Egy magyar nyelvű beszédfelismerő rendszer szószintű hibáinak elemzése

Gosztolya Gábor[1,2], Vincze Veronika[1], Grósz Tamás[2], Tóth László[1]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: {tothl, ggabor, vinczev}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail: groszt@inf.u-szeged.hu

Kivonat Az automatikus beszédfelismerő rendszerek szószintű hibáját hagyományosan egy illesztési távolságon alapuló metrikával mérjük, amely a szóalakok pontos egyezésének vizsgálatán alapszik. Mint a legtöbb beszédfelismerési technika, ez is jól illeszkedik az angol nyelvre, más (pl. ragozó) nyelvekre azonban ez nem feltétlenül igaz. Ebben a cikkben azt vizsgáljuk, hogy egy hagyományosnak számító beszédfelismerő rendszer (mély neuronhálós akusztikus modell és szó-trigram nyelvi modell) milyen jellegű hibákat vét. Ehhez száz hangfelvétel hibáit gyűjtöttük ki, annotáltuk manuálisan, majd elemeztük. Végkövetkeztetésünk, hogy a szótárban nem szereplő elemek mellett nagy gondot okoz a magyar nyelvben az egybe- és különírások kezelése, melyet a hagyományos pontosságmetrika különösen nagy mértékben büntet. [1]

Kulcsszavak: beszédfelismerés, illesztési távolság, N-gram modell

1. Bevezetés

Az automatikus beszédfelismerő rendszerek fejlesztésében hamar dominánssá vált, és mindmáig az is maradt az angol nyelvű beszéd felismerése. Emiatt a beszédfelismeréssel foglalkozó kutatók általában olyan technikákat dolgoztak ki, melyeknél a fő kritérium az volt, hogy angol nyelvre jól működjenek; természetes, hogy a más nyelvekre beszédfelismerő rendszert fejlesztők és a téma kutatói ugyanezeket vagy nagyon hasonló technikákat alkalmaznak saját nyelvükre. Ez azonban sokszor nem a legszerencsésebb, hiszen egy-egy nyelv speciális tulajdonságai indokolhatják, hogy az adott problémát árnyaltabban közelítsük meg. Erre egy jó példa az ún. N-gram nyelvi modell, melyben egy-egy szóalak előfordulási valószínűségét statisztikai módszerekkel, a megelőző $N-1$ szóalak alapján becsüljük; ez a megközelítés értelemszerűen elég jól illeszkedik az angol nyelvhez, egy ragozó nyelvnél (mint pl. a magyar) viszont indokolt (lenne) az eljárás finomítása (pl. nyelvtani kategóriákra [1] vagy morfokra [2] számolt N-gram).

[1] A jelen kutatás során használt TITAN X grafikus kártyát az NVIDIA Corporation ajándékozta csoportunknak.

Egy másik elterjedt beszédfelismerési technika, mely szintén a szóalakok különbözőségének elvén alapszik, maga a beszédfelismerő rendszer pontosságát mérő metrika: általában az illesztési (vagy Levenshtein-) távolságon [3] alapuló pontosságértéket szokás alkalmazni. Ebben a beszédfelismerő rendszer kimenetét a helyes szöveges átirathoz hasonlítva megszámoljuk a beszúrt, törölt illetve kicserélt szóalakok számát, és ezen értékek (valamint a helyes átiratban szereplő szóalak-előfordulások száma) alapján számítjuk ki a rendszer százalékban kifejezett pontosságát. Nyilvánvaló, hogy egy agglutinatív nyelv esetében ez a megközelítés is vezethet problémákhoz, ugyanakkor tudtunkkal itt fel sem merült más metrika használata. (Bár az eljárás finomításaként értékelhetjük, hogy a távolkeleti nyelvek (pl. japán vagy kínai) esetén ugyanilyen módon, de nem szóalak-, hanem karakteralapon számítják ki a beszédfelismerő rendszer pontosságát.)

