Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása

Tarján Balázs[1,2], Varga Ádám[2], Tobler Zoltán[2], Szaszák György[1,2], Fegyó Tibor[1,3], Bordás Csaba[4], Mihajlik Péter[1,2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
tarjanb@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.
mihajlik@thinktech.hu
[3] SpeechTex Kft.
tfegyo@speechtex.com
[4] Médiaszolgáltatás-támogató és Vagyonkezelő Alap (MTVA)

Kivonat: Cikkünkben egy valós idejű, kis erőforrás-igényű gépi beszéd-szöveg átalakító rendszert mutatunk be, melyet elsősorban televíziós közéleti társalgási beszéd feliratozására fejlesztettünk ki. Megoldásunkat összevetjük a tématerületen legelterjedtebben használt nyílt forráskódú keretrendszer, a Kaldi dekóderével is. Ezen felül különböző adatbázis-méretek mellett és újrabeszélés alkalmazásával is végzünk felismerési kísérleteket. Kísérleti rendszerünkkel, mely egy több mint 70 millió szót tartalmazó szövegkorpuszon és egy közel 500 órás beszédadatbázison lett tanítva sikerült az eddig publikált legalacsonyabb szóhibaarányt elérnünk magyar nyelvű, televíziós híradók és közéleti társalgási beszéd témakörén.

1. Bevezetés

Világszerte egyre szigorúbb törvények írják elő a televíziós társaságoknak, hogy a kép és hang mellett feliratot is sugározzanak, melynek célja a műsorok akadálymentesítése a siket és nagyothalló nézők számára. Bizonyos műsorok feliratozása kis ráfordítással is megoldható, mert a feliratok rendelkezésre állnak (pl. filmek) vagy elkészíthetőek kézi úton (pl. felvett műsorok). **Élő műsorok** esetén azonban nincs, vagy csak nagyon korlátozottan van lehetőség a hagyományos, manuális módszerek alkalmazására. Cikkünkben bemutatunk egy nagyszótáras beszédfelismerő rendszert, melyet elsősorban közéleti- és hírműsorok gépi úton történő, élő feliratozásához fejlesztettünk a Médiaszolgáltatás-támogató és Vagyonkezelő Alappal (MTVA) folytatott kutatás-fejlesztési együttműködésünk keretében.

Egy ilyen automata feliratozó rendszer fejlesztése többféle kihívást tartogat. A nemzetközi irodalomban a legtöbb eredmény híradók felismerésével született, mely jól artikulált, felolvasott szövegen alapuló beszédnek tekinthető. Ezzel szemben a közéleti, politikai műsorok gyakran két- vagy akár többszereplős párbeszédeket, illetve spontán megfogalmazásokat tartalmaznak, melynek felismerése a **párhuzamos beszédszakaszok** és a **lazább artikuláció** miatt jóval nehezebb feladat. További kihívás, hogy az

élő műsorok feliratozásához valós időben működő rendszert kellett terveznünk, mely ráadásul akár öt közszolgálati csatorna párhuzamos feliratozására is képes. Mindez még egy manapság korszerű szerveren sem egyszerű feladat a nagyszótáras felismerő rendszerek magas erőforrásigénye miatt.

Célunk tehát, hogy bemutassuk a rendszer fejlesztése során kipróbált módszereket és azok hatását a felismerési hibára valamint az erőforrásigényre. Összehasonlítjuk az igen elterjedt, nyílt forráskódú **Kaldi** programcsomag [1], valamint a SpeechTex Kft. által rendelkezésünkre bocsátott **VOXerver** [2] súlyozott véges állapotú átalakítókon (Weighted Finite State Transducer – WFST) alapuló beszédfelismerő dekódereket. Megvizsgáljuk továbbá, hogy mekkora előnnyel járhat, ha a műsorokat nem közvetlenül feliratozzuk, hanem közbeiktatunk egy ún. **újrabeszélőt**, aki elismétli az elhangzottakat. Mindezek mellett különböző méretű akusztikus és szöveges tanítókorpusz mellett is meghatározzuk a rendszer hibáját és erőforrásigényét, valamint **mély neurális hálózatokon** (Deep Neural Network - DNN) alapuló akusztikus modelleket is alkalmazunk a további hibacsökkentés érdekében.

