Beszédfelismerési kísérletek hangoskönyvekkel

Tóth László

MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport
tothl@inf.u-szeged.hu

Kivonat: Valós körülmények között a gépi beszédfelismerést számos tényező nehezíti, például a háttérzaj, a beszélő hangjának egyéni sajátosságai, a spontán artikuláció vagy a beszéd érzelmi töltete. A gyakorlatban is alkalmazható felismerőrendszerek készítéséhez természetesen ezeket a problémákat mind tudni kell kezelni, egyelőre azonban a jóval egyszerűbb feladatokat sem tudjuk tökéletesen megoldani. Jelen cikkben annak megvizsgálása a célkitűzésünk, hogy vajon mire képes a jelenlegi technológia „ideális" körülmények között. Az optimális viszonyok szimulálásához egy hangoskönyv hanganyagával dolgozunk, mivel ennek rögzítése során az említett hátráltató tényezők többsége nem, vagy csak minimális mértékben jelentkezik. A kiértékelés segítése érdekében a kapott eredményeket egy korábbi, telefonos adatbázison végzett hasonló kísérletsorozat eredményeivel állítjuk párhuzamba. Méréseink szerint a hangoskönyvön kapott fonetikai kimenet pontossága már minimális nyelvi támogatással is 86% fölött van, és emberi szemmel is majdnem tökéletesen olvasható.

1 Bevezetés

A piac használható alkalmazások iránti igénye a beszédtechnológiai kutatást erőteljesen kényszeríti az egyre nehezebb, komplexebb problémák irányába – példa erre a zajos beszéd felismerése iránti igény vagy az utóbbi időben a természetes, spontán beszéd vizsgálatának fókuszba kerülése. A piaci elvárások persze jogosak olyan értelemben, hogy a gyakorlati használhatósághoz valóban túl kell lépni a csak steril laboratóriumi körülmények között működő rendszereken. Ez nincs alapvető ellentmondásban a kutatók vágyaival, hiszen végcélnak ők is a teljesen kötetlen beszéd felismerését tekintik. A baj inkább az, hogy egyelőre még az egyszerűbb, „redukált" felismerési feladatok sincsenek teljesen megoldva, így az ipar egyfajta „előremenekülésre" kényszeríti a kutatókat – miközben a problémamegoldás íratlan szabályai sokkal inkább az egyszerűbb feladatokra való visszalépést írnák elő. Épp ezért azt gondoljuk, hogy nem szabad abbahagyni az egyszerűsített felismerési szituációk vizsgálatát sem, mivel a nehézségeket okozó tényezőket szétválasztva könnyebb azokat elemezni és megérteni. Az egyszerűbb, könnyebb feladatok vizsgálatát továbbá azért sem érdemes feladni, más számos olyan értelmes alkalmazás létezik, ahol ezeknek is létjogosultságuk lehet (például egy rádióhíreket figyelő vagy TV-híradót feliratozó rendszer esetén mind a stúdióminőségű felvétel, mind a fegyelmezett artikuláció feltételezhető).

Jelen cikkünkben hangoskönyveken végzünk beszédfelismerési kísérleteket. A tesztekkel annak megvizsgálása a célunk, hogy a jelenlegi beszédfelismerők (főleg az akusztikus komponens) mire lennének képesek ideális körülmények közt, vagyis ha a zavaró tényezők nagy részét ki tudnánk zárni. A hangoskönyvek tartalma „ideális" beszédnek tekinthető olyan értelemben, hogy a beszédfelismerést valós helyzetben megnehezítő tényezők közül a legtöbb nem jelentkezik a hanganyagukban. A 2. fejezetben áttekintjük ezeket a tényezőket, és megpróbáljuk érzékeltetni a beszédfelismerőkre tett hatásukat. Az érzékeltetést fogja szolgálni az is, hogy felismerési eredményeinket párhuzamba állítjuk a 2008-as Interspeech konferencián publikált értékekkel, melyeket ugyanazon felismerési technikával értünk el, de az MTBA telefonbeszéd-adatbázison. Az eredmények 5. fejezetbeli közlése előtt azonban természetesen részletesen ismertetjük a hanganyag feldolgozásának lépéseit a 3., majd az alkalmazott ún. „tandem" felismerési technológiát a 4. fejezetben. Cikkünk az eredmények elemzésével és a következmények levonásával zárul a 6-7. fejezetekben.

