e-Magyar beszédarchívum

Kornai András[1], Szekrényes István[2]

[1] MTA Nyelvtudományi Intézet,
1068 Budapest, Benczur u. 33., e-mail: andras@kornai.com
[2] Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
4032 Debrecen, Egyetem tér 1, e-mail: szerkenyes.istvan@arts.unideb.hu

Kivonat Cikkünkben az e-magyar digitális nyelvfeldolgozó rendszer részeként létrejött nyílt forráskódú és szabad felhasználású beszédarchívum jelenlegi állapotáról és további terveiről számolunk be.

Kulcsszavak: beszédtechnológia, beszédarchívum, e-magyar

1. Céljaink

A beszédarchívum[3] létrehozásával három fő célunk volt. Az első és legfontosabb a magyar beszédtechnológiára annak kezdetei óta jellemző zárt kutatási és publikációs modell felváltása egy szabad, nyílt forrású (Free and Open Source Software, FOSS) modellel. Második célunk a hagyományos, gondosan felcímkézett és mind artikulációsan mind akusztikailag tiszta adatokon alapuló felügyelt tanulási módszerek felváltása gyengén felügyelt illetve felügyeletlen (weakly supervised, unsupervised) módszerekkel. Harmadik, az első kettőtől nem mindig könnyen elválasztható célunk pedig egy a digitális bölcsészeti munkát, elsősorban a szociológiát, történelemtudományt, folklorisztikát, és néprajzot beszédtechnológiai oldalról támogató platform alapjainak megteremtése.

2. Kiinduló állapot

Az `e-magyar` pályázat a nyelvtechnológiában, különösen a szószintű eszközök (morfológiai elemzés és generálás), de kisebb részben már a frázis- és mondatszintű eszközök területén teljessé tette a nyílt forrású adatok és eszközök bevezetését (Várad et al, ugyane kötetben), ennek minden, a fejlődést katalizáló előnyével együtt. Ez csak úgy volt lehetséges, hogy az évtizedek során komoly FOSS eszközök halmozódtak fel, melyek közül a teljesség igénye nélkül kiemeljük a `Hun*` és a `Magyarlánc` eszközláncokat, a monolinguális `Webkorpuszt` és a `Hunglish` párhuzamus korpuszt. Mostani projektünk elkezdése előtt a magyar beszédtechnológia szabadon letölthető adatokat nem tett közzé (az egyedi mérlegelésen alapuló hozzáférés-engedélyezést nem sorolhatjuk a FOSS paradigmába) sem a világszerte közismert beszédtechnológiai eszközök magyar honosításai nem voltak elérhetőek, annak ellenére, hogy a létező szoftverek, különösen a beszédfelismerés terén, elsősorban ilyeneken alapultak (ennek pontos mértéke természetesen csak a szoftverek nyilvánosságra kerülésével lesz megállapítható).

[3] `http://e-magyar.hu/hu`

3. A projekt eredményei

Elmondhatjuk, hogy a FOSS beszédarchívum megjelenésével a helyzet gyökeresen megváltozott. Az adatok szintjén elérhetővé vált sok ezer órányi jogtiszta adásminőségű (broadcast quality) és sokszáz órányi ennél rosszabb (communication quality) anyag. Ezeknél sokkal jobb minőséget képvisel a `BEA` spontánbeszéd-adatbázis [2], de kisebb, és nem teljesen FOSS. Hangsúlyoznánk, hogy a korszerű beszédtfelismerésben a jobb akusztikai minőség nem követelmény, sőt, immár több évtizedes tapasztalat, hogy a legjobban azok a beszédfelismerő rendszerek teljesítenek, melyeket reális, az alkalmazásban valóban fellépő akusztikai körülményeket tükröző adatokon tanítottak be.

