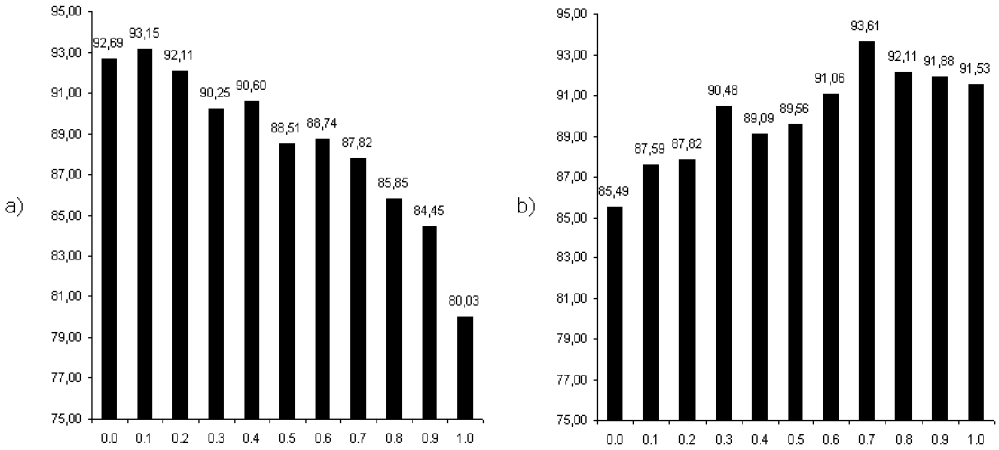
Az MTBA korpuszon kapott szófelismerési eredményeket a 10. táblázatban láthatjuk. A számok azt mutatják, hogy – jól megválasztott α_D és I értékek esetén – hosszmodell nélkül is pontosan olyan eredmények hozhatók ki, mint a kétféle exponenciális hosszmodellel. Csak a gamma hosszmodell tudott a többiekénél jobb eredményt elérni – utóbbi viszont mind a hagyományos, mind az átlagolós hibrid teljesítményén javított.

A hosszmodellekkel végzett kísérleteket [19]-ben publikáltuk.

HMM/ANN hibridek tanítása újramintavételezéssel

Jelen disszertációban mindenütt neuronhálókat alkalmaztunk az oszályok posterior valószínűségének becslésére. A neuronhálók ilyetén felhasználását egy elméleti eredmény teszi lehetővé, amely bizonyítja, hogy megfelelő feltételek esetén a neuronháló kimenetei az egyes osztályok valószínűségeit adják meg [2]. A gyakorlatban azonban a bizonyításban szereplő feltételek – végtelen sok tanító adat, többek között – nem teljesíthetők. Így a háló a valóságban csak közelítéseket nyújt – melyek pontatlansága sok esetben karakterisztikus. Egy ilyen esetre példa, amikor a különböző osztályokból rendelkezésünkre álló tanítópéldák száma nagyon eltérő. Gyakran tapasztaljuk ilyenkor, hogy a kevés példával reprezentált osztályokat a háló hajlamos rosszabbul megtanulni. Gyakori megoldás erre ezen osztályok példáit többször beadni a hálóba. Ezeket a technikákat összefoglaló néven 'újramintavételezési eljárásoknak' nevezi az irodalom [22]. Ezek közül mi konkrétan egyetlen technikával, az ún. 'valószínűségi mintavételezéssel' [11] foglalkoztuk, pontosabban megvizsgáltuk annak alkalmazhatóságát a HMM/ANN hibridek neuronháló-komponensének betanításában. Ez a módszer két lépésben választja ki a következő tanítópéldát: először egy c_k osztályt választ valamely $P'(c_k)$ eloszlást követve, majd egy konrét tanítópéldát ebből az osztályból a $p(x|c_k)$ eloszlás alapján.

Először is megnéztük, hogyan befolyásolja a $P'(c_k)$ eloszlás megválasztása a posterior valószínűségek közelítésére vonatkozó tételt. Könnyen belátható, hogy általános esetben a tanítópéldák eloszlásának manipulálása érvényteleníti a tételt. Van azonban két speciális eset, amelyekről tudunk mondani valamit. Az egyik természetesen az, amikor $P'(c_k)$ megegyezik $P(c_k)$-val, az osztályok természetes eloszlásával. Ilyenkor a valószínűségi mintavételezési séma ugyanazt a példaeloszlást eredményezi, mint a hagyományos 'teljes mintavételezés', így a háló kimenetei a $P(c_k|x)$ értékeket közelítik. A másik speciális eset, amikor $P'(c_k) = \frac{1}{K}$, vagyis egyforma valószínűséggel választjuk az egyes osztályokat.