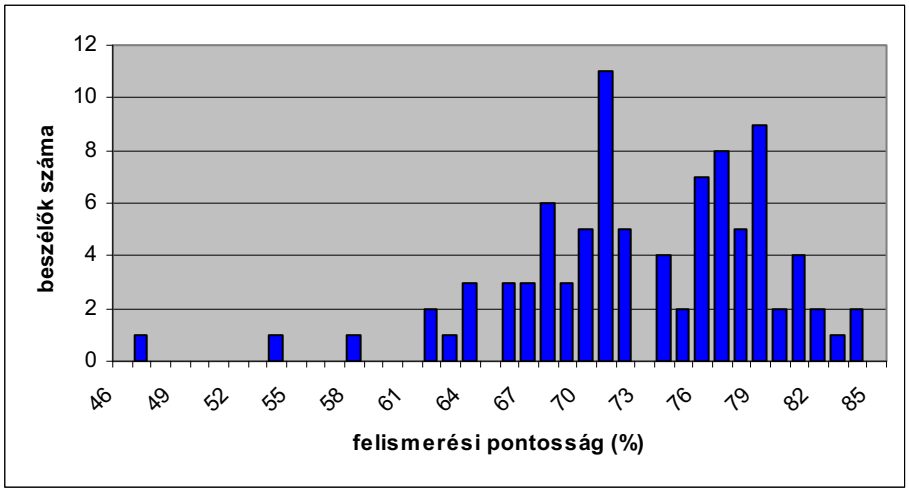
2 A beszédfelismerést megnehezítő tényezők

Az alábbiakban áttekintjük a beszédfelismerést valós szituációkban megnehezítő fő tényezőket, és az irodalomból vett adatokkal kíséreljük meg érzékeltetni jelentőségüket. Megvizsgáljuk továbbá, hogy a hangoskönyvre és az összehasonlítási alapként szolgáló MTBA adatbázisra az adott tényező milyen mértékben jellemző.

Gyakorlati körülmények lényegében nincs olyan helyzet, amelyben a háttérzaj beszűrődése teljesen megakadályozható lenne. Tapasztalataink szerint még a híradók stúdióban rögzített felvételein is akad háttérzaj, behallatszik például, ahogy a bemondó a papírjait rendezgeti. És sok olyan alkalmazás van, amely kifejezetten erős háttérzaj mellett kísérel meg beszédfelismerést használni (pl. egy vadászgép pilótafülkéjében). Hagyományosan zajnak (ún. konvolutív zaj) tekintjük továbbá az átviteli közeg (pl. telefonvonal) okozta torzítást is, amely bizonyos frekvenciakomponensek gyengülését-erősödését okozza. A háttérzaj a beszédfelismerők felismerési pontosságát drasztikusan csökkenti, főleg ha az emberi beszédpercepcióval párhuzamba állítva vizsgáljuk [6]. A konvolutív zaj hatása még kiábrándítóbb, ugyanis ezt mi emberek szinte nem is érzékeljük (legfeljebb a hangszín változásaként), miközben a felismerők hatékonyságát meglepő fokban le tudja rontani. Ezt jól példázza, hogy ha diktálószoftvert vásárolunk, általában mikrofont is kapunk hozzá, mivel már pusztán másik mikrofon használata is érzékelhető teljesítménycsökkenéssel járhatna. Esetünkben az összehasonlítási alapként szolgáló MTBA adatbázis különféle vonalakon rögzített telefonos felvételeket tartalmaz, a telefon szokásos torzításával és frekvenciavágásával. Háttérzaj is beszűrődik a felvételekbe, bár tapasztalataink szerint viszonylag ritkán (az adatközlők érzékelhetően nyugodt körülményeket választottak a híváshoz). Ezzel szemben a feldolgozott hangoskönyv stúdióban készült, így háttérzajt gyakorlatilag nem tartalmaz, és feltehetően professzionális mikrofonnal vették fel (bár azt nem tudhatjuk, hogy végig ugyanazzal-e).

A gépi beszédfelismerők közismerten érzékenyek a beszélő személyére, azaz az egyes beszélők hangja közt adódó eltérésekre. Az MTBA adatbázis 500 adatközlő felvételeit tartalmazza, és mindenkitől csak 12-12 mondatot, így a beszélő személye

igen gyakran változik. Az 1. ábrán bemutatott hisztogram a különböző adatközlőkre kapott felismerési pontosságok szórását érzékelteti egy konkrét, az MTBA adatbázison végzett kísérlet esetén. Látható, hogy a 74%-os átlaghoz képest a 10-10% körüli kitérés sem ritka egyik irányban sem, sőt, a legjobb és legrosszabb beszélő közti különbség több mint 36%! Habár az egyes felvételek közt nem csak a beszélő személye, hanem a telefonvonal, így a zajviszonyok is változnak, úgy véljük, hogy a kapott nagy szórást alapvetően a beszélők közti eltérések okozzák (mint mondtuk, a felvételek zajszintje jellemzően alacsony). Az MTBA 500 beszélőjével szemben a bemutatandó kísérletekben feldolgozott hangoskönyvet egyetlen ember olvassa fel, így a beszélőváltás mint zavaró tényező teljesen ki lesz zárva.



1. ábra. Beszédhang-felismerési pontosság eloszlása az MTBA adatbázison a beszélő személy függvényében.