Ha felmérjük, hogy beszédfelismerő rendszerünk milyen jellegű hibából vét (relatíve) sokat, akkor az adott jelenséget célirányosan kezelve jelentősen javíthatjuk a felismerés pontosságát. Emiatt érdekes lehet, hogy egy beszédfelismerő rendszer szószintű kimenetében milyen jellegű tévesztések fordulnak elő gyakrabban; ugyanakkor a jellemző hibatípusokat nyugodtan nevezhetjük közismertnek. Köztudott, hogy a hibák egy jelentős része visszavezethető a felismerési szótárból hiányzó (Out-of-vocabulary, OOV) szóalakokra, azon belül különösen két nagyobb csoportra: a tulajdonnevekére és a számnevekére (ez utóbbiba beleértünk egyéb, számokhoz kapcsolódó szóalakokat is, pl. *harmincezren*). Tulajdonnevekből és egyéb névelemekből nagyon sok alak létezik, ráadásul ezek gyakorisága a beszéd témájától, sőt a beszéd elhangzásának idejétől függően változik [4]. A számnevek esetében szintén a szóalakok nagy (gyakorlatilag végtelen) száma okoz gondot. A két típusban közös, hogy egyrészt nem lehet az összes lehetséges szóalakot felvenni a szótárba, másrészt az OOV szó helyére beerőltetett, hibás szóalak a nyelvi modell által a következő szavakra adott becslést is lerontja.

Az OOV szóalakok további köztudott tulajdonsága, hogy hajlamosak „elrontani" az előfordulásuk közvetlen környezetét is. Ha egy beszédfelismerő rendszer egy általa ismeretlen (és így felismerhetetlen) szóalakkal találkozik, jellemző, hogy az érintett részre valamely akusztikailag nagyon hasonló, ám a szótárban szereplő szóalakot illeszt. Amennyiben az akusztikailag hasonló szó rövidebb vagy hosszabb, akkor a szótévesztések a szomszédos szavakra is kiterjednek (pl. *biztosít többséget* vs. *biztosítók siket*). A másik mellékhatás, mikor a rendszer az ismeretlen szót több, a szótárban szereplő szóból próbálja meg kirakni (pl. *huszonkilencedikére* vs. *ózon kilencedikére*), ekkor ugyanis két vagy több szótévesztést (egy szócserét és egy vagy több szóbeszúrást) tapasztalunk, amelyek nagyobb súllyal esnek latba a rendszer szószintű pontosságának számításakor.

Jelen cikkünkben azt vizsgáljuk, hogy egy, a fönti szempontok alapján hagyományosnak számító megközelítés hogyan viselkedik magyar nyelvre. Ehhez egy, ma a legmodernebbnek számító akusztikus, és egy hagyományos szóalak trigram (3-gram) nyelvi modellel rendelkező beszédfelismerő rendszerrel végzünk felismerést magyar nyelvű híradófelvételeken [5], majd a szószintű hibákat manuálisan kategorizáljuk és elemezzük. Végül kísérletet teszünk arra is, hogy felmérjük a felismerő kimenetének olvashatóságát.

2. A tesztkörnyezet

A következőkben bemutatjuk a tesztkörülményeket: a beszédfelismerő rendszer akusztikus modelljét, a hangfelvételek szöveges átiratainak készítését, valamint az alkalmazott nyelvi modellt.

2.1. Akusztikus modell

Akusztikus modellként egy mély egyenirányított neurális hálót alkalmaztunk [5]. Mély neuronhálónk tanításához egy teljesen GMM-mentes módszert használtunk [6], melynek lényege, hogy az átirat időbeli illesztését és a környezetfüggő állapotok klaszterezését is csak mély neuronhálókra támaszkodó módszerek segítségével végezzük.

Mivel kezdetben nem állt rendelkezésünkre a szöveges átirat időbeli illesztése, első lépésként egy maximális kölcsönös információn (Maximum Mutual Information, MMI [7]) alapuló módszerrel tanítottunk egy környezetfüggetlen neuronhálót. A betanított háló kimenetei alapján elvégeztük az annotáció kényszerített illesztését, majd az illesztések további finomításához tanítottunk egy újabb hálót immár a hagyományos keresztentrópia-alapú kritériumra optimalizálva. Az új háló alapján újraillesztettük a címkéket, és a továbbiakban az így kapott keretszintű címkézést használtuk.