3. ábra. Szófelismerési pontosságok (%) λ függvényében, a $P(c_k)$-val való osztással ill. anélkül

Ezt az esetet 'osztályonkénti mintavételezésnek' neveztük el és megmutattuk, hogy ilyenkor a neuronháló kimenetei $p(x|c_k)$-szel arányos értékeket adnak. Emlékezzünk vissza rá, hogy a hagyományos hibridnek pontosan ilyen értékekre van szüksége, s ezeket a $P(c_k|x)/P(c_k)$ tört formájában állítja elő (ld. (8)). A valószínűségi mintavételezési séma egy alternatív megoldást kínál: a hálót taníthatjuk osztályonkénti mintavételezéssel, s ez esetben a $P(c_k)$-val való osztásra sem lesz szükség.

A bizonyítás osztályonkénti mintavételezéshez igazított változata pontosan ugyanazokat a gyakorlatban teljesíthetetlen feltevéseket tartalmazza, mint az eredeti. A kísérletekben ezért egy folytonos átmenetet állítottunk elő a két speciális eset között $P'(c_k)$-et $(1-\lambda)P(c_k)+\lambda\frac{1}{K}$ alakban felírva, ahol λ-t 0 és 1 között változtattuk. A felismerőt minden λ értéknél lefuttattuk a $P(c_k)$-val való osztással és anélkül is. A 3. ábra összegzi az MTBA korpuszon elért szófelismerési pontosságokat. Az eredmények azt mutatják, hogy nem azokkal a λ értékekkel – $\lambda = 0.0$ és $\lambda = 1.0$ – kaptunk optimális teljesítményt, amelyekkel a tételek szerint kellett volna, hanem a két végpont közé eső beállításokkal. Ez a példa jól érzékelteti a beszédfelismerés elmélete és gyakorlata közt gyakran tapasztalt különbséget.

Az újramintavételezéses tanítási sémával végzett kísérleteket [20]-ban publikáltuk.

Összegzés

A Szerző véleménye szerint a disszertáció legfontosabb eredményét nem a javasolt modellek maguk képezik, hanem az általuk nyert betekintés a felismerők működésébe. A szegmentum-alapú modell rávilágított, hogy a fonetikai szegmentumok osztályozása a HMM-nél jóval egyszerűbb és ésszerűbb módszerrel is megoldható. Habár a dekódolási feladatokban a szegmentális rendszer nem viselkedett túl meggyőzően, az általa nyújtott látásmód segített jobban megérteni, hogyan működnek a keretalapú rendszerek. Mind a hagyományos HMM-et, mind a HMM/ANN hibridet megvizsgáltuk a szegmentum-alapú szemszögből, s mindkét esetben azt találtuk, hogy a keret-valószínűségek szorzással történő kombinálása oldja meg a szegmentálást is, nem csak a szegmentumok osztályba sorolását. Figyelembe véve, hogy a hangok osztályozásában a szegmentális modell és az átlagolás is hasonlóan jól (vagy jobban) működött, mint a szorzás, összességében azt kell mondanunk, hogy a naív Bayes-szabály fő hozzájárulása a felismeréshez nem az *osztályozás*, hanem a *szegmentálás* megoldásában rejlik. Ez elég meglepő eredmény, s jól mutatja, hogy néha érdemes egy régi problémát egy új aspektusból megnézni.

Az eredmények tézisszerű összefoglalása

Az alábbiakban hét tézispontba rendezve összegezzük a Szerző kutatási eredményeit.

I.) A Szerző összeállított egy fonetikai szegmentumok parametrizálására alkalmas szegmentális jellemzőkészletet. A jellemzőkészlet reprezentációs erejét többféle adatbázison, többféle tanulóalgoritmussal kombinálva tesztelte. Ezek alapján azt találta, hogy a jellemzőkészlet segítségével általában jobb beszédhang-felismerési eredmények érhetők el, mint a hagyományos HMM beszédhangmodellekkel.