A következő fejezetben cikkünk témaköréhez legjobban illeszkedő nemzetközi és hazai eredményeket mutatjuk be. Ezután a kísérleti feliratozó rendszerünk felépítését, valamint tanító- és tesztadatbázisait, majd a negyedik fejezetben az erőforrás-igényeket és pontosságokat meghatározó méréseink eredményét ismertetjük. Végül az utolsó fejezetben összefoglalását adjuk vizsgálataink legfontosabb eredményeinek.

2. Kapcsolódó eredmények

A legtöbb televíziós műsor felismerésével kapcsolatos publikáció **híradók** leiratozásával foglalkozik. Kutatócsoportunk korábbi eredményei [2–4] 10-50 óra híradós kézi leirat alapján tanított Gaussian mixture modell (GMM) alapú akusztikus modellel és webről gyűjtött szövegeken alapuló nyelvi modellekkel készültek, melyekkel átlagosan 21-27%-os szóhiba-arányt értünk el híradókon. Hasonló mértékű, kb. 24%-os szóhibaarányt említenek [5]-ben, ahol a felügyelet nélküli tanításra helyezték a hangsúlyt. Sajnos azonban ezt az eredményt nem könnyű összehasonlítani másokéval, mivel a tesztanyag egyszerre tartalmazott híradókat és közéleti társalgási beszédet is. Az eddigi legjobb magyar nyelvű híradó-felismerési eredmény legjobb tudomásunk szerint [6]-ban található, ahol 17%-os szóhiba-arányról számoltak be, melyet web-alapú tanítószöveg és DNN akusztikus modell segítségével kaptak.

Közéleti társalgási beszéd közvetlen, tehát újrabeszélés nélküli átírására kevesebb példa van, különösen magyar nyelven. Az egyetlen, melyről tudunk egy korábbi munkánk [2], ahol a híradók felismerésére optimalizált rendszerünket teszteltük televíziós beszélgetéseken is. Az elért 50%-os hibaarány azonban magasnak mondható. Általában a nemzetközi sztenderd ezen a feladattípuson 20-30% között mozog [7–9]. Természetesen az ilyen kis hibájú rendszerek nagy mennyiségű feladatspecifikus hang- és szöveganyagon lettek tanítva, ám szerencsére ilyenek a jelenlegi feladatnál már számunkra is rendelkezésre állnak.

Cikkünkben bemutatott **újrabeszélés**i eredményeket nehéz összehasonlítani a külföldi megoldásokkal. A gyakorlatban is működő újrabeszélt feliratozás általában meg-

lepően nagy, 5-12 másodperces, de eseteként akár 18 másodpercesnél is nagyobb késleltetéssel dolgozik [10]. Ennek egyrészt az az oka, hogy minőségbiztosítási célból az újramondott és felismert feliratok még egy manuális hibajavítási fázison is átesnek, mely nyilvánvalóan késleltetéssel jár. Másrészről a tapasztalatok szerint a feliratok információtartalmát érdemes 125-160 szó/perc körüli értékre csökkenti, hogy ne vonja el a néző figyelmét túlzottan a képről [11]. Ez utóbbi azonban szintén azt jelenti, hogy az újrabeszélőnek be kell várnia bizonyos mennyiségű információt, hogy aztán abból **kivonatot** készíthessen. Ezzel szemben jelenlegi kísérleteinkben **szó szerinti** újrabeszélést kértünk az újrabeszélőktől, és a teljesen **valós idejű** működésre koncentráltunk.

3. A kísérleti rendszer

Ebben a fejezetben a gépi beszéd-szöveg átalakító rendszerünk tanítása és tesztelése során felhasznált adatokat és módszereket ismertetjük.

3.1. Akusztikai modellezés

3.1.1. Akusztikai tanító-adatbázisok

A kísérleteink során alkalmazott akusztikus modelleket két különböző méretű beszéd-adatbázison tanítottuk. Az első adatbázis 64 óra, kézzel annotált, webes híradót tartalmazott (**webes híradók**). Ezt az adatbázist használtuk a kezdeti modellek tanításához és a dekóderek erőforrásigényének felméréséhez.