A kényszerített illesztés elvégzése után a tavaly bemutatott Kullback-Leibler–divergencián alapuló klaszterezést alkalmazva állítottuk elő az összevont környezetfüggő állapotokat [8]. A végső akusztikus modellként használt mély neuronhálót az így kapott környezetfüggő osztályok felismerésére tanítottuk a hagyományos keresztentrópiát optimalizáló hiba-visszaterjesztéses algoritmussal. A konkrét modellünkben 1843 környezetfüggő állapotot használtunk; a neuronháló öt rejtett rétegből állt, rétegenként ezer-ezer ún. rectifier neuronnal.

2.2. Nyelvi átírás és -modellezés

A beszédadatbázis ortografikus átiratának elkészítésekor a következő alapszabályokat alkalmaztuk. A szöveges átiratok csak nagybetűket tartalmaznak, hogy a tulajdonnevek nagy kezdőbetűjéből eredő szótévesztéseket kizárjuk. A számneveket minden esetben ortografikusan írtuk át. A kötőjelet tartalmazó összetett szavak esetén a kötőjel helyett szóközt írtunk. A rendhagyó szóalakokat (pl. idegen szavak) kiejtés szerint írtuk le, hogy a fonetikus átíró feladatát megkönnyítsük. A két utóbbi lépés a beszédkorpuszban szereplő átiratok és a nyelvi modell szókészlete között nyilván eltéréseket okoz, ennek hatását a továbbiakban elemezni fogjuk.

A szótár és a nyelvi modell kialakításához két forrást használtunk fel. A nyelvi modell alapját az origo hírportál anyagából készített szövegkorpusz képezte, ami kb. 50 millió szövegszóból állt. Mivel a korpusz sok hibás szóalakot tartalmazott, szükségünk volt egy megoldásra a hibák kiszűrésére. Ezért a rendszer által elfogadott szóalakok listáját leszűkítettük a Magyar Kiejtési Szótár szólistájára [9]. Ez kb. másfél millió, korpuszokból kigyűjtött szóalakot tartalmaz,

ezért azt feltételeztük, hogy a legtöbb fontos szóalak fellelhető lesz benne. A origo korpuszon megvizsgáltuk ezen másfél milló szóalak gyakoriságát, és csak a legalább kétszer előforduló alakokat tartottuk meg. Ez a lépés felismerőnk szótárának méretét 486 982 szóra redukálta. Az origo korpusz alapján trigram nyelvi modellt készítettünk a HTK nyelvi szubrutinjait az alapértelmezett értékekkel használva [10], míg a szótár szavainak kiejtését a Magyar Kiejtési Szótárból [9] vettük.

Úgy véljük, hogy a fönti (etalon) szöveges átirat és a nyelvi modell tekinthetőek standard és ésszerű módon elkészítettnek, így a cikkünkben ezután megjelenő tapasztalatok és következtetések tekinthetőek általános érvényűnek (legalábbis magyar nyelvre).

2.3. Egyéb kísérleti körülmények

Kísérleteinket a „Szeged" magyar nyelvű híradós beszédadatbázison [5] végeztük. Az adatbázis összesen 28 órányi hangzóanyagot tartalmaz, melyet a szokásos felosztásban használtunk: 22 órányi anyag volt a betanítási rész, 2 órányi a validációs halmaz, a maradék 4 órányi hanganyag pedig a tesztelésre szolgáló blokk.

Az elemzést az adatbázis teszthalmazán, annak is egy 100 hangfelvételből álló részhalmazán végeztük. A teljes teszthalmazon a felismerési pontosság 84, 14% volt, míg a vizsgált részen 85, 69%; utóbbiban összesen 4214 szóelőfordulást számoltunk.

3. A hibák elemzése

Az előforduló hibatípusok elemzéséhez manuálisan néztük át a teszthalmaz egy részének annotációját és az adott részre a beszédfelismerő rendszer kimenetét. Ehhez automatikusan kigyűjtöttük a tévesztett részeket, majd azokat és a felismerő illesztett eredményét egy-egy szomszédos szóval kiegészítve megjelenítettük. Ezután az egyes tévesztéseket manuálisan kategóriákba soroltuk.