Mivel a beszédadatbázisok sokáig úgy készültek, hogy kísérleti alanyokat kértek fel valamely szöveganyag felolvasására, így viszonylag későn tudatosult a kutatókban, hogy milyen jelentős eltérések vannak az olvasott és a spontán beszéd artikulációja között. Eleinte csak az tűnt fel, hogy a laboratóriumi körülmények közt elfogadhatóan működő felismerők a gyakorlatban sokkal rosszabbul teljesítenek, de csak az utóbbi 5-10 évben kezdték el a spontán beszéd jellegzetességeit közelebbről tanulmányozni. Hogy kézzelfogható értékeket is mondjunk, végeztek például olyan tesztet, melyben egy tárgyaláson felvett hanganyagot újraolvastattak ugyanazon résztvevőkkel. Az olvasott és a spontán felvételeken mért felismerési hiba között közel kétszeres faktort kaptak [12]. Magyar nyelvre Mihajlik és társai próbálkoztak spontán és tervezett beszéd (hírműsorok) ugyanazon technológiával való felismerésével [8]. Habár az eredmények nem precízen összemérhetők, hiszen a két feladat közt a beszédmódon kívül más eltérések is voltak, az általuk kapott bő kétszeres hibatényező is jól érzékelteti, hogy milyen jelentős hatékonyságromlás lép fel spontán beszéd esetén. Ez a hatékonyságromlás épp elég ahhoz, hogy a felismerők átessenek az „éppen használható” kategóriából a használhatatlanba, ezért olyan megoldással is találkozni – például egy japán tévéműsor-feliratozó rendszerben –, hogy a zajos vagy spontán részeket egy képzett beszélő megfelelő artikulációval újramondja [14]. Esetünkben mindkét

adatbázis olvasott beszédet tartalmaz, de míg a hangoskönyveket színészek olvassák lemezre, az MTBA adatbázisban bőven akadnak igénytelen beszédmódú adatközlők. Így ebből a szempontból is könnyebbnek ígérkezik a hangoskönyvek felismerése, habár az MTBA sem tartozik a legnehezebb (azaz spontán beszéd) kategóriába.

Egyetlen embertől származó hangfelvétel esetén is változhat a beszéd hangminősége, akár fizikai (pl. rekedtség), akár lelki okokból (pl. érzelmi felindultság). Ebből a szempontból talán kivételesen a hangoskönyv a rosszabb, az MTBA esetében ugyanis olyan rövid hanganyagunk van egy-egy beszélőtől, hogy ezt a jelenséget nem igazán van mód megfigyelni. Egy hangoskönyvben természetesen előfordulhat, hogy a színész hangszínének megváltoztatását kifejezőeszközként használja, de az általunk választott felvétel esetén ez kevésbé jellemző. Néhol fordul csak elő egyfajta suttogás jellegű, visszamerengő beszédstílus.

3 A hanganyag és feldolgozása

A viszonyítási alapként közölt felismerési eredményeket az MTBA adatbázison értük el, és részben már publikáltuk korábban [9]. Az MTBA adatbázisról is közöltünk már részletes leírást [10]; mint már kiderült, ez egy telefonon át rögzített korpusz, mely 500 beszélőtől tartalmaz felvételeket, melyekből mi itt az olvasott mondatokat és szavakat tartalmazó blokkot használtuk fel. A teljes adatbázis manuális fonetikai szegmentáláson és címkézésen esett át; az ennek során használt 58 címkéből viszont némelyik olyan ritkán fordul elő, hogy kénytelenek voltunk néhány összevonást eszközölni, így a kísérletekben 52 címkével dolgoztunk. A felvételekből elhagytunk bizonyos, a kézi címkézés során zajosnak talált felvételeket, így az eredeti 8000 fájl helyett csak 6935-öt használtunk fel. Ezt úgy osztottuk fel tanító és tesztelő részre, hogy előbbibe 408, utóbbiba 91 beszélő került (1 beszélő esetén az összes felvétel túl zajosnak bizonyult).

Feldolgozandó hangoskönyvnek olyan felvételt választottunk, amelynek eredeti, írott változata is jogdíjmentesen elérhető. Választásunk Krúdy Gyula Szindbád-történeteinek „Szindbád utazásai” című gyűjteményes kiadására esett, Gáspár Sándor előadásában (Kossuth kiadó – Mojzer kiadó). A felvétel teljes játékideje 212 perc, ami körülbelül fele az MTBA adatbázis időtartamának. A hanganyagot szinkronba kellett hoznunk a szöveganyaggal, ennek lépéseit ismertetjük az alábbiakban. Először is a hanganyagot végighallgattuk, a szöveghez képesti esetleges eltéréseket keresve. Ilyet kb. tucatnyi esetben találtunk csak, és viszonylag rövid szavakat érintve (többnyire indulatszavak, pl. „óh”, „ah” elhagyása vagy beszúrása a felolvasó által). A lehallgatás során vágtuk ki az egyes fejezetek végén elhangzó zenei szignált, valamint az idegen szavakat is kigyűjtöttük a fonetikai átírás előkészítéseként.