A második adatbázisba (**kiterjesztett adatbázis**) a kezdeti modell elemei mellé más beszéd-adatbázisokat is felvettünk, melyek reményeink szerint tovább növelik a feliratozó rendszer pontosságát és robosztusságát. A négy kiegészítő adatbázis tartalma és jelölése a következő:

- *Közéleti társalgási beszéd leirata* (**Közéleti hírműsorok**): 31 óra manuálisan címkézett közéleti televíziós beszélgetést tartalmaz, melyet az MTVA bocsátott rendelkezésünkre
- *Félig felügyelten annotált MTVA adatbázis* (**FF**): Az MTVA által rendelkezésünkre bocsátott televíziós felvételek egy részéhez felirat is tartozott. Ezek a feliratok nem mindenhol követték hűen a műsorban elhangzottakat, így közvetlenül nem voltak alkalmasak akusztikus modell tanítására. Ezt a problémát félig felügyelt tanítóanyag-válogatás [12] segítségével kezeltük, így a 136 órányi felirattal rendelkező hanganyagból összesen 100 órát válogattunk ki tanítási célra.
- *Egri Katolikus Rádió adatbázis* (**EKR**): 65 órányi beszélgetést és hírfelolvasást tartalmaz ez az adatbázis, melyet az Egri Katolikus Rádióban rögzítettek.
- *Speecon beszédadatbázis* (**Speecon**): A 2000-ben indult Speecon projekt [13] célja az volt, hogy változatos környezetben rögzített beszédadatbázisokkal segítse a beszédfelismerő rendszerek tanítását. Mi ennek az adatbázisnak a magyar nyelvű változatát használtuk, azon belül is az irodai és otthoni környezetben gyűjtött felvételeket. A négy rögzített mikrofonjel közül kettőt használtunk fel az akusztikus modellünkben, így adódott a 2x114 órás adatbázisméret.

A kísérleteink utolsó fázisában használt rendszer akusztikai tanító-adatbázisa így összesen közel 500 óra hanganyagot tartalmazott (lásd **1. táblázat**).

1. táblázat: Az akusztikai tanító-adatbázisok mérete.

	Kezdeti adatbázis	**+Kiterjesztett adatbázis**			
	Webes híradók	Közéleti hírműsorok	FF	EKR	Speecon	∑
Időtartam [óra]	64	31	100	65	228	488

3.1.2. Akusztikus modellek tanítása

Az akusztikus modellek tanítása Kaldi keretrendszerben [1] történt, de front-endként a VOXerver [2] lényegkiemelő modulját használtuk. Emellett a VOXerver dekóderét is alkalmassá tettük az elkészült akusztikus modellek fogadására.

A 64 órás korpuszon a tanítás a state-of-the-artnak megfelelő módon, MFCC39 jellemzőkön, trifón GMM/HMM modellek elkészítésével indult. A tanított modellek 9453 osztott állapottal, állapotonként átlagosan 10 Gauss-komponenssel rendelkeztek. Az ebből kiindulva készített DNN bemeneti rétege 351 dimenziós (aktuális keret ±4 keret összefűzve), 3 rejtett rétege 400 egységből, egységenként 5 neuronból állt (összesen 2000 neuron rejtett rétegenként), p-norm aktivációs függvényekkel. A p-norm nemlinearitást a maxout nemlinearitás általánosításaként kapjuk [14]:

$$y = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}$$

ahol p szabad paraméter, N pedig a neuronok száma egységenként. Esetünkben p=4, N=5.

Az 500 órás korpuszon tanított modellek jellemzővektorait MFCC13 kiindulási jellemzőkön, a front-endben hét (aktuális ±3 keret) összefűzése után alkalmazott LDA eljárás után 40 dimenziós keretenként kaptuk. A state-of-the-art GMM/HMM 9677 osztott állapotot, állapotonként átlagosan 10 Gauss komponenst tartalmazott. Az ebből kiindulva tanított DNN bementi rétege 9 keret összefűzésével 360 dimenziós, 6 rejtett rétegű, rejtett rétegenként 400x5=2000 neuront tartalmazott, p-norm (p=4) aktivációs függvényekkel.