Ugyanilyen változást hozott a projekt a követő szoftverek terén is. Több tucatnyi alternatíva telepítésével és összemérésével választottuk ki a legjobbakat. Számos okból utasítottunk el szoftvereket:

- Egzotikus nyelvet igényel (pl. Luá-t mint a corona[4])
- Elöregedett modulokat használ (pl. tcl/tk-t mint a snack[5])
- Zárt modulokat használ (pl. a pysonic[6])
- Rendszerspecifikus (leggyakrabban Windows)
- Dokumentálatlan (pl. a RawAudioSocket)
- Csak kutatásra használható (pl. az OpenSmile[7])
- Elhagyott (pl. a LiUM[8])
- Fontos formátumokat nem támogat (pl. az AudioLazy)

Tucatjával találtunk olyan szoftvereket, melyek egyszerre több szempontból is problematikusak, és van még egy pár olyan, amivel változatlanul próbálkozunk, ilyen pl. a `bob.bio.spear`[9] és a `Brno phoneme recognizer`[10].

A hangformátumok konverziójára végül a `SoX` és `ffmpeg` eszközöket, a beszédaktivitás detektálására és naplózás (diarization) céljára a `shout` programot (ld. 4.1), végül statisztikai nyelvmodellezésre az `srilm` eszközt (Nemeskey, ugyane kötetben) használtuk fel. Ez utóbbihoz olyan modelleket tettünk elérhetővé, melyek perplexitása 56, tudtunkkal az összes publikált (de le azért nem tölthető) modell perplexitását lényegesen megjavítva. Eredeti vállalásunkkal ellentétben *nem készült el*, de terveink között változatlanul szerepel az automatikus nyelvazonosítást lehetővé tevő szoftver.

[4] `https://docs.coronalabs.com/api/library/audio/play.html`
[5] `http://www.speech.kth.se/snack`
[6] `http://pysonic.sourceforge.net`
[7] `http://audeering.com/research/opensmile`
[8] `http://www-lium.univ-lemans.fr/diarization/doku.php/download`
[9] `https://pypi.python.org/pypi/bob.bio.spear/2.0.4`
[10] `http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context`

4. Új modulok integrációja

4.1. emDia, emSad

Az `emDia` beszélő diarizáló modul a 'ki, mikor beszélt' kérdésre ad választ (tehát a beszélőváltásokat állapítja meg), ez a nyílt forráskódú (GPL), C++-ban írt `SHOUT` Speech recognition toolkit [4] 'shout_segment' és 'shout_cluster' programjainak a használatával történik. A modul a bemeneti audio fájlt a `SoX` (Sound Exchange, GPL) eszközt használva konvertálja, így minden olyan formátumot elfogad, amit ez kezel (pl. mp3, wav). A diarizáló modul kimenete két RTTM (Rich Transcription Time Marked) kompatibilis fájl, amelyek a megtalált beszéd-zaj-csend, illetve a különböző beszélőkhöz tartalmazó audio szegmenseket írják le.

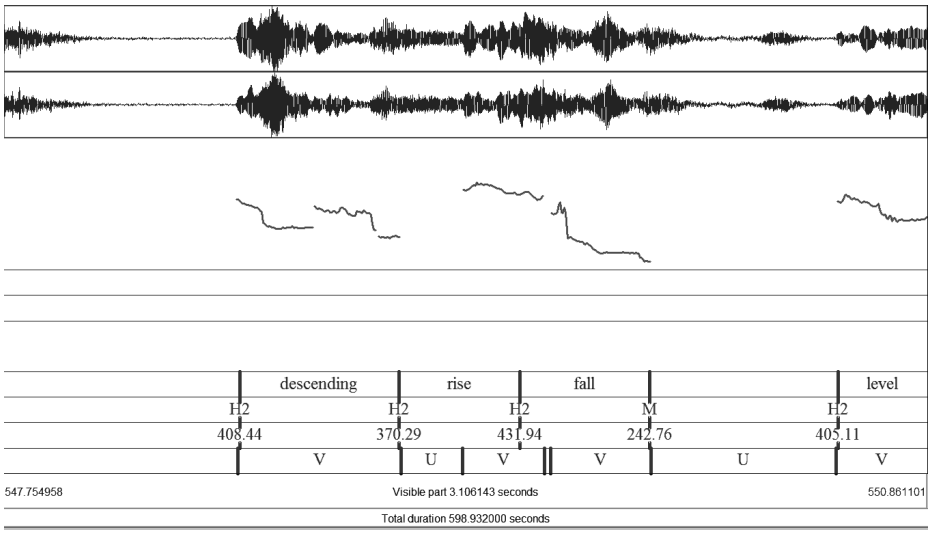
Az `emSad` modul a diarizáló modul első lépésének, a beszédtevékenység detekciónak az önálló futtatását teszi lehetővé. Szintén a `SoX` eszköz felhasználásával többféle bemeneti formátumot támogat. A modul funkciói közé tartozik még az azonos típusú szegmensek egyetlen hangfájllá konvertálása, ami pl. alkalmas egy beszédet, zajt és csendet vegyesen tartalmazó fájlból a beszéd kinyerésére.