Mivel az MTBA-éhoz hasonló fonetikai szintű címkézést szerettünk volna készíteni a hangoskönyvhöz, így a következő lépés a szöveganyag fonetikai átírása lett volna. Erre egy elég sajátos megoldást alkalmaztunk, több szempontot is figyelembe véve. A szokványos út az előforduló szóalakok kigyűjtése, majd azok átírása. Az átírás azonban nem triviális dolog, több okból sem [7]. Az egyik problémát a kettős betűk okozzák, melyek azonosításához morfológiai elemzésre lenne szükség (lásd pl. „pácsó”). A másik probléma, hogy bizonyos hasonulási folyamatok fellépése szintén

függ a morfémahatárok helyétől (erre példa a /tj/ kapcsolat az „látják”, illetve „átjáró” szavakban). Ráadásul a hasonulás sok esetben opcionális, azaz többféle ejtés is helyes lehet. Erre tényleg csak az a megoldás létezik, hogy az adott szóhoz több lehetséges kiejtést is megadunk. Tipikus ilyen opcionális hasonulási pozíció a szóhatár, ahol akár kis szünetet is tarthatunk, de kiejthetjük a szomszédos szavakat szünet nélkül is, sőt a szóvégi hangok hasonulásával is. Hogy melyik következik be, az leginkább az artikuláció igényességén múlik, azaz a szövegből többnyire megjósolhatatlan. A szóhatárokon fellépő hasonulásokat a szavak izoláltan történő átírásával dolgozó módszerek többnyire nem is képesek kezelni.

A fenti okból, valamint mivel nem állt rendelkezésünkre egy kifinomult, morfológiai elemzést is figyelembe vevő fonetikus átíró, a szavankénti átírás helyett egy mássalhangzó-kapcsolatokra épülő fonetikai átírást alkalmaztunk. Ehhez abból indultunk ki, hogy a szóköz csak az írott szövegben jelent triviális tagolási határt – az akusztikumban viszont a szóhatár az egyik legkiszámíthatatlanabbul viselkedő jelenség. Miért nem választunk hát inkább olyan tagolást, amelynek határai akusztikailag stabilak? Ebből kiindulva a szöveget nem a szóközöknél, hanem a magánhangzóknál tördeltük el. Egyrészt azért esett a magánhangzókra a választásunk, mert szép artikuláció esetén nem jellemző, hogy kiesnek vagy redukálódnak (a hossz módosulásától eltekintve). Másrészt pedig a hasonulás alapvetően a mássalhangzó-klasztereket érinti, a magánhangzókon nem terjed át, így egyfajta természetes határt képez. Előnyt jelentett továbbá az is, hogy mássalhangzó-kapcsolatból jóval kevesebb van, mint szóalakból: esetünkben a 7186 különböző szóalakhoz képest csak 809 különböző mássalhangzó-kapcsolatot találtunk (a szóhatárokon átívelő kapcsolatokat is beleértve!). Így az elemek automatikus, szabályalapú fonetikai átírása után az összes elemet át tudtuk nézni, és szükség esetén kézileg korrigálni. Ezzel a megoldással a szóhatárokat kényelmesen tudtuk kezelni, például a „T SZ” betűsorhoz három lehetséges átiratot rendeltünk:

t sil s
t s
ts:

ahol “*sil*” a csend fonetikai címkéje. A módszernek természetesen van egy olyan hátránya, hogy mivel a teljes szót nem látja, így olyankor is megenged alternatívákat, amikor nem kellene, például a *pácsó* szóhoz a helyes [pa:tʃo:] mellett a hibás [pa:tjo:] átírást is fel fogja kínálni. Mindenesetre úgy gondoltuk, hogy ez kevésbé rontja a felismerő hatásfokát, mint ha egy szóhoz csak egyetlen, de esetleg hibás átirat van megengedve.

A fonetikai átírással kapott, a fentiekben ismertetett módon alternatívákat is megengedő szimbólumsorozatnak a hanganyaghoz való legjobb illeszkedését ún. kényszerített illesztéssel [5] határoztuk meg. Ehhez a HTK beszédfelismerő csomagot használtuk [13], melyet az MRBA adatbázison tanítottunk be. Ez az adatbázis szerkezetében nagyon hasonlít az MTBA-hoz, a lényeges különbség, hogy nem telefonvonalon, hanem személyi számítógépekbe dugott mikrofonokon keresztül rögzítettük [11]. Emiatt úgy éreztük, hogy felvételi körülményei jobban igazodnak a hangoskönyvéhez, és ezért talán megfelelőbb a feladathoz.

A kényszerített illesztés révén előállt annotált adatbázist kb. 80%-20% arányban osztottuk fel tanító és tesztelő részre, egész pontosan a hangoskönyv tíz Szindbád-történetéből nyolcat jelöltünk ki tanításra és kettőt tesztelésre.