Az újrabeszélt felvételek kiértékelése során egy a tesztanyaghoz maximum a posteriori (MAP) módszerrel adaptált GMM modellt alkalmaztunk. A lexikai elemek fonetikus átírását a magyar nyelv hasonulási tulajdonságait figyelembe vevő, automatikus eljárással készítettük.

3.2. Nyelvi modellezés

3.2.1. Szöveges tanító-adatbázisok

Kísérleti rendszerünk nyelvi modelljének betanításához négy, különböző forrásból származó szövegkorpuszt használtunk fel (lásd **2. táblázat**):

- *MTVA feliratok*: Az MTVA által rendelkezésünkre bocsátott televíziós felvételekhez tartozó feliratok közül kiválogattuk a közéleti- és hírműsorokhoz tartozókat. Az így nyert 15 millió szót tartalmazó szöveges adatbázis képezi a kezdeti rendszer tanítószövegét, melyet a dekódolási eszközök összevetése során használtunk
- *Webes híradók*: Az azonos nevű akusztikai tanító-adatbázisunk szöveges leirata alkotja ezt a szövegkorpuszt
- *Közéleti hírműsorok*: A webes híradókhoz hasonlóan ez is az azonos nevű beszédadatbázis kézi leiratát tartalmazza
- *Webkorpusz*: A kisebb méretű, de feladatspecifikus tanítószövegek mellett kiegészítő adatbázisként egy webes hírportálokról gyűjtött, 55 millió tokent tartalmazó korpuszt is felhasználtunk a kísérleti rendszerben

2. táblázat: A szöveges tanító-adatbázisok statisztikai adatai.

	Kezdeti adatbázis	**+Kiterjesztett adatbázis**			
	MTVA feliratok	Webes híradók	Közéleti hírműsorok	Webkorpusz	∑
Token [millió szó]	15,1	0,454	0,409	54,8	70,8
Type [ezer szó]	586	67	49	613	931

3.2.2. Nyelvi modellek tanítása

A szövegkorpuszok előkészítése során eltávolítottuk a nem fonetizálható elemeket, meghatároztuk a mondathatárokat, majd statisztikai módszer segítségével átalakítottuk a mondatkezdő szavakat, oly módon, hogy csak a feltételezhető tulajdonnevek őrizzék meg a nagy kezdőbetűs írásmódot. Ezután bizonyos, nem hagyományos lexikai elemeket (pl. számok) átírtunk kiejtett alakjukra, így segítve a kiejtési modell generálását.

A normalizált tanítószövegek alapján minden korpuszon független, 3-gram nyelvi modellt tanítottunk az SRI nyelvi modellező eszköz segítségével [15]. A kísérletek során felhasznált nyelvi modellek ezután úgy készültek, hogy az egyes modelleket lineáris interpoláció segítségével a beszédfelismerési feladathoz adaptáltunk egy paraméterhangolási célokra elkülönített tesztanyagon. Entrópia-alapú modellmetszést nem alkalmaztunk, azonban a webkorpuszból készített nyelvi modell szótárából eltávolítottuk az egyszer előforduló szavakat.

3.3. Tesztelés

A feliratozó rendszer tesztelésére összesen 3,75 óra televíziós közéleti társalgási beszédet különítettünk el, melyből 2,75 órát használtuk a közvetlen feliratozó rendszer tesztelésére és 1 órát az újrabeszélt felvételek kiértékelésére. Ezen felül 10 televíziós híradón is teszteltük a feliratozót, mely összesen további 3 óra tesztanyagot jelentett.

A kísérletek során két súlyozott, véges állapotú átalakítót alkalmazó dekóder teljesítményét vetettük össze. Az első a népszerű Kaldi [1] toolkit **FasterDecoder** nevű eszköze, melyet a SpeechTex Kft. WFST dekóderével a **VOXerver**-rel [2] hasonlítottunk össze. Az eredmények minél pontosabb összevethetősége érdekében minden tesztet ugyanazon a számítógépen (3.5 GHz Core i7), és ugyanazon az operációs rendszeren (Ubuntu 12.04) futtattuk. A különböző implementációk dekódolási sebességét, memóriaigényét és pontosságát is mértük. Az MTVA által rendelkezésükre bocsátott szerveren összesen 24 GB memóriát allokáltak az 5 csatornát feliratozó rendszer üzemeltetésére, így legfeljebb **4.8 GB** állt rendelkezésünkre egy csatorna feliratozásához.