Az egyes hibákat először nyelvészeti szempontok alapján kategorizáltuk. Ilyen volt például az egybeírás/különírás: ez esetben a beszédfelismerő rendszer által készített átirat és az etalon szöveg mindössze egy (vagy több) szóköznyi eltérést mutatott (pl. *a két százmilliárdos tétel* vs. *a kétszáz milliárdos tétel*, *az exportdinamikája is* vs. *az export dinamikája is*). Visszatérő hiba volt az is, ha egymás után következett két azonos hang, melyet egy (hosszú) hangnak tekintett a rendszer. Ebben a kategóriában különösen gyakran egy *a*-ra végződő szót követő *a* névelő okozta a hibát (*mondja bankszövetség* vs. *mondja a bankszövetség*). Sok esetben magát a szót/szótövet jól felismerte a rendszer, azonban a hozzá kapcsolódó toldalékok esetében hibázott: lemaradt a toldalék (*Mezőtúr polgármester* vs. *Mezőtúr polgármestere*), esetleg hibás toldalék került a szó végére (*szétdarabolják* vs. *szétdarabolták*). Bizonyos esetekben az átiratból kimaradt egy szó (*terén erősítik* vs. *terén ha erősítik*). Két olyan hibatípussal is találkozhattunk, amikor a beszédfelismerő kimenete helyes volt, mégis eltért az

etalontól. Az egyik hibatípusnál maga az etalon átirat tartalmazott hibát (*a szennyezetett víztől* vs. *a szennyezett víztől*), míg a másik hibatípus esetében az etalon átirat készítési elveinek megfelelően a rendhagyó kiejtésű tulajdonnevek fonemikus átiratban szerepeltek a szövegben, ugyanakkor a beszédfelismerő az eredeti helyesírás szerint tüntette fel ezeket (*Magyar Helszinki Bizottság* vs. *Magyar Helsinki Bizottság*). A szótárban nem szereplő szavak esetében megfigyelhetjük, hogy azokat a rendszer gyakran fonetikailag hasonló tulajdonságokkal bíró hangokból álló szóval helyettesíti, például a *be* és *de* szavak összecserélése során ugyanúgy zöngés zárhangot találunk a szó elején.

A hibatípusok megoszlásán kívül azt is vizsgáltuk, hogy az egyes hibatípusok jellemzően milyen jellegű szavak környezetében fordulnak elő. Hogy ezt megtehessük, négy tényezőt vizsgáltunk. Amennyiben az adott tévesztéshez tartozó etalon-átiratban bármelyik szóra igaz volt a vizsgált feltétel (pl. a három érintett szóból az egyik hiányzott a szótárból), az érintett hibaelőfordulásra bejelöltük az adott tulajdonságot.

Először azt vizsgáltuk, hogy szerepel-e az etalonban névelem (pl. *Balogh*, *Fidesz*, *tálibok*). Másodszor azt ellenőriztük, hogy szerepel-e benne számnév vagy számmal kapcsolatos szóalak (pl. *ezeréves*, *kétmilliós*, *ezerkilencszázötvenhatos*). Ezután azt is megnéztük, hogy van-e az adott annotációban olyan szó, amely nem szerepel a beszédfelismerő rendszer szótárában (OOV). Végül azt vizsgáltuk, hogy az etalonnak tekintett annotáció helyes-e, vagy esetleg hibát tartalmaz. Ez jellemzően egybe-különírási hiba volt; természetesen ez nem feltétlenül jelenti azt, hogy az etalon valóban hibás, hanem tükrözheti azt is, hogy az annotáció más elvek szerint készült, mint ahogyan a szótár és a nyelvi modell felépült.