4 Akusztikai modellezés a tandem technológiával

A hanganyag előfeldolgozása eltérőképpen zajlott a két adatbázis esetén, ugyanis az MTBA-s kísérletekben alkalmazkodnunk kellett egy angol rendszerhez [9]. Így ott PLP-vektorokkal reprezentáltuk a beszédjelet, míg a hangoskönyv esetében a szokványos 39 elemű kepsztrális (MFCC) együtthatóvektorok sorozatát nyertük ki [5]. Korábbi tapasztalataink alapján ez nem okoz nagy eltérést, egyik reprezentáció sem nevezhető szignifikánsan jobbnak a másiknál.

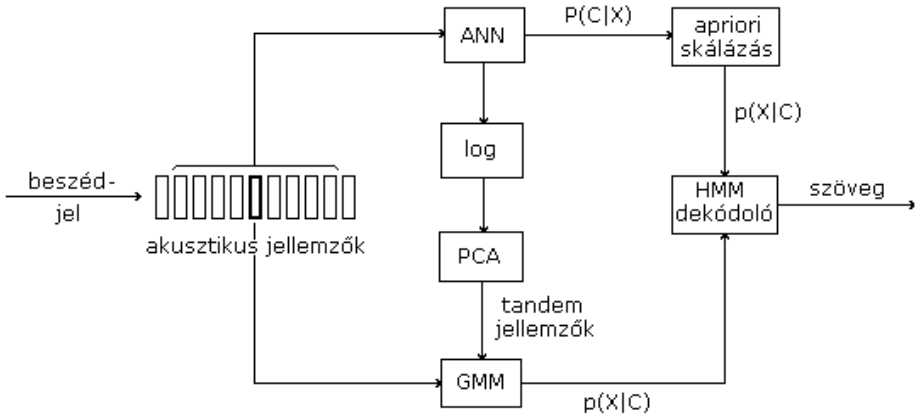
A hagyományos rejtett Markov-modelles (HMM) technológia a jellemzővektorok alapján, Gauss-keverékeloszlások illesztésével ad közelítést az egyes építőelemek (modell-állapotok) valószínűségére [5]. Mi egy másik fajta technikát használtuk, amely a gaussos modellek helyett mesterséges neuronhálót alkalmaz a lokális valószínűségek becslésére. Ez a megoldás két fő előnyt kínál a hagyományoshoz képest: egyrészt a neuronháló tanítása diszkriminatív – szemben a Gauss-keverékmodell hagyományos generatív tanításával –, ezért általában valamivel nagyobb osztályozási pontosságot tud elérni. Másrészt a neuronhálót általában nem csak egyetlen adatvektoron, hanem több (esetünkben 9) szomszédos vektoron szokták tanítani, a nagyobb környezet figyelembe vétele pedig szignifikáns javulást tud hozni. Meg kell jegyeznünk azonban, hogy a Gauss-keverékmodellhez is létezik diszkriminatív tanítási algoritmus, és annak sincs elvi akadálya, hogy több szomszédos vektoron tanítsák – egyszerűen csak valami oknál fogva ez nem terjedt el.

A neuronháló által adott kimenetek bizonyos kritériumok teljesülése esetén valószínűségi becslésként értelmezhetők, és egy apró módosítással beépíthetők a hagyományos HMM-sémába; így kapjuk az ún. HMM/ANN hibrid modellt [2]. A hibrid technológiát – főként kisebb feladatok esetén – sokan találták jobbnak, mint a hagyományos HMM-et, de a nagyobb rendszerekben mégsem bírt elterjedni. Saját tapasztalatunk az, hogy bár akusztikai szinten tényleg pontosabb, a nyelvi modellel kombinálva mégis leromlik a teljes rendszer hatékonysága. Ennek oka sejtésünk szerint az lehet, hogy a másfajta modellezési és tanítási technika miatt a neuronhálós akusztikus modellt máshogy kellene kombinálni a nyelvi modellel, mint ahogy azt a hagyományos HMM teszi.

Ezt a problémát egy huszárvágással oldja meg az ún. HMM/ANN tandem technológia [4]. Ez a neuronhálótól kapott értékeket nem valószínűségi becslésként értelmezi, hanem úgy tekinti, hogy a neuronháló egy nemlineáris transzformációt hajtott végre az akusztikus jellemzőkön; vagyis a kimenet továbbra is akusztikus jellemzővektor, pusztán egyfajta transzformált formában. Ez esetben viszont be lehet rajta tanítani egy teljesen hagyományos, Gauss-komponensekkel dolgozó rejtett Markov-modellt. Ezzel a trükkös megoldással azt mondhatjuk, hogy a rendszerben csak az akusztikai előfeldolgozó modult cseréltük le, így semmit nem kell módosítani a hagyományos, jól bevált és ezerszer letesztelt rejtett Markov-modellünkön. A megoldás

egyetlen hátránya az, hogy a rendszert duplán kell tanítani, és nyilván a kiértékeléskor is lassabb lesz.

A 2. ábra blokkdiagramja összefoglalja a hagyományos, a hibrid és a tandem modellek számítási lépéseit.



2. ábra. A hagyományos modell (alsó útvonal), a hibrid (felső útvonal) és a tandem modell (középső útvonal) sematikus összevetése.