4.2. emPros

Az `emPros` (eredeti nevén: `ProsoTool`) egy a `Praat` beszédfeldolgozó program [1] szkript nyelvén implementált, az élőnyelvi kommunikációban előforduló verbális megnyilatkozások prozódiájának elemzésére és lejegyzésére szolgáló algoritmus, amely a `HuComTech` projekt alapkutási céljai [5] érdekében került (gépi annotálást végző, offline eszközként) kifejlesztésre. A fejlesztés kezdeti szakaszának – még csak a terveket és a lehetőségeket feltáró – részeredményei a VIII. Magyar Számítógépes Nyelvészeti Konferencián kaptak először nyilvánosságot [10]. A későbbi publikációk elsősorban a beszéddallam automatikus lejegyzésére szolgáló, az `e-magyar`[11] projekt weboldalán is elérhető modul hátterét [9] és működését [8] tárgyalják. A további tervek között szereplő, a beszéd hangerőváltozásait és tempóját elemző modulok jelenleg is fejlesztés alatt állnak. Az algoritmus tesztelése a `Langua Archive`[12] és a `Meta-Share`[13] projekteken keresztül kutatási célokra közzétett `HuComTech korpusz`[14] magyar nyelvű, formális és informális dialógusokat rögzítő hangfelvételeinek és szöveges átiratainak felhasználásával, a korpusz széleskörű elemzési szempontokat átfogó annotációnak további bővítése céljából történt. A legfrissebb (eddig nem publikált) javítások és átdolgozások, melyeknek a program jelenlegi flexibilitását köszönheti, a `SegCor` projekt[15] közreműködésével, a `FOLK` korpusz [7] német nyelvű, változatos kondíciók között (2–14 adatközlővel) készített hangfelvételeinek elemzése során valósultak meg.

[11] `http://e-magyar.hu/hu/speechmodules/emPros`
[12] `https://tla.mpi.nl/`
[13] `http://metashare.nytud.hu/`
[14] `https://hdl.handle.net/1839/00-0000-0000-001A-E17C-1@view`
[15] `http://www1.ids-mannheim.de/prag/muendlichekorpora/segcor.html`

Az alkalmazás fejlesztését leginkább Piet Mertens szintén a `Praat` program szkript nyelvén, `Prosogram`[16] néven implementált, a tonális kontúrok pszichoakusztikai alapokon [3] történő stilizálását végző eljárása inspirálta [6], de az alaphang modulációinak kategorizálására használt módszereket tekintve a `Tilt`[17] intonációs modell paramétereiből is merített. Fontos különbség, hogy az intonáció elemzését az `emPros` a beszéd szegmentális szerkezetétől függetlenül, nem a szótagok szintjén végzi, így nem is igényli a szótaghatárok előzetes detektációját. A szegmentáció alapját az alapfrekvencia kontúr (a `Praat` program beépített funkcióival történő) simítása és stilizálása eredményeként kapott, a percepció számára nem releváns mikro-intonációs mozgásokat a beszéd hosszabb egységein átívelő intonációs trendekben integráló dallammenetek képezik. A dallammenetek kategorizálása és címkézése azok időtartama és Hertzben mérhető „amplitúdója" alapján történik, amely a vizsgált beszélő öt részre felosztott hangterjedelmével és átlagos hangmagasság ingadozásával kerül összevetésre.



1. ábra. A ProsoTool kimenete a Praat program szerkesztő felületén

Mivel a szkript beszélőnként végzi az intonáció elemzését, a hangfelvétel mellett egy olyan (Praat TextGrid formátumú) annotáció is bemeneti követelmény, amely a megnyilatkozások időbeli pozícióját beszélőnként külön tengelyen (annotációs szinten) tartalmazva reprezentálja a fordulóváltások akusztikai szerkezetét. Az `e-magyar` beszédfeldolgozó moduljai között helyett kapó `emDia` pontosan a föntebbi információkat szolgáltatja kimenetként, így az `emPros` a beszélő diarizáló kimenetén alkalmazott eljárásként integrálható, amelyben a beszélők hangjának izolált akusztikai elemzését egy a prozódiai moduloktól különválasz-

[16] `http://bach.arts.kuleuven.be/pmertens/prosogram/`
[17] `http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/c16909.htm`

tott előfeldolgozó algoritmus készíti elő. A kimenet a bemenetben jelölt beszélők szerint elkülönítve, `Praat TextGrid` formátumban kódolja a hanglejtés dallammenetekre szegmentált elemzését. A lejegyzés négy, a(z) 1. ábrán is látható, időben párhuzamos szintből áll. Az első szint a stilizálás eredményeként kapott dallammeneteket a „rise" (szökő), „fall" (lebegő), „ascending" (emelkedő), „descending" (ereszkedő), „level" (szinttartó) kategóriák valamelyikébe sorolja. A második szint a dallammenetek mozgását a beszélő 5 szintre ($L_2 < L_1 < M < H_1 < H_2$) felosztott hangterjedelmében pozícionálja. A harmadik szint az előző szint relatív értékeihez az eredeti, Hertzben mért értékeket társítja hozzá. A negyedik szint pedig a beszéd zöngés („V") és zöngétlen („U") szakaszait különíti el.

5. Együttműködésben várható eredmények

Az archívum bővülése több irányból is várható anélkül, hogy ez újabb anyagi vagy emberi erőforrásokat igényelne. Támogatásukról biztosítottak a NAVA, az OGyK, az OSzK, az MTA TK, Kisebbségkutató, és Szociológiai intézetek és más intézmények is, sőt az intézmények egy részétől már kaptunk is anyagokat.

Különösen fontos a hazai és környező országokbeli társadalomtudósok támogatása. A teljesség igénye nélkül: Havas Gábor, Lengyel Gabriella, Németh Szilvia, Zolnay János, Virág Tünde; a kolozsvári kisebbségkutató (Fosztó László, Kiss Tamás, Vitos Katalin, Lőrincz József), a marosvásárhelyi Sapientia (Gagyi József), a kolozsvári Kriza Társaság (Szabó Töhötöm), a Babes-Bolyai Egyetem (Tánczos Vilmos, Pozsony Ferenc), a kolozsvári, marosvásárhelyi rádiók anyagai (Maksay Ágnes, Tibád Zoltán).

Külön említést igényel Molnár Gusztáv hatalmas interjúanyaga (mintegy (70 óra, nagyrészt magyarul, de több mint 20 óra románul) a XX század olyan jelentős személyeivel mint Balogh Edgár vagy Szabó T. Attila. Sajnos ezen anyagok nagy része ma még kazettán van, de ezek átjátszását folyamatosan végezzük.

Különösebb plusz befektetés nélkül, csupán a meglevő folyamatok folytatásával az archívum még éveken át bővülni fog.

6. A továbblépés főbb irányai

Számítunk a közösség támogatására abban, hogy a beszédarchívum még jobban használható legyen. Az első és legfontosabb lépés ebben egy *adatkezelési* modell (data curation model) kialakítása kell legyen.

Kik adják az adatokat? A google kérdőív[18] kitöltésével bárki, aki szeretné adatait nyilvánosan hozzáférhetővé tenni.