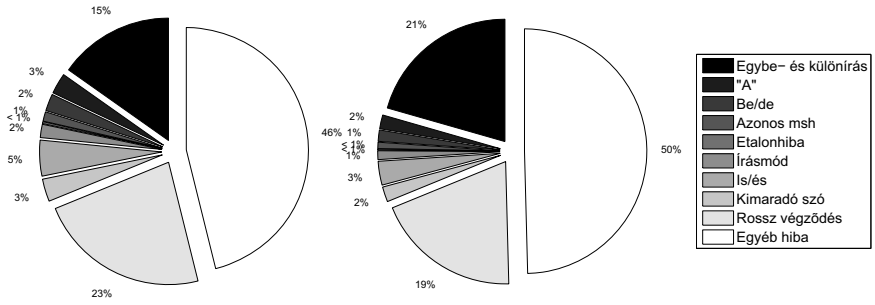
Mivel a beszédfelismerő rendszer szószintű hibáját a korábban ismertetett, illesztési távolságon alapuló módszerrel szokás mérni, logikus a hibák darabszáma mellett a hozzájuk tartozó *szótévesztések* számát is vizsgálni, ezért ezeket is feljegyeztük.

4. Eredmények

Az 1. ábrán látható az egyes hibatípusok megoszlása a hibák darabszámának és a szótévesztések számának arányában.

Az egyes hibatípusok, illetve azokon belül az egyes annotált szótípusok megoszlása az 1. és 2. táblázatokban található. Látható, hogy a tévesztések kicsit több mint 50%-át lehetett besorolni valamilyen informatív hibakategóriába, így kb. 46%-uk az „Egyéb hibák” közé került. A szótévesztéseknek ez valamivel nagyobb részét, szinte pontosan a felét tette ki, ami arra vezethető vissza, hogy bizonyos hibakategóriák (pl. be/de vagy is/és tévesztés, egymás után két „a”, kimaradó szó, írásmódeltérés) esetében jellemzően egy hibára egyetlen szótévesztés jut, míg ez az érték átlagosan 1,5. Az egybe- és különírás hibatípus azonban a felismerési hiba nagyobb részéért felelős, hiszen ilyenkor legalább két szótévesztés jut minden felismerési hibára.

A névelemeket érintő hibák között természetszerűleg voltak írásmódeltérési hibák, illetve gyakori volt a rossz végződés is. Ez nem meglepő, hiszen az egyes



1. ábra. A hibák megoszlása az egyes hibakategóriák között a hibák darabszámának (balra) és a szótévesztések számának (jobbra) arányában

1. táblázat. Az egyes hibatípusok előfordulásának száma, illetve ezen belül az egyes annotált szótípusok előfordulásának száma

Hibatípus	Névelem	Számnév	OOV	Annot.	Össz.
Egybe- és különírás	3	25	25	14	61
Egymás utáni két „a”	0	0	0	0	11
Be/de tévesztés	0	0	9	0	9
Egymás után két azonos msh	0	0	0	0	5
Etalonhiba	0	0	1	1	1
Írásmódeltérés	7	0	6	0	7
Is/és tévesztés	0	0	0	0	19
Kimaradó szó	0	0	0	0	12
Rossz végződés	7	3	20	0	91
Egyéb hiba	96	6	114	0	185
Hibák összesen	113	34	175	15	401

névelemek eleve elég ritkán fordulnak elő a tanítószövegben, így a ragozott alakjaik sem túl gyakoriak. Mégis, a névelemek nagy részét érintő hibák az Egyéb kategóriába estek.

A számneveket érintő hibák nagy többsége egybe- és különírási tévesztés volt. Kézenfekvő lenne ezt betudni annak, hogy nagyon sok számnévi szóalak képezhető, melyeket képtelenség felsorolni egy szótárban, ugyanakkor a 25 esetből csak 5 olyan volt, ahol egyúttal OOV szó is szerepelt az átiratban. A gondot a számneveknél valószínűleg az okozta, hogy a nyelvi modell *mindkét* írásmódot képes előállítani (pl. a *kétszázharmincezer* szó esetén mind a *kétszázharminc*, mind az *ezer* szó szerepelhet (és szerepelt is) a szótárban); illetve tizenegy esetben a számneveket érintő egybe- és különírási hiba annotációs hibával is egybeesett.