4. Eredmények

A fejezet első felében a Kaldi és a VOXerver dekóder erőforrásigényét hasonlítjuk össze közéleti társalgási beszéd felismerési feladatán. Utána bemutatjuk az összes tanítóanyag felhasználásával készült feliratozó rendszerünk pontosságát és erőforrásigényét immáron nem csak közéleti beszélgetések, hanem híradók esetén is. Végül az újrabeszélők alkalmazásával kapott eredményeket ismertetjük.

4.1. Erőforrásigények összehasonlítása

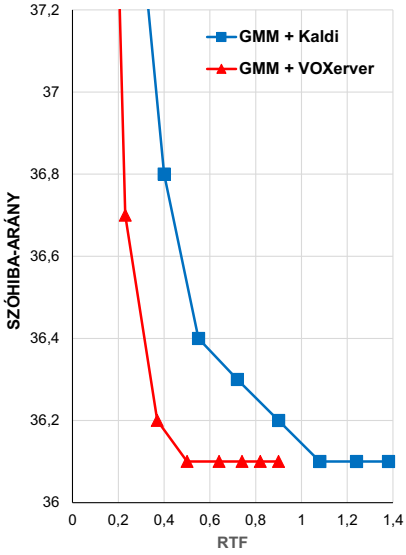
Ennél a vizsgálatnál mind az akusztikus, mind a nyelvi modell tanításához az előző fejezetben bemutatott kezdeti adatbázisokat használtuk, és a közéleti társalgási beszéd közel 3 órás tesztanyagán értékeltük ki őket. Az egyes dekóder és akusztikus modell párosokkal a szaturációs pontban mérhető szóhiba-arányokat és dekódolási sebességeket a **3. táblázat**ban foglaltuk össze. Míg GMM esetén az alkalmazott dekódertől nem függ a hibaarány, VOXerver-rel dekódolva 1%-kal jobb hibaarányt kapunk DNN modell esetén, melynek okát egyelőre vizsgáljuk. A futási sebességek között már azonban lényegesebb különbséget látunk. A Kaldi dekóderével több mint **kétszer annyi időbe telik** a szaturációs pont elérése, mint VOXerver-rel. Talán még ennél is szembetűnőbb a különbség a két dekóder memóriahatékonysága között, ugyanis itt majdnem **hatszoros a különbség,** ismét a VOXerver javára. A Kaldi több mint 7 GB-os memóriaigénye azt jelenti, hogy már egy ilyen szűkített modellel sem tudnánk kiszolgálni az MTVA szerverén az összes csatornát.

A VOXerver kiemelkedő erőforrás-hatékonysága az adattárolási és dekódolási stratégiájában rejlik. A VOXerver-ben az akusztikai állapotok nem részei a WFST-nek, hanem az akusztikus modellek és a CLG (fonetikai környezet, szótár, nyelvtan) szintű WFST kerülnek együtt eltárolásra egy speciális, bináris struktúrában. Ez a struktúra a gyors hozzáférésre, az optimális cache-használatra és a modellek nagyon kompakt reprezentációjára lett kifejlesztve.

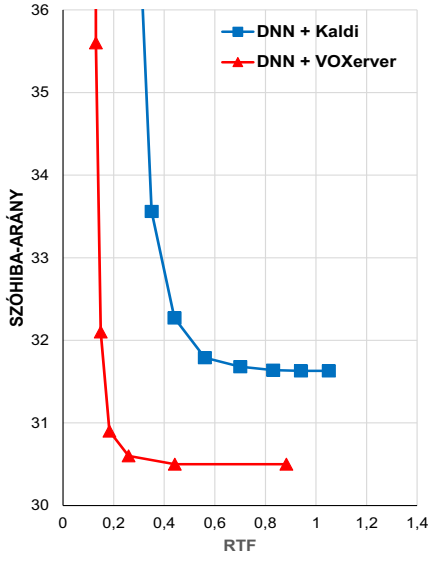
3. táblázat: A közéleti társalgási beszéd közvetlen felismerésével nyerhető legjobb eredmények (**RTF: Real Time Factor**, a dekódoláshoz szükséges idő és a tesztfelvétel hosszának hányadosa, ha RTF ≤ 1 a rendszer képes valós idejű feldolgozásra).