[18] `https://docs.google.com/forms/d/e/1FAIpQLSdwBoeLh_g2A6FO5VbKONGIBYJ-CfWb83KXFClVodr68Bhm5w/viewform?c=0&w=1`

Kik őrzik az adatokat? Ennek infrastrukturális hátterét legalább 10 évre megadta az `e-magyar` finanszírozású hardver-fejlesztés, a szervezeti hátteret biztosítja az MTA Nyelvtudományi Intézet és az MTA SZTAKI közti megállapodás. Természetesen teljes idejű, vagy akár részidejű digitális könyvtáros felvétele a folyamatot nagyban gyorsítaná, erre azonban a pályázat egyszeri jellege nem adott módot.

Milyen metaadatokat tároljunk, és milyen sémában? A rendszer rugalmas, itt elsősorban az érdekelt felhasználók véleményét várjuk ahhoz, hogy igényeiknek a leginkább megfelelő adatbázis-sémát és keresési eszközöket illesszünk az adatokhoz. Terveink szerint ez nem kézi címkézéssel nyert „gold", hanem az `emSad`, az `emDia`[19], és a `emPros` az egész adaton való átfuttatásával keletkező „silver" adatokon fog alapulni.

A második kérdés a további szoftverek fejlesztése. Mint az `emPros` (ProsoTool)[20] példája mutatja, független `github` szoftver-repozitórium minden nehézség nélkül kapcsolható az e-magyar-hoz, és nagy örömmel várjuk a többi FOSS szoftver megjelenését.

7. Köszönetnyilvánítás

Köszönettel tartozunk Uwe Reichelnek és Mády Katalinnak (NYTI), továbbá a `speech@lists.mokk.bme.hu` levelezőlista minden tagjának számos hasznos ötletért és tanácsért, Pajkossy Katalinnak és Ács Juditnak (BME) az `emDia` és az `emSad` beüzemeléséért, Takács Dávidnak (Meltwater) és Gerőcs Mátyásnak (NYTI) a webes arculatért. Külön köszönet Schreiner Józsefnek (interNetWire Communications) a határon túli kutatások anyagának áttekintéséért és a digitalizáció beindításáért, és Both Zsoltnak (SZTAKI) a hardver beüzemeléséért.

Az `e-magyar` eszközlánc az MTA 2015. évi Infrastruktúra-fejlesztési Pályázat 2. kategóriájában elnyert támogatás segítségével valósult meg.

Hivatkozások

1. Boersma, Paul & Weenink, D.: Praat: doing phonetics by computer [computer program]. version 6.0.22. `http://www.praat.org/` (2016), retrieved 15 November 2016
2. Gósy, M. (ed.): Beszéd, adatbázis, kutatások. Akadémia (2012)
3. Hart, J.t.: Psychoacoustic backgrounds of pitch contour stylisation. IPO-APR 11, 11–19 (1976)
4. Huijbregts, M.: Segmentation, diarization and speech transcription: surprise data unraveled. Ph.D. thesis (2008)

[19] `https://github.com/hlt-bme-hu/hunspeech`
[20] `https://github.com/szekrenyesi/prosotool`

5. Hunyadi, L., Földesi, A., Szekrényes, I., Staudt, A., Kiss, H., Abuczki, A., Bódog, A.: Az ember–gép kommunikáció elméleti–technológiai modellje és nyelvtechnológiai vonatkozásai. In: Általános Nyelvészeti Tanulmányok XXIV: Nyelvtechnológiai kutatások, pp. 265–309. Akadémiai Kiadó, Budapest (2012)
6. Mertens, P.: The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In: Proceedings of Speech Prosody (2004)
7. Schmidt, T.: Good practices in the compilation of folk, the research and teaching corpus of spoken german. In: Kirk, J.M., Andersen, G. (eds.) Compilation, transcription, markup and annotation of spoken corpora, Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], pp. 396–418 (2016)
8. Szekrenyes, I.: Prosotool, a method for automatic annotation of fundamental frequency. In: 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 291–296. IEEE, New York (2015)
9. Szekrényes, I.: Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. Journal on Multimodal User Interfaces 8:(2), 143–150 (2014)
10. Szekrényes, I., Csipkés, L., Oravecz, C.: A hucomtech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei, automatikus prozódiai annotáció. In: Tanács, A., Vincze, V. (eds.) VIII. Magyar Számítógépes Nyelvészeti Konferencia, pp. 190–198. JATEPress (2011)