Az OOV szónál történt tévesztéseknek együtt kb. negyedét tették ki az egybe- és különírási, valamint a szuffixhibák, a nagy többségüket az egyéb hibák közé soroltuk. Ennek valószínűleg az a magyarázata, hogy ehhez a két kategóriához az szükséges, hogy legalább a szó egy eltérő ragozású alakja szerepeljen a szótárban;

2. táblázat. Az egyes hibatípusokhoz tartozó szótévesztések száma, illetve az egyes hibatípusokon belül az egyes annotált szótípusokhoz tartozó szótévesztések száma

Hibatípus	Névelem	Számnév	OOV	Annot.	Össz.
Egybe- és különírás	6	52	52	28	124
Egymás utáni két „a”	0	0	0	0	11
Be/de tévesztés	0	0	9	0	9
Egymás után két azonos msh	0	0	0	0	5
Etalonhiba	0	0	1	1	1
Írásmódeltérés	7	0	6	0	7
Is/és tévesztés	0	0	0	0	19
Kimaradó szó	0	0	0	0	12
Rossz végződés	11	5	32	0	116
Egyéb hiba	157	11	188	0	299
Hibák összesen	181	68	288	29	603

3. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák száma

Szótípus	Névelem	Számnév	OOV	Annot.	Össz.
Névelem	113	0	99	0	113
Számnév	0	34	10	11	34
OOV	99	10	175	1	175
Annot.	0	11	1	15	15
Összesen	113	34	175	15	216

amennyiben ez sem áll fenn, a beszédfelismerő rendszer valamilyen egyéb, hasonló hangzású szót fog beerőltetni az adott helyre (és ezzel esetleg a környezetet is elrontja). Az olyan tévesztési helyek, ahol az annotáció nem volt helyes (vagy konzisztens), általában egybe- és különírási hibához vezettek; egy esetben pedig az annotáció egyszerűen el lett gépelve (*szennyezetett*).

Az egyes tévesztési típusok felől közelítve látható, hogy az egybe- és különírási tévesztések nagyon nagy része történik olyan helyeken, ahol valamelyik jelölt szótípus előfordul az annotációban; ezek teszik ki az ilyen típusú hibák kb. 80%-át. A be/de tévesztések mindegyike egyúttal OOV hiba is, aminek az a triviális oka van, hogy a „be” szó valahogyan kimaradt a szótárból. Nem meglepő, hogy az írásmódeltérések kizárólag névelemeket érintenek, az már annál inkább, hogy egy esetben nincs szó OOV-ről. Ennek az az oka, hogy mind az *Attilának*, mind az *Atillának* szóalak szerepelt a szótárban.

Az egyéb, máshova besorolhatatlan hibák több mint felében névelem is előfordult, kétharmadukban pedig a szótárban nem szereplő szó is. Az összes előforduló hibát tekintve is magas (bár ennél alacsonyabb) arányokat láthatunk; összességében csak a tévesztések kb. 54%-ánál nem volt jelen egyik jelölt szókategória sem, igaz, ezek adták a felismerési hiba kb. 60%-át.

A 3. és 4. táblázat mutatja, hogy az egyes jelölt szótípusok mennyire estek egybe. (Értelemszerűen a táblázat főátlója megegyezik az összesítő sorral és -

4. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák szótévesztéseinek összege

Szótípus	Névelem	Számnév	OOV	Annot.	Össz.
Névelem	181	0	161	0	181
Számnév	0	68	22	22	68
OOV	161	22	288	1	288
Annot.	0	22	1	29	29
Összesen	181	68	288	29	360

oszloppal.) Látható, hogy a névelemmel egybeeső tévesztések nagyon nagy része (87%-a) egyúttal OOV is; fordítva ez értelemszerűen jóval kisebb (53%), hiszen sok más szóalak-típusra is jellemző lehet, hogy hiányzik a szótárból (pl. ragozott alakok). A számnevek kb. egy-egyharmada OOV és annotálási hiba. Föltűnő még, hogy az annotálási hibák milyen nagy része számnév; ez valószínűleg a számnevek helyesírásának bonyolultságára vezethető vissza (hiszen a szavakat a kötőjelek mentén feldaraboltuk, így a kötőjelezési hibák is egybe- és különírási hibaként jelennek meg).

Az egyes hibakategóriákra néhány példát láthatunk az 5. táblázatban.

Összességében, tapasztalataink szerint a hibák egy jelentős része arra vezethető vissza, hogy az átiratot és a szótárat (részben) eltérő módon állítottuk össze. A tulajdonnevek fonetikus átírása segített a bemondások fonetikai címkéinek meghatározásakor (és így az akusztikai modell tanításakor), a felismerő szótárába azonban ezek a szavak más alakban kerültek be, így, még ha meg is találta a kérdéses szavakat a beszédfelismerő rendszer, a kimenet az eltérő írásmód miatt hibásnak számított. Valószínűleg a kiejtési szótár és az N-gram modell felépítésére használt korpusz időnként eltérő írásmódja is felelős azért, hogy egy sor rövidítés és tulajdonnév végül kimaradt a nyelvi modellből; ilyenek voltak (az egy szál Fidesz kivételével) a pártok nevei, melyek pedig a híradófelvételeinkben erőteljesen felülreprezentáltak. Emellett valamilyen rejtélyes okból néhány igen gyakori szó (pl. *be*, *legalább*) is hiányzott a szótárból (vagyis a Magyar Kiejtési Szótárból).

A fentieken felül a felismerési hibák meglepően nagy része vezethető vissza egybe- és különírási hibákra, főleg számnevek esetében. Ekkor a felismerő kimenete „gyakorlatilag” helyes, jól olvasható és értelmezhető, „csak” helyesírási hibát tartalmaz. Ez a jelenség nem (vagy csak elhanyagolható mértékben) jelentkezik az angol nyelv esetén; magyar nyelvre végzett felismerésnél azonban ez fokozottan jelen van. Természetesen a rosszul tagolt szavak is hibának számítanak, azonban ezeket logikus lenne kisebb súllyal figyelembe venni, mint ha egy teljesen más jelentésű szót ismertünk volna fel az adott helyen. Véleményünk szerint ez a beszédfelismerés területén gyakorlatilag egyeduralkodó pontosságmetrika (magyar) nyelvspecifikus hiányossága.

5. táblázat. Példák az egyes hibatípusokra

Hibakategória	Etalon szöveg	Felismert szöveg
Egybe- és különírás	kettőszáz milliárdot	kettő százmilliárdot
	százhatvannégyezer	százhatvannégy ezer
	bankszektortól	bank szektortól
	feladatszabó állománygyűlésére	feladat szabó állomány gyűlésére
Kimaradó „a”	leszakította a vihar	leszakította vihar
Írásmódeltérés	smitt pál	schmidt pál
	balog andrást	balogh andrást
Egymás után két azonos msh.	ülést tart	ülés tart
	alkotmánybírók kiválasztásáról	alkotmánybíró kiválasztásáról
Rossz végződés	tihamért	tihamér
	miniszterelnököt	miniszterelnökhöz
	kivégzésére	kivégzését
Egyéb	védőhálóként	védőháló kint
	blogjában	blokkjában
	nem tervez tömeges	nem tervezte meg és
	huszonkilenc pontot	huszonkilenc bontott
	miniszterelnököket	miniszterelnök őket
	húsfeldolgozóba jártak	húsfeldolgozó bejártak
	képest	képes
	tévéjeiket	tévéje éket
Egyéb (szn.)	nulla egész kilenc tized	nyúl egész keresztttüzet
	huszonkilencedikére	ózon kilencedikére
Egyéb (ne.)	alkaida tag	ajkaid adtak
	az atévé híradóban	az a tévéhíradóban
	szuzuki szvift modell	hozó kiszűrt modell
	rogán antal	jogán antal
	kósa lajos	koós alajos
	emeszpés	ám ezt és
	be az emeszpé	bazár messzi
	oszama bin láden	asszam a világon
	biszku béla	whisky béla
	biszku béla	büszke béla
	vargasovszki	varga sóz ki
	európai unió	euró pari unió
	robert ficó	róbert fikció
	fidesz kádéenpé	fidesz káld ilyen ki

5. A kimenet olvashatóságának vizsgálata

Habár a pontos szóalak-írásmód is elvárható egy jól működő beszédfelismerő rendszertől, nyilvánvaló, hogy az (emberi) olvasó is képes valamilyen szintű hi-

6. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák szótévesztéseinek összege

Javított hibák	Hibák száma	Relatív csökk.	Szótév. száma	Relatív csökk.
Semmi (eredeti kimenet)	401	—	603	—
Könnyű hibák	229	43%	400	34%
Könnyű és nehéz hibák	219	45%	368	39%

bajavításra. Ennek vizsgálatához a beszédfelismerő rendszer által szolgáltatott szószintű átiratokat egy tesztalanynak mutattuk meg, és megkértük, hogy próbálja megtalálni és kijavítani a hibákat. A javított hibákat könnyen és nehezen javítható csoportokra osztottuk, ezután két kijavított változatot vizsgáltunk: az egyikben csak a könnyen javítható hibákat korrigáltuk, a másikban pedig mindkét kategóriát. Ezután mindkét változatra kiszámítottuk a pontosságmetrikát.

A 6. táblázat mutatja, hogyan alakult a hibák és a szótévesztések száma a könnyen, illetve a könnyen és nehezen javítható hibák korrigálása után. Látható, hogy a hibák több mint 40%-át korrigálni lehetett olvasás közben, ezek azonban csak a szótévesztések egyharmadáért voltak felelősek. A nehezen korrigálható hibák kijavításával csak 10-zel csökkent a hibák száma, a szótévesztéseké azonban ennél sokkal jobban; erre részben az a magyarázat, hogy bizonyos hibák javítása nem sikerült tökéletesen, azonban szokszor a szétdaraboltan „felismert” szót egyetlen, ám helytelen szóra cserélt a tesztalany (így az illesztési távolságnál egy szócsere és több szóbeszúrás helyett már csak egyetlen szócsere jelent meg).

Amellett, hogy a kísérlet relevanciáját csökkenti, hogy csak egyetlen tesztalannyal végeztük, a fönti számértékeket egyébként is fenntartással kell kezelnünk. Ennek oka részben az, hogy bizonyos hibákat az emberi olvasó sokszor észre sem vesz (pl. bizonyos ragozási, egybe- és különírási hibák), ami mellett természetesen a szöveget tökéletesen megértette.

6. Konklúzió

Ebben a cikkben azt vizsgáltuk, hogy egy hagyományos felépítésű magyar nyelvű beszédfelismerő rendszer milyen jellegű hibákat vét. Ehhez a teszthalmaz egy részén előforduló szószintű hibákat kigyűjtöttük, majd manuálisan kategorizáltuk és elemeztük. Tapasztalataink szerint a hibák egy jelentős része vezethető vissza OOV szavakra, amely megfelel a várakozásoknak (bár a szótárból hiányzó szavak köre szokatlanul tág). Ugyanakkor a tévesztések egy jelentős része egybe- és különírási hiba, amely sokszor arra vezethető vissza, hogy a nyelvi modell mind az egybe-, mind a különírt formát megengedi. A szóalakok pontos egyezésén alapuló pontosságmetrika ráadásul ezt a típusú hibát nem enyhébb, hanem súlyosabb tévesztésnek tekinti, mint hogyha egy teljesen más jelentésű szót ismertünk volna

fel az adott helyen, mely a magyar (és feltehetőleg az egyéb agglutinatív) nyelvű beszédfelismerőket fokozottan sújtja.

Hivatkozások

1. Bánhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével. In: MSZNY, Szeged (2005) 337–347
2. Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T.: Improved recognition of spontaneous Hungarian speech: Morphological and acoustic modeling techniques for a less resourced task. IEEE Transactions on Audio, Speech, and Language Processing **18**(6) (2010) 1588–1600
3. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8) (1966) 707–710
4. Gosztolya, G., Tóth, L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal. In: MSZNY, Szeged (2010) 224–235
5. Grósz, T., Kovács, Gy.., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY, Szeged (2014) 3–13
6. Grósz, T., Gosztolya, G., Tóth, L.: GMM-free ASR using flat start sequence-discriminative DNN training and Kullback-Leibler divergence based state tying. In: ICASSP. (2016) (beküldve)
7. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: ICASSP. (2009) 3761–3764
8. Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel. In: MSZNY, Szeged (2015) 174–181
9. Abari, K., Olaszy, G., Zainkó, Cs., Kiss, G.: Magyar kiejtési szótár az Interneten. In: MSZNY, Szeged (2006) 223–230
10. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)