Az elvi áttekintés után lássuk a tandem modell megvalósításának technikai részleteit. Az alkalmazott neuronháló 9 szomszédos jellemzővektoron tanult, kimenetként pedig az 52 fonetikai címke mindegyikéhez rendeltünk egy-egy kimenő neuront. A MTBA-n végzett tesztek során a rejtett réteg neuronjainak száma 4800 volt, ugyanis szinkronban kellett lennünk az említett angol modellel. A hangoskönyv esetén csupán 500 rejtett neuronnal dolgoztunk, mivel a neuronszám további növelése nagyobb futásidő-növekedéssel jár, mint amennyit az eredményeken javít. A neuronhálót mindkét esetben backpropagation algoritmussal tanítottuk be, az adatok 10%-án számított keresztvalidációt használva megállási kritériumként. A tanítási célértékeket természetesen a kényszerített illesztés során kapott fonetikai címkék képezték.

A neuronháló által kiadott vektorokon a HTK csomag rejtett Markov-modelljét tanítottuk be [13]. Akusztikus komponensként 3-állapotú monofón beszédhangmodelleket képeztünk, állapotonként 9-9 Gauss-eloszlással. Az irodalom javaslata szerint a neuronháló kimenő értékeit a HMM-be való beengedés előtt érdemes logaritmizálással Gauss-görbéhez jobban igazodó alakúra hozni, valamint főkomponens-analízissel dekorrelálni. Mi is így tettünk, ugyanis saját méréseink is alátámasztották az említett trükkök hasznosságát [9]. Egy további trükk a neuronhálókimeneteknek az eredeti akusztikai vektorokkal együtt való használata, azaz a két vektor konkatenálása. Habár a két vektor elvileg redundáns, a gyakorlatban egy minimális javulást ez a fogás is tud hozni, így mi is alkalmaztuk. Így összességében a HMM inputját képező jellemzővektor 91 komponensű volt a szokványos 39 helyett. A beszédhangmodellek tanításához a hagyományos, maximum-likelihood kritériumot optimalizáló algoritmusok mellett diszkriminatív (MMI-hibakritériumot alkalmazó)

tanítást is bevetettük [3]; szerencsére a HTK csomag tartalmazza ennek implementációját.

A novellák szöveganyagát kevésnek éreztük egy szószintű nyelvi modell (*N*-gram) betanításához, egy általános, kortárs korpuszokon tanított nyelvi modell pedig nem igazán illett volna a regény majd' száz éves szókincséhez. Ezért nyelvi modellként beszédhangszintű modellezéssel próbálkoztunk: a HTK eszköztárát használva a tanítókorpusz fonetikai címkéiből beszédhang-bigramokat számoltunk. Továbbá mivel a kényszerített illesztésnél alkalmazott módszer miatt rendelkezésünkre állt a szöveganyag magánhangzó-mássalhangzókapcsolat elemekre való felbontása, kézenfekvően adódott, hogy ezekből is megpróbáljunk bigramot képezni. Erre a nyelvi modellre jobb híján „szótag"-bigramként fogunk hivatkozni, bár az elemei csak méretükben hasonlítanak a nyelvészeti értelemben vett szótagokhoz.

5 Eredmények és diszkusszió

Legelső lépésként a rejtett Markov-modellt teljesen hagyományos módon, azaz közvetlenül az akusztikus jellemzővektorokon tanítottuk be. Az így kapott értékeket viszonyítási alapként használhatjuk a tandem-reprezentáció, azaz a neuronháló segítségével végzett transzformáció hasznosságának megítélésében. Az első tesztekben semmiféle nyelvi modellt nem használtunk, hogy az eredmények tisztán az akusztikus modellek hatékonyságát tükrözzék. Az MTBA adatbázison 53,37%-os, míg a hangoskönyv esetén 72,18%-os pontossággal egyezett a felismerő által kiadott és a címkézés szerint a fájlhoz tartozó leirat (pontosságon a két sztring szokványos, angol terminológiával „accuracy"-nek nevezett illeszkedését értve). Már magában ez az érték is jól mutatja, hogy a hangoskönyv mennyivel könnyebb felismerési feladatot jelent.

A következő lépés a hagyományos jellemzőkről a tandem jellemzőkre való áttérés volt. Ennek első fázisa a neuronháló betanítása az osztálycímkék felismerésére. Ennek eredményessége a rejtett Markov-modellbe való beépítés előtt is tesztelhető, bár ilyenkor persze még csak az egyes adatvektorokra vonatkozó osztályozási pontosságot tudjuk vizsgálni. A neuronháló 74,11%-os felismerést tudott elérni az MTBA esetén, míg a hangoskönyvön 85,24%-ot produkált. Mivel ezek a pusztán lokális értékek jóval magasabbak, mint a HMM-mel kapott globális eredmények, jó eséllyel várhattuk, hogy az ezekre épülő teljes modell is lényegesen jobb lesz.