Akusztikus modell	Dekóder	Szóhiba-arány	RTF	Memória használat
GMM	Kaldi	36,1 %	1,23	7,6 GB
	VOXerver	36,1 %	0,5	1,2 GB
DNN	Kaldi	31,6 %	0,9	7,6 GB
	VOXerver	30,5 %	0,44	1,2 GB

A két dekóder futási sebességének alaposabb összevethetőségének érdekében különböző beam szélességgel is futtattunk dekódolást, melynek eredményét az **1.** és **2. ábrán** mutatjuk be. Mint látható mindkét esetben nagyjából fele akkora RTF-nél történik meg a szaturáció a VOXerver-t használva. Ha a lehető legnagyobb pontosság érdekében szaturáció közelében szeretnénk üzemeltetni a feliratozó rendszert, akkor a Kaldi dekóderével a valósidejű működés határán (RTF ~ 1) mozgunk, ezzel szemben a VOXerver képes valósidejű feldolgozásra (RTF ~ 0,4-0,5) az akusztikus modell típusától függetlenül.



1. ábra: Szóhiba-arány a futásidő függvényében közéleti társalgási beszéden mérve, GMM-alapú akusztikus modellel.



2. ábra: Szóhiba-arány a futásidő függvényében közéleti társalgási beszéden mérve, DNN-alapú akusztikus modellel.

4.2. Teljes rendszer

A teljes méretű nyelvi modelleket csak VOXerver környezetben tudtunk futtatni (lásd **4. táblázat**), ugyanis nem állt rendelkezésünkre olyan szerver, mellyel ki tudtuk volna elégíteni a Kaldi keretrendszer WFST hálózat építése közben keletkező memóriaigényét. Becsléseink szerint Kaldi hálózatot használva 25 GB memóriára lett volna szükségünk egyetlen felismerési szál futtatására, és a modell megépítéséhez szükséges memória ennek akár háromszorosa is lehetett volna.

A 4. táblázat adatai alapján elmondhatjuk, hogy a nyelvi modell bővítésével 6%-os, további akusztikus tanítóanyagok bevonásával 8%-os, összegségében pedig mintegy **13%-os relatív hibaarány csökkenést** értünk el a közéleti társalgási beszéd feladatán. Annak érdekében, hogy láthassuk a különbséget a két feladat nehézsége között, egy 3 órás híradókat tartalmazó adatbázissal is teszteltük a modelleket. Látható, hogy még a kezdeti modellekkel is jelentősen alacsonyabb hibaarány érhető el híradókon (~12%), a kiterjesztett modellekkel pedig ez tovább csökkenthető. Az így elért kicsivel **10% alatti szóhiba-arány** annyira alacsonynak mondható, hogy akár újrabeszélő nélküli, közvetlen feliratozását is lehetővé teszi a híradóknak.

Felmerül a kérdés, hogy a nagyobb nyelvi modell és a több rejtett réteget használó DNN akusztikus modell milyen hatással van a feliratozó rendszer erőforrásigényére. A kiterjesztett nyelvi modell hatására a VOXerver memóriaigénye **4 GB-ra növekedett**, mely azonban még így is alatta marad a kitűzött 4,8 GB-os határnak. A dekódolási sebesség tekintetében méréseink alapján nincs változás a kezdeti rendszerhez képest.

4. táblázat: Közéleti társalgási beszéd és híradók feliratozásának hibaaránya a kiterjesztett adatbázisokkal tanított modellek alapján.

Tesztadatbázis	Tanítószöveg	Szóhiba-arány	
		Kezdeti DNN (64 óra)	Kiterjesztett DNN (500 óra)
Közéleti társalgási beszéd	Kezdeti	30,5 %	28,0 %
	+Kiterjesztett	28,7 %	26,4 %
Híradók	Kezdeti	12,4 %	11,1 %
	+Kiterjesztett	10,6 %	9,9 %

4.3. Újrabeszélési kísérletek

Bár a 26%-os hibaarány még nemzetközi összehasonlításban is alacsonynak mondható, ez még nem jelenti azt, hogy a mostani megoldás közvetlenül alkalmas lenne közéleti társalgási beszéd feliratozására. A hibaarány további csökkentése céljából úgy döntöttünk, hogy kipróbáljuk a más országokban már nagy népszerűségnek örvendő **újrabeszélés**t (re-speaking) [11]. Ehhez 1 órányi közéleti társalgási beszéd valós körülményeket szimuláló újrabeszélését rögzítette számunkra az MTVA. Mivel gyakorlott, szakképzett újrabeszélő jelenleg nem érhető el Magyarországon, az MTVA-val közösen úgy döntöttünk, hogy első kísérleteinket tömörítés nélkül, szószerinti újramondással végezzük. Így lehetőségünk nyílt szóhiba-arány számítására, azonban egy gyakorlatban is

működő rendszer esetén a jövőben meg kell fontolni az összefoglaló jellegű újrabeszélés alkalmazását, ugyanis az így létrejött nagy mennyiségű szöveg mind a befogadó, mind az előállítói oldalon problémát jelenthet.

Az újrabeszélt műsorokat két módon teszteltük. Először elkészítettük a feliratot közvetlenül a hangsávból (**közv.**), majd utána az újrabeszélt változatból is (**újrab.**). Akusztikus modellként a közvetlen feliratozás esetén az 4.1 pontban használt GMM és DNN modellt alkalmaztuk, újrabeszélt változaton pedig a DNN-t, illetve a GMM modell egy MAP módszerrel beszélőadaptált változatát. Az eredményeket az **5. táblázat**ban foglaltuk össze. Látható, hogy összesen négy felismerési feladatot és három újrabeszélőt vizsgáltunk, és mindkét tényezőtől erősen függött az újrabeszéléssel kapható javulás. DNN modellek esetén 2-30% között mozog a relatív hibaarány csökkenés, mely tisztán az újrabeszélésnek tulajdonítható. A legjobb eredményt az adaptált GMM modellel kaptuk, mely a közvetlenül feliratozott DNN-es eredményhez képest átlagosan 9%-kal jobb. Elmondható tehát, a mostani egyszerű kísérletek is igazolták az újrabeszélés hatékonyságát. Meggyőződésünk, hogy az újrabeszélők képzésével a jelenleginél sokkal jobb eredmények is elérhetők lesznek a jövőben.

5. táblázat: Közéleti társalgási beszéd közvetlen és újrabeszélés utáni gépi feliratának szóhiba-aránya. A műsor és újrabeszélő azonosítókat az első oszlopban jelöltük arab illetve római számmal.

Műsor / Újrab.	**GMM (közv.)**	**DNN (közv.)**	**DNN (újrab.)**	**GMM + MAP (újrab.)**
1 / I	31,5 %	23,4 %	20,5 %	19,2 %
2 / I	30,7 %	24,8 %	23,4 %	18,8 %
3 / II	33,6 %	25,4 %	18,1 %	17,6 %
4 / III	34,1 %	26,8 %	26,2 %	25,1 %
Átlag	**32,5** %	**25,1** %	**22,1** %	**20,2** %

5. Összefoglalás

Cikkünkben bemutattunk egy elsősorban televíziós közéleti társalgási beszéd feliratozására optimalizált gépi beszéd-szöveg átalakító rendszert, melyet az MTVA-val együttműködésben fejlesztünk. Többféle adatbázis alapján, többféle technikával tanítottunk beszédfelismerő modelleket, melyeket kétféle WFST dekóder segítségével értékeltünk ki. Méréseink azt mutatták, hogy az akusztikai modellezéstől függetlenül a VOXerver **kétszer gyorsabb és hatszor kevesebb memóriát fogyaszt**, mint a Kaldi keretrendszer FasterDecoder nevű eszköze. A több mint 70 millió szón és 500 óra beszéden, mély neurális hálózatok felhasználásával tanított rendszerünk kb. 26%-os szóhiba-aránnyal ismerte fel a közéleti társalgási beszédet és kevesebb mint 10%-os hibával a híradókat. Legjobb tudomásunk szerint ezek a **legalacsonyabb publikált értékek** mindkét magyar nyelvű beszédfelismerési feladaton.

A közéleti társalgási beszéd feliratának javítása céljából újrabeszéléssel is végeztünk kísérleteket. Bár ezzel a technikával sikerült 20% körülire csökkenteni a feliratozás hibaarányát, ez még mindig túl magas a gyakorlati alkalmazhatósághoz. Véleményünk szerint azonban képzettebb, gyakorlottabb újrabeszélőkkel és az összefoglaló jellegű újramondás megengedésével a jövőben jó minőségű feliratok hozhatóak majd létre.

Köszönetnyilvánítás

Ezúton is szeretnénk megköszönni a Médiaszolgáltatás-támogató és Vagyonkezelő Alapnak minden segítséget, mellyel munkánkat támogatta. Kutatásunk részben a Patimedia (PIAC_13-1-2013-0234) projekt támogatásával készült.

Bibliográfia

1. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Others: The Kaldi speech recognition toolkit. In: Proc. ASRU (2011).
2. Tarján, B., Mihajlik, P., Balog, A., Fegyó, T.: Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection. In: 2nd International Conference on Cognitive Infocommunications (CogInfoCom). pp. 1–5. , Budapest, Hungary (2011).
3. Tarján, B., Mihajlik, P.: On morph-based LVCSR improvements. In: Spoken Language Technologies for Under-Resourced Languages (SLTU-2010). pp. 10–16. , Penang, Malaysia (2010).
4. Tarján, B., Fegyó, T., Mihajlik, P.: A Bilingual Study on the Prediction of Morph-based Improvement. In: SLTU 2014: 4th International Workshop on Spoken Languages Technologies for Under-Resourced Languages. pp. 131–138. , Saint Petersburg (2014).
5. Roy, A., Lamel, L., Fraga, T., Gauvain, J., Oparin, I.: Some Issues affecting the Transcription of Hungarian Broadcast Audio. In: 14th Annual Conference of the International Speech Communication Association (Interspeech 2013). pp. 3102–3106 (2013).
6. Tamás, G., György, K., László, T.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY. pp. 3–13 (2014).
7. Sundermeyer, M., Nussbaum-Thom, M., Wiesler, S., Plahl, C., Mousa, A.E.-D., Hahn, S., Nolden, D., Schluter, R., Ney, H.: The RWTH 2010 Quaero ASR evaluation system for English, French, and German. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 2212–2215 (2011).
8. Winebarger, J., Nguyen, B., Gehring, J., Stüker, S., Waibel, A.: The 2013 KIT Quaero Speech-to-Text System for French. In: Proceedings of the 10th International Workshop for Spoken Language Translation (IWSLT 2013). , Heidelberg (2013).
9. Kobayashi, A., Oku, T., Imai, T., Nakagawa, S.: Risk-Based Semi-Supervised Discriminative Language Modeling for Broadcast Transcription. {IEICE} Trans. 95-D, 2674–2681 (2012).
10. Ofcom: Measuring live subtitling quality: Results from the first sampling excercise. (2014).
11. Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., Remael, A.: Live Subtitling with Speech Recognition Causes and Consequences of Text Reduction. (2010).
12. Mihajlik, P., Balog, A.: Lightly supervised acoustic model training for imprecisely and asynchronously transcribed speech. In: Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on. pp. 1–5 (2013).

13. Siemund, R., Höge, H., Kunzmann, S., Marasek, K.: SPEECON-speech data for consumer devices. Second Int. Conf. Lang. Resour. Eval. 883–886 (2000).
14. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 215–219. IEEE (2014).
15. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proceedings International Conference on Spoken Language Processing. pp. 901–904. , Denver, US (2002).