II.) A Szerző különféle, az ún. 'antifón' koncepción alapuló stratégiákat fejlesztett ki a posterior-alapú szegmentális beszédmodell szegmentálási valószínűségeket leíró komponensének közelítésére. A javasolt modellezési módszerek beszédfelismerési képességeit több magyar beszédkorpuszon is kiértékelte.

III.) A Szerző megvizsgálta a replikátor neuronhálók alkalmazhatóságát a szegmentum-alapú modell szegmentálási valószínűségének közelítésére.

IV.) A Szerző tanulmányozta a naív Bayes feltevés által okozott modellezési torzítás kihatását a HMM beszédhangmodellek beszédfelismerési képességeire. A vizsgálatok alapján amellett érvelt, hogy a torzítás tulajdonságainál fogva csak kis mértékben ártalmas a HMM beszédhang-azonosítási képességére nézve, miközben számottevően segíti a modellt a beszédjelek fonetikai szegmentumokra tagolásában. Az eredmények megmagyarázzák, hogyan lehet képes a HMM nagyon jó felismerési eredmények elérésére annak ellenére, hogy a valószínűségi becslései meglehetősen pontatlanok.

V.) A Szerző megvizsgálta a HMM/ANN hibrid modell működését, összevetve azt a szegmentum-alapú modellel. A levont következtetések alapján bevezetett egy újszerű hibrid struktúrát, amely a keretekhez rendelt posterior valószínűségi értékeket szorzás helyett átlagolással kombinálja. Felismerési kísérletekkel igazolta, hogy az 'átlagolós hibrid' hasonló vagy jobb felismerési eredményekre képes, mint a hagyományos hibrid.

VI.) A Szerző megvizsgálta a gamma-eloszláson alapuló explicit hosszmodellek alkalmazhatóságát hibrid HMM/ANN beszédfelismerőkben. Azt találta, hogy a gamma-eloszlást használó hosszmodell mind a hagyományos, mind az átlagolós hibrid felismerési teljesítményén javít.

VII.) A Szerző egy újramintavételezési sémán alapuló tanítási technológiát javasolt a hibrid modellekben alkalmazott neuronhálók betanításához. A kísérletekben a javasolt módszer a felismerési eredmények kis mértékű javulását eredményezte.

A kutatásokból származó publikációkat, valamint azok tartalmának az egyes tézispontokhoz való viszonyát a 11. táblázat tekinti át.

	[10]	[15]	[16]	[17]	[18]	[19]	[20]
I.	•	•	•				
II.		•	•				
III.				•			
IV.					•		
V.						•	
VI.						•	
VII.							•

11. táblázat. A tézispontok és a Szerző publikációinak viszonya

Hivatkozások

[1] Austin, S., Zavaliagkos, G., Makhoul, J., Schwartz, R., Speech Recognition using Segmental Neural Nets, Proceedings of ICASSP'92, Vol. 1, pp. 625-628, 1992.

[2] Bishop C. M., Neural Networks for Pattern Recognition, Clarendon Press, 1995.

[3] Bourlard, H. A., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach, Kluwer Academic, 1994.

[4] Bourlard, H., Hermansky, H. and Morgan, N., Towards Increasing Speech Recognition Error Rates, Speech Communication, pp. 205-231, May 1996.

[5] Domingos P. and Pazzani M., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Vol. 29, pp. 103-130, 1997.

[6] Glass, J. R.: A Probabilistic framework for segment-based speech recognition, Computer Speech and Language, Vol. 17, pp. 137-152, 2003.

[7] Hawkins, S., He, H. X., Williams, G. J., Baxter,R. A., Outlier Detection Using Replicator Neural Networks, Proc. DaWak'02, 2002.

[8] Huyer, W., Neumaier, A., SNOBFIT - Stable Noisy Optimization by Branch and Fit, Submitted for Publication

[9] Kleinschmidt, M., Localized Spectro-Temporal Features for Automatic Speech Recognition, Proceedings of EuroSpeech 2003, pp. 2573-2576., 2003.

[10] Kocsor, A. and Tóth, L., Kernel-Based Feature Extraction with a Speech Technology Application, IEEE Transactions on Signal Processing, Vol. 52, No. 8, pp. 2250-2263, 2004.

[11] Lawrence, S., Burns, I, Back, A, Tsoi, A. C., Giles, C. L., Neural Network Classification and Prior Class Probabilities, In: Orr et al. (eds.), Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science State-of-the-Art Surveys, Springer, pp. 299-314, 1998.

[12] Leung, H. C., Hetherington, I. L., Zue, V. W., Speech Recognition using Stochastic Segment Neural Networks, Proceedings of ICASSP'92, Vol 1., pp. 613-616, 1992.

[13] Ostendorf, M., Digalakis, V., Kimball, O. A., From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, IEEE Trans. ASSP, Vol. 4. pp. 360-378, 1996.

[14] Pylkönnen, J., Kurimo, M., Duration Modeling Techniques for Continuous Speech Recognition, Proceedings of ICSLP' 2004, pp. 385-388, 2004.

[15] Tóth, L., Kocsor, A., Kovács, K., 2000. A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, In: Sojka, P. et al. (eds.), Proceedings of Int. Conf. on Text, Speech and Dialogue TSD'2000, Lecture Notes in Artificial Intelligence Vol. 1902, pp. 307-313, Springer, 2000.

[16] Tóth, L., Kocsor, A., Gosztolya, G., Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, pp. 643-657, 2004.

[17] Tóth, L., Gosztolya, G., Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition, In: Yin, F. et al. (eds.), Proceedings of Int. Symp. on Neural Networks ISNN'2004, Lecture Notes in Computer Science Vol. 3173, pp. 996-1001, Springer, 2004.

[18] Tóth, L., Kocsor, A., Csirik, J., On Naive Bayes in Speech Recognition, International Journal of Applied Mathematics and Computer Science, Vol. 15, No. 2, pp. 287-294, 2005.

[19] Tóth, L., Kocsor, A., Explicit Duration Modelling in HMM/ANN Hybrids, In: Matousek et al. (eds.), Proceedings of Int Conf. on Text, Speech and Dialogue TSD 2005, Lecture Notes in Artificial Intelligence Vol. 3658, pp. 310-317, Springer, 2005.

[20] Tóth, L., Kocsor, A., Training HMM/ANN Hybrid Speech Recognizers by Probabilistic Sampling, In: Duch et al. (eds.), Proceedings of Int. Conf. on Artificial Neural Networks ICANN'2005, Lecture Notes on Computer Science Vol. 3696, pp. 597-603, Springer, 2005.

[21] Verhasselt, J., Illina, I., Martens, J.-P., Gong, Y., Haton, J.-P., Assessing the importance of the segmentation probability in segment-based speech recognition, Speech Communication, Vol. 24, No. 1., pp. 51-72, 1998.

[22] Weiss, G. M., Provost, F., The Effect of Class Distribution on Classifier Learning: An Empirical Study, Tech. Report ML-TR-44, Dep. Comp. Sci., Rutgers Univ, 2002.

Társszerzői nyilatkozatok

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő I. tézispontra, illetve az azt lefedő

> Kocsor, A., Tóth, L., Kernel-Based Feature Extraction with a Speech Technology Application, IEEE Transactions on Signal Processing, Vol. 52., No. 8, pp. 2250-2263, 2004.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

> A szegmentum-alapú jellemzőkészlet megtervezése, a kinyeréséhez szükséges algoritmusok implementálása, a jellemzőkészleten alapuló beszédhang-felismerési tesztek megtervezése és végrehajtása.

A következő – Tóth László disszertációjában fel nem használt – eredményekben az én hozzájárulásom volt a meghatározó:

> A kernel-alapú tulajdonságkinyerő algoritmusok bevezetése és egységes tárgyalásmódjának kialakítása, a kernel-alapú algoritmusokra vonatkozó tesztek megtervezése és végrehajtása.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő I-II. tézispontokra, illetve az azokat lefedő

> Tóth, L., Kocsor, A., Kovács, K., A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition. In: Sojka, P. et al. (eds.), Proceedings of Int. Conf. on Text, Speech and Dialogue TSD'2000, Lecture Notes on Artificial Intelligence Vol. 1902, pp. 307-313, Springer, 2000.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

> A szegmentum-alapú jellemzőkészlet megtervezése, a kinyeréséhez szükséges algoritmusok implementálása, a jellemzőkészleten alapuló beszédhang-felismerési tesztek megtervezése és végrehajtása.
>
> A posterior valószínűségekre épülő, szegmentum-alapú beszédmodell szegmentálási valószínűségének becslését célzó, „anti-fonémákon" alapuló közelítési sémák kidolgozása és tesztelése.

A következő – Tóth László disszertációjában fel nem használt – eredményekben az én hozzájárulásom volt a meghatározó:

> A kernel-alapú tulajdonságkinyerő algoritmusokra vonatkozó beszédhang-felismerési tesztek megtervezése és végrehajtása.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András

Kovács Kornél

Kovács Kornél

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő I-II. tézispontokra, illetve az azokat lefedő

Tóth, L., Kocsor, A., Gosztolya, G., Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, pp. 643-657, 2004.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

A szegmentum-alapú jellemzőkészlet megtervezése, a kinyeréséhez szükséges algoritmusok implementálása, a jellemzőkészleten alapuló beszédhang-felismerési tesztek megtervezése és végrehajtása.

A posterior valószínűségekre épülő, szegmentum-alapú beszédmodell beszédhang-dekódolási képességeire vonatkozó tesztek megtervezése és végrehajtása.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András

Gosztolya Gábor

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő III. tézispontra, illetve az azt lefedő

> Tóth, L., Gosztolya, G., Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition, In: Yin, F. et al. (eds.), Proceedings of Int. Symp. on Neural Networks ISNN'2004, Lecture Notes on Computer Science Vol. 3173, pp. 996-1001, Springer, 2004.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

> A replikátor neuronháló (RNN) megtervezése és implementálása, a neuronhálót felhasználó beszédfelismerési tesztek megtervezése.

Mivel jómagam csak a tesztek lefuttatásában segédkeztem, ezért az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Gosztolya Gábor

Kocsor András (mint Gosztolya Gábor témavezetője)

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő IV. tézispontra, illetve az azt lefedő

Tóth, L., Kocsor, A., Csirik, J., On Naive Bayes in Speech Recognition, International Journal of Applied Mathematics and Computer Science, Vol. 15, No. 2, pp. 287-294, 2005.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

A „naív Bayes" feltevés elemzéséhez kapcsolódó kísérletek megtervezése, a kísérletek végrehajtása és kiértékelése, a „naív Bayes" feltevésre vonatkozó következtetések levonása.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András

Csirik János

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő V-VI. tézispontokra, illetve az azokat lefedő

Tóth, L., Kocsor, A., Explicit Duration Modelling in HMM/ANN Hybrids, Matousek et al. (eds.): Proceedings of Int Conf. on Text, Speech and Dialogue TSD 2005, Lecture Notes on Artificial Intelligence Vol. 3658, pp. 310-317, Springer, 2005.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

Az átlagolós hibrid HMM/ANN modell bevezetése, az elemzéséhez kapcsolódó kísérletek megtervezése, a kísérletek végrehajtása és kiértékelése.

A Gamma-eloszláson alapuló hosszmodell implementálása és tesztelése.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Tóth László Ph.D-fokozatra pályázó „Posterior-Based Speech Models and their Application to Hungarian Speech Recognition" című disszertációját. A disszertációban szereplő VII. tézispontra, illetve az azt lefedő

> Tóth, L., Kocsor, A., Training HMM/ANN Hybrid Speech Recognizers by Probabilistic Sampling, In: Duch et al. (eds.), Proceedings of Int. Conf. on Artificial Neural Networks ICANN'2005}, Lecture Notes on Computer Science Vol. 3696, pp. 597-603, Springer, 2005.

c. publikációra vonatkozóan kijelentem, hogy a következő eredményekben a pályázó hozzájárulása volt a meghatározó:

> Az újramintavételezésen alapuló tanítási séma modellezési tulajdonságainak matematikai elemzése, a hatékonyságának empirikus kiértékelését célzó tesztek megtervezése, a tesztek végrehajtása és kiértékelése.

Az említett publikációt fokozatszerzésre nem használtam fel és a későbbiekben sem fogom felhasználni.

Szeged, 2005. 12. 27.

Kocsor András