A rejtett Markov-modell tandem jellemzőkkel történt betanítása után kapott eredményeket az 1. táblázat 2. sorában találhatjuk. Látható, hogy a tandem technikának köszönhetően mindkét adatbázison jelentősen, és körülbelül ugyanolyan mértékben (kb. 25%-kal) csökkent a felismerési hiba.

Harmadik finomítási lépésként a nyelvi modellek bevetésével folytattuk. A táblázat 3. sora mutatja a beszédhang-bigrammal kapott eredményeket. Mivel a „szótag"-jellegű felbontást csak a hangoskönyvön csináltuk meg, így az ezekre épülő bigramot is csak ezen az adatbázison értékeltük ki; az eredmény a táblázat 5. sorában található. Mint az várható volt, a kétféle nyelvi modell közül a nagyobb egységekkel dolgozó szótagalapú hozott nagyobb javulást.

1. táblázat: beszédhang-felismerési pontosságok a két adatbázison, különféle akusztikai és nyelvi modellek esetén.

	MTBA	Hangoskönyv
HMM hagyományos jellemzőkkel (nyelvi modell nélkül)	53,37%	72,18%
HMM tandem jellemzőkkel (nyelvi modell nélkül)	65,09%	79,49%
Tandem beszédhang-bigram nyelvi modellel	69,67%	83,62%
Tandem + beszédhang-bigram + diszkriminatív tanítás	73,93%	86,26%
Tandem szótag-bigram nyelvi modellel	---	84,58%
Tandem + szótag-bigram + diszkriminatív tanítás	---	86,33%

Utolsó lépésként bevetettük a HTK diszkriminatív tanítási algoritmusát. Mivel ez a módszer a teljes rendszert finomítja, így mindkét nyelvi modell mellett le kellett futtatnunk a tanítást. A diszkriminatív tanítás újabb 13-15 százalékkal csökkentette a hiba mértékét, ennek köszönhetően a telefonos adatbázison sikerült megközelíteni a 75%-os pontosságot. A hangoskönyvön a kétfajta nyelvi modell között csökkent a különbség, a végeredményként kapott 86,26% és 86,33% közt nincs jelentős eltérés.

A táblázat értékei jól mutatják a két adatbázis által prezentált felismerési feladat nehézségi különbségét: az MTBA adatbázison elért legjobb eredmény alig jobb, mint a hangoskönyvön a legegyszerűbb megoldással kapott érték! Érdekességképp megjegyezzük, hogy a korábban az 1. ábrán bemutatott hisztogram az MTBA-n elért 73,93%-os átlaghoz tartozik, és az ábrán szereplő legmagasabb, 85%-os érték gyakorlatilag megegyezik a hangoskönyvön kapott pontossággal. Tehát az MTBA-n betanított modell is el tudta érni ugyanazt a hatékonyságot, de csak a számára „legszimpatikusabb” beszélőn – a többiek sajnos lehúzták az átlagot.

SZ I N B Á D A Z E L Ő T M E SS Ú T A K A I S H A J D O N D Ó
V O L T H E CC O K NY A F O D R O CS K A CS A L G O TT A -
- M E N P E S T Ö BB U D Á R A -- A N É P R G E TT Ó L A M A
R I CC G E T I V A TT A L E MM É K T O V A B I S -- D E M O S
T -- A L E KK Ö Z E L E BB É S A L O K I K S E M E N T H A --
E S O N M E LL E TT E TY K E D V E S T É S N Ő -- A K I N E K
F E H É R F Á TTY A L A V O L T É SS A L G O S F É R C I P Ő

3. ábra. Példa a beszédfelismerő fonetikai szintű kimenetére.

Az eredmények jól érzékeltetik a tandem technológia hasznosságát is. Meg kell azonban jegyeznünk, hogy az összes kísérletben kizárólag monofón HMM-eket alkalmaztunk. A táblázat 1. sorában összehasonlításként szereplő eredmények feltehetően sokkal magasabbak lennének trifón modelleket használva. A tandem eredmények viszont kevésbé javulnának, ugyanis a neuronháló tanítása elég nehezen házasítható össze a trifón modellezéssel, és ennek optimális megoldása jelenleg a tandem-jellegű módszerekkel foglalkozók egyik legfontosabb kutatási problémája (lásd pl. [1]).

A tandem technológia egyik sajátossága, hogy a neuronháló révén rögtön az adatvektorok szintjén is tudunk mondani részeredményt; a hagyományos HMM-es technológiában ez nem szokás (bár megoldható lenne). Pedig érdekes tanulságokat kínálna annak részletes kielemzése is, hogy vajon a bigram modellel is megtámogatott globális eredmény miért nem jobb lényegesen, mint a neuronháló által a puszta adatvektorokon elért pontosság (73,93% vs. 74,11%, illetve 86,33% vs. 85,24%). Ez a meglepő megfigyelés azt sejteti, hogy a lokális hibák nem egyenletesen oszlanak el, hanem bizonyos hosszabb-rövidebb szakaszokon felhalmozódnak. E hipotézis igazolása azonban mélyre hatóbb kivizsgálást igényelne.

A fonetikai szintű kimenet mellett természetesen nagyon érdekes lenne szószintű eredményeket is látni, a fent kapott értékekből ugyanis nem lehet tudni, hogy vajon a szavakat milyen arányban tudná eltalálni egy szómodelleket is tartalmazó rendszer. A korábban ismertetett okok miatt sajnos nem állt módunkban komolyabb nyelvi modellel is kipróbálni a felismerést; végeztünk azonban egy olvasási tesztet, melynek során a kísérleti személy azt a feladatot kapta, hogy „fejtse meg” a felismerő kimenetét, azaz legjobb tudása szerint javítsa értelmes magyar szöveggé. Erre tetszés szerinti idő állt rendelkezésére, és a szövegben is oda-vissza ugrálhatott. Feladványként a tesztadatbázisba került két Szindbád-történet egyikét kapta meg (melyet korábban még nem olvasott). A 3. ábra egy részletet mutat a dekódolandó betűsorozatból. Kísérleti alanyunknak a szöveg 1337 szövegszavának 94,24%-át sikerült eltalálnia. Meg kell jegyezzük, hogy bár szigorú értelemben 77 szót nem talált el, a hibák túlnyomó többsége csak egyetlen betű vagy morféma eltéréséből állt, és értelemzavarónak csak szűk tucatnyit lehetne nevezni. Ez elég elgondolkoztató arra nézve, hogy az eltalált szavak száma mennyire értelmes mérőszáma a felismerés pontosságának. További észrevételünk, hogy habár az ember által szimulált „nyelvi-szemantikai modell” nyilván összehasonlíthatatlanul ügyesebb, mint a gép, valós szituációban az utóbbi annyival könnyebb helyzetben van, hogy az akusztikai modelltől nem csak a legvalószínűbb megoldást kapja meg, hanem további lehetőségeket is (ún. *N*-best list vagy lattice). Hasonló segítség birtokában feltehetően kísérleti személyünk is még jobb eredményt tudott volna elérni.

6 Összegzés

Cikkünkben egy hangoskönyvön végeztünk beszédfelismerési kísérleteket annak felmérésére, hogy egy ilyen gyakorlatilag optimálisnak nevezhető hangfelvétel esetén milyen felismerési pontosságra képes rendszerünk. Az eredményeket az MTBA telefonbeszéd-adatbázison végzett hasonló tesztek eredményeivel párhuzamba állítva

igazolódott sejtésünk, hogy a hangoskönyv lényegesen egyszerűbb felismerési feladatot jelent. Fonetikai szinten 86%-os pontosságot sikerült elérnünk, ami már szabad szemmel is jórészt értelmezhető kimenetnek felel meg. További legfontosabb feladatnak a tesztek magasabb szintű nyelvi modellel való megtámogatását tartjuk, illetve tervezzük a felismerési hibák jellegzetességeinek elemzését is, ami rálátást adhat az akusztikai modell további javításához.

Hivatkozások

1. Aradilla, G., Bourlard, H., Magimai-Doss, M.: Using KL-based Acoustic Models in a Large Vocabulary Recognition Task. In: Proceedings of Interspeech 2008 (2208) 928–931
2. Bourlard, B., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic (1994)
3. He, X., Deng, L.: Discriminative Learning for Speech Recognition: Theory and Practice. Morgan & Claypool (2008)
4. Hermansky, H., Ellis, D., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of ICASSP 2000 (2000) 1635–1638
5. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing. Prentice Hall (2001)
6. Lippmann, R. P.: Speech Recognition by Machines and Humans. Speech Communication, 22(1) (1997) 1–15
7. Mihajlik P., Tatai, P.: Automatikus fonetikus átírás magyar nyelvű beszédhez. Beszédkutatás 2001 (2001) 172–185
8. Mihajlik P., Tarján B., Tüske Z., Fegyó T.: Investigation of Morph-based Speech Recognition Improvements across Speech Genres. In: Proceedings of Interspeech 2009 (2009) 2687–2690
9. Tóth L., Frankel, J., Gosztolya G., King, S.: Cross-lingual Portability of MLP-Based Tandem Features - A Case Study for English and Hungarian. In: Proceedings of Interspeech 2008 (2008) 2695–2698
10. Vicsi K., Tóth L., Kocsor A., Gordos G., Csirik J.: MTBA - magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika, Vol. LVII, No.8 (2002) 35–43
11. Vicsi K., Kocsor A., Teleki Cs., Tóth L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: II. Magyar Számítógépes Nyelvészeti Konferencia (2004) 315–318
12. Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A.: Effect of speaking style on LVCSR performance. In: Proceedings of ICSLP 1996 (1996) 16–19
13. Young, S. et al.: The HMM Toolkit (HTK) – software and manual. http://htk.eng.cam.ac.uk (1995)
14. Zhao, Y.: Speech-Recognition Technology in Health Care and Special-Needs Assistance. IEEE Signal Processing Magazine, 26(3) (2009) 87–90