FORvoice 120+: magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra

Beke András, Szaszák György, Sztahó Dávid

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
1117 Budapest, Magyar tudósok körútja 2.
beke.andras@gmail.com, {szaszak, sztaho}@tmit.bme.hu

Kivonat: A jelen tanulmányban elsőként kerül bemutatása a FORvoice 120+ magyar nyelvű kriminalisztikai célú utánkövetéses adatbázis. A FORvoice célkitűzése egy kriminalisztikai szempontból megbízható, követéses, reprezentatív beszélői adatbázis elkészítése magyar nyelven. Az adatbázis vizsgálati anyagot biztosít a magyar nyelven történő kriminalisztikai fonetikai kutatásokhoz, illetve a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez. Az adatbázis 120 beszélő (60 női és 60 férfi) felvételét fogja tartalmazni. A felvételek szigorú protokoll szerint történnek, amelyek követik a nemzetközi irányokat. A FORvoice lehetőséget biztosít, hogy azon akusztikai, fonetikai, nyelvészeti, beszédtechnológiai kutatásokat végezhessenek, külön tekintettel az adatközlő egyéni beszéd tulajdonságára, továbbá a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez, új, egyéni akusztikai-fonetikai jellemzők megállapításához.

1 Bevezetés

Az elmúlt évtizedekben egyre növekedett az a szakmai erőfeszítés, amely azt a lehetőséget vizsgálja, hogy vajon a beszéd milyen egyéni, személyhez köthető tulajdonságai hordozzák a beszélőspecifikus jellemzőket (vö: Tomi Kinnunen 2018: An Overview of Text-Independent Speaker Recognition: from Features to Supervectors). Mindezen kérdések még nagyobb figyelmet kapnak a törvényszéki munka során, ahol a cél egy kérdéses mintán hallható személy kilétének megbízható azonosítása objektív, statisztikai, megismételhető módszerek segítségével, ahogy azok a DNS-tesztek módszertanában ismeretesek. A beszélők személyének felismeréséhez szükséges egy olyan magyar nyelvű, sok beszélőt tartalmazó adatbázis, amely kriminalisztikai céloknak megfelel, és amely lehetőséget biztosít kriminalisztikai fonetikai, nyelvészeti és beszédtechnológiai kutatások elvégzéséhez. A jelen tanulmány egy ilyen adatbázis fejlesztését mutatja be, amely legalább 120 beszélőt tartalmaz, szigorú protokoll mentén rögzített, az egyes beszélőktől időben eltérő hangmintákat tartalmaz, illetve különböző beszédtípusokat. Az adatbázis jelentősége igen nagy, mivel lehetőséget biztosít a beszélők beszédének személyspecifikus jellemzőinek vizsgálatára. Ugyanakkor az adatbázis és az azon elvégzett kutatássorozat nem csak a kutatók számára hasznos, hanem a társadalom szá-

mára is, mivel egy ilyen adatbázis lehetőséget biztosít a rendőri szerveknek, nemzetbiztonsági szerveknek, hogy a törvényszéki munka során használt rendszereket megbízhatóbbá tegyék, újakat fejlesszenek.

A kriminalisztikai hang-összehasonlítás során arra keressük a választ, hogy az időben korábban a hivatalos szervekhez érkezett hangminta (jellemzően telefonos spontán beszédet tartalmazó minta) ugyanattól a személytől származik-e, akitől később az eljárás során interjú szituációban vettek hangmintát (jellemzően nem spontán stúdió hangminőségű hangminta). A kriminalisztikai hang-összehasonlításkor a beszédmintákat akusztikailag elemzik, és ezen az elemzésen alapulva mutatják be, hogy a hasonlóságot milyen mértékben növeli vagy csökkenti a keletkezett bizonyíték ezzel segítve a bírói döntési mechanizmust.

A kriminalisztikai tudományban bekövetkezett paradigmaváltás (Saks és Koehler, 2005) előtt az elemzés során elégséges volt csak a két hangminta akusztikai összevetése, vagyis annak prezentálása, hogy az adott akusztikai jegy (pl. alaphangmagasság) nagy hasonlóságot vagy különbséget mutat-e. Ugyanakkor belátható, hogy fennállhat az az eset is, hogy egy populációból véletlenszerűen vett két beszélőnél ugyanezt a hasonlóságot vagy különbséget találnánk. A kérdés tehát az, hogy az adott akusztikai jegy(ek) mennyire hasonló(ak), illetve mennyire tipikus(ak) az adott egyénre, illetve a populációra nézve. Ezt a kérdés oldotta fel a kriminalisztikai tudományban bekövetkezett paradigmaváltás (Saks és Koehler, 2005), amely a bizonyíték kiértékelésében, illetve prezentálásában hozott változásokat, és amely forradalmasította a kérdéses és a gyanúsítottól származó minta összehasonlításának módszertanát (vö. DNS-profil stb.). Az új paradigmát a valószínűségi-arány keretrendszer (likelihood-ratio framework) kvantitatív implementációjaként lehet jellemezni, amely az eredmények megbízhatóságának kvantitatív úton történő kiértékelését biztosítja. A likelihood-ratio framework során két alapvető hipotézis kell megvizsgálni. A jogalkalmazó által az igazságügyi szakértőnek feltett alapkérdés: „Mekkora valószínűséggel származik a kérdéses minta a gyanúsított személytől?”, illetve az ún. ellenhipotézis: „Mekkora valószínűséggel származik a kérdéses minta az adott népességből véletlenszerűen kiválasztott másik személytől?”. Mindkét, ún. posterior valószínűség kiszámításához a Bayes-elv alapján hipotézisenként két valószínűségi értéket kell kiszámolni, majd a kapott valószínűségeket egymással elosztani. Az igazságügyi kérdés a likelihood framework tükrében tehát az, hogy „Mennyiszer tűnik valószerűbbnek, hogy a megfigyelt különbségek az ismert és a kérdéses minták között azt a feltételezést támogatja, hogy a kérdéses mintának és az ismert mintának azonos az eredete, mint azt a feltételezést, hogy az eredete különböző?” (lásd bővebben Geoffrey Stewart Morrison munkáit).

Ahhoz, hogy a LR keretein belüli kísérleteket elvégezhessük, szükséges egy olyan adatbázis, amely az új paradigma alapfeltevéseinek megfelel (Morrison és mtsai, 2012): 1):

(i) minden beszélőtől időben eltérő mintákat kell rögzíteni (hasonlóság modellezése),
(ii) sok beszélőt kell tartalmaznia lehetőleg a populációra reprezentatíven (tipikusság modellezéséhez),
(iii) különböző módon rögzített hangmintákat kell felvenni (un. channel mismatch kompenzálására, pl. telefonos vagy stúdió minőségű),

(iv) egy beszélőtől különböző beszédtípusokat kell rögzíteni a beszédstílus különbségeiből fakadó beszélőn belül is megjelenő eltérések kompenzálására (speech style mismatch compensation).

Mindezen kihívásoknak megfelelően terveztük meg a jelen tanulmányban bemutatott FORvoice adatbázist, amely a jelenleg elérhető hazai adatbázisok között egyedülálló (vö. MTBA (Vicsi és mtsai, 2004), MRBA (Vicsi és mtsai, 2004), BABEL (Vicsi és Vig, 1998), BEA (Gósy és mtsai, 2012)).

2 Célkitűtések

A fejlesztett FORvoice adatbázis a következők célkitűzések mentén épül fel. Célunk egy olyan adatbázis létrehozása, amely

(i) illeszkedik a kriminalisztikában bekövetkezett új paradigmaváltásban megfogalmazott kritériumokhoz, így a rajta végzett elemzések fontos alapkövei lehetnek a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez, új, egyéni akusztikai-fonetikai jellemzők megállapításához,

(ii) annotált és lekérdezhető, így lehetőséget biztosít a szakemberek számára, hogy azon akusztikai, fonetikai, nyelvészeti, beszédtechnológiai kutatásokat végezhessenek, külön tekintettel az adatközlő egyéni beszéd tulajdonságaira, továbbá

(iii) olyan alap adatbázis legyen, amelyen új kutatási irányokat lehessen megvalósítani a fonetikában, a beszédtechnológiában és olyan kutatási kérdések megválaszolására adjon alapot, amelyeket korábban nem lehetett tanulmányozni magyar nyelven (pl. a beszélőn belüli és a beszélők közötti variancia szisztematikus elemzése hosszabb időtávon, stb.).

3 Anyag, módszer és kísérleti személyek

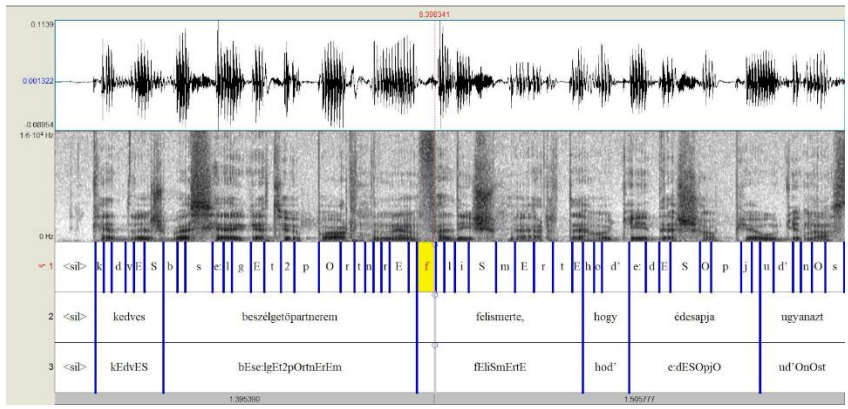
Az adatbázis készítése során a fő szempontok igazodnak a nemzetközi irodalomhoz (Morrison és mtsai, 2012): 1) minden beszélőtől legalább két, időben relatíve távoli felvételt kell tartalmaznia; 2) az egyes személyektől különböző beszédtípusokat kell rögzíteni: alkalmi beszélgetés, irányított beszélgetés, ál-rendőrségi-kihallgatás (monológ formájában); 3) az adatbázisnak meg kell felelnie a kutatások és a kriminalisztikai esetek követelményeinek (a felvételi és adatátviteli csatorna közötti különbségek modellezése). A felvételi és adatátviteli csatorna eltérésének kritériuma utólag kerül modellezésre digitális jelfeldolgozás segítségével.

A FORvoice 120 beszélő hangmintáit fogja tartalmazni (60 női és 60 férfi). A beszélők 18-65 év közötti magyar anyanyelvű személyek, változatos születési hellyel. Minden személytől kétszer készül felvétel, a két alkalom között legalább két hét különbséggel. A felvételek előtt az alanyok írásban nyilatkoznak a hangjuk rögzítéséhez való hozzájárulásukról. A hangfelvételek mellett rögzítjük a beszélők fittségi és stressz szintjét (önbevallásos alapon), illetve, hogy dohányoznak-e, valamint azt, hogy van-e hangképzést befolyásoló betegségük.

A felvételek párosával történnek egy csendes irodai szobában. A beszélő párok 2-3 m távolságra ülnek egymástól (1. ábra). A hangrögzítés USB-s külső audió interfész és fejmikrofonok segítségével történik, a kezdeti jó minőség biztosítása érdekében (2. ábra). A felvétel során a beszélőknek négy feladatot kell elvégezniük:

(i) *Szabad párbeszéd*. Teljesen szabad párbeszéd, kötöttségek nélkül.
(ii) *Célzott információcsere*. A beszélőpárosok egy-egy hibás terméklista számukra 4-4 olvashatatlan soraiban lévő információit kell megszereznie a beszélgetőtárstól.
(iii) *Monológ*. A beszélőknek az előző napjukat kell elmesélniük tárgyilagosan.
(iv) *Állandó válaszú kérdések*. A beszélőknek 5-5 rögzített kérdésre kell 'Nem emlékszem', illetve 'Nem kívánok válaszolni' választ adniuk.

1. ábra. A hangfelvétel helyszíne



2. ábra. Spektrogram és szegmentálás minta egy elkészült jó minőségű felvételről

A FORvoice jelenleg 60 beszélő teljes protokolját tartalmazza. Elkészült a hanganyag szó- és hangszintű átiratozása is. Mindemellett további kiegészítésként az intonációs frázisok jelölése is megvalósult.

4 A FORvoice-on tervezett kutatások

A korai kutatások során elsőként az alaprendszereket (baseline) szeretnénk rögzíteni és publikálni, vagyis a jelenleg széles körben használt i-vector alapú beszélőfelismerő algoritmusok kiértékelése történik meg a FORvoice-on.

Emellett leíró jellegű temporális akusztikai-fonetikai elemzéseket is tervezünk az egyéni beszédjellemzők vizsgálatára. Későbbiekben vizsgáljuk a kriminalisztikai szempontú beszélőspecifikus akusztikai-fonetikai paramétereket. Elemezzük, hogy az új beszélőjellemzők az általunk létrehozott rendszer eredményeit milyen módon javítják. Alapvetően három nagyobb területen tervezünk kísérleteket végezni: i) temporális jellemzők, ii) prozódiai jellemzők, iii) mély neurális hálózatokon alapuló jellemzők. A kutatás során olyan akusztikai paramétereket vizsgálunk, amelyek dinamikus változása jól tükrözi az artikulációs szervek egyéni működését, ilyen módon beszélőspecifikus jellemzőként működnek, valamint jól vizsgálhatók az agglutináló típusú nyelv esetében. A temporális vizsgálatok során az egyes hangfelvételekből különböző típusú időzítési információk kinyerését végezzük el, amelyek alapján elemezzük a beszélők közötti és az egyes beszélőkre jellemző variancia mértékét. Kísérletet tervezünk a mély neurális hálók alkalmazására a törvényszéki hang-összehasonlító rendszer jellemzőkinyerésére. A mély neurális hálók segítségével lehetőség nyílik további lokális jellemzőkinyerésre is. Egy ilyen technika a mély hálók utolsó rejtett réteg kimeneteinek használata (Bacchiani és Rybach, 2014). Több tesztben magukat a hálók kimeneteit is sikerrel alkalmazták (pl. Senior és mtsai, 2014); ugyanakkor ahhoz, hogy a már bejáratott, gaussi modellezési technikákat minél nagyobb pontossággal lehessen alkalmazni az így kinyert jellemzőkön, szükséges lehet azok transzformálása (Zhang és Woodland, 2014).

Az egyénre jellemző intonációs vagy hangsúlyozási mintázatok automatikus analízise mellett a szövegtagolás és központozás egyéni specifikumait is vizsgájuk: menynyiben tárhatók fel egyénre jellemző mintázatok, milyen konfidenciával. Automatikus eszközök használatára törekszünk (prozódiai esemény detektálókra és intonációs osztályozókra), amelyeket a kísérleti rendszerbe is integrálunk. Vizsgáljuk továbbá, hogy milyen módon lehet az egyes akusztikai-fonetikai jellemzőket egymással kombinálni, és ezeknek milyen hatás van a rendszer kimenetére. Elemezzük tovább, hogy a különböző akusztikai jellemzőkkel kapott kimeneti értékeket (valószínűségi-arány érték: Likelihood Ratio Score) milyen módon lehet kombinálni a rendszer végleges eredményének javításához.

Az i-vektorok számításának egy neurális hálózattal megvalósított enkóder-dekódereken alapuló alternatív eljárását is kidolgozzuk, amitől a beszélők mélyebb jellemzését, leírását várjuk. Az enkóder bemenetére a beszélőtől származó, a dekóder kimenetére univerzális vagy beszélőfüggetlen mintát téve a rejtett rétegen kinyerhető a tömör,

átlaghanghoz viszonyított beszélőreprezentáció. Tervezzük, hogy ezzel a módszerrel végzünk kísérleteket az adatbázison mérve annak performanciáját.

5 A FORvoice elérhetősége

Az adatbázist előreláthatólag 2022-ben fogjuk nyilvánosan elérhetővé tenni.

6 Összegzés

A FORvoice fontos mérföldköve lehet a magyarországi kriminalisztikai tudományának. Lehetőséget biztosít a beszélőre specifikus akusztikai-fonetikai, nyelvészeti jegyek vizsgálatához magyar nyelven a fonetikai, a nyelvészeti és a beszédtechnológiai szakemberek számára. Társadalmi hasznossága kiemelkedő, mivel ez lesz az első olyan magyar beszédkorpusz, amely kriminalisztikai szempontoknak megfelel és így standardizálttá válik.

Az adatbázis lehetőséget biztosít mind a tudományos szakmai közönség, mind pedig a rendészeti szervek számára, a törvényszéki hang-összehasonlítás módszertanának vizsgálatára, illetve általánosságban beszélőfelismerő rendszerek kifejlesztésére és kiértékelésére. A FORvoice adatbázison végzett kutatások új ismereteket nyújtanak az adatközlő egyéni beszéd-sajátosságairól, és alapot adnak további nyelvészeti, beszédtechnológiai vizsgálatokhoz. A fejlesztendő adatbázis – szigorú protokolljának, felvételi módszertanának, az annotálásnak és mennyiségi jellemzőinek köszönhetően (sok beszélő, utánkövetéses eljárás, beszélőnként több felvétel, különböző beszédtípusok) – kiválóan alkalmazható elsősorban

- a bűnügyi beszélőazonosításban,
- az automatikus beszédfelismerő rendszerekben,
- a beszédszintézisben és beszédfelismerésben a beszélőadaptációban,
- de tágabb felhasználási területként minden követéses adatot igénylő kutatásban vagy fejlesztésben – így a fonetika, a beszéd alapú egészségügyi diagnosztika, stb.

A tervezett adatbázis nemzetközi tudományos értéke mellett, jelentős nemzeti értéket is képvisel.

Köszönetnyilvánítás

Az FK128615 számú projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, az FK pályázati program finanszírozásában valósult meg.

Hivatkozások

Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 230-234). IEEE. (2014)

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: Beszélt nyelvi adatbázis [BEA – A Hungarian Spontaneous Speech Database]. In Gósy, M. (ed.): Beszéd, adatbázis, kutatások. Budapest: Akadémiai Kiadó. pp. 9-24. (2012)

Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40. (2010)

Morrison, G. S., Rose, P., Zhang, C.: Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Australian Journal of Forensic Sciences, 44(2), 155-167. (2012)

Saks, M. Koehler, J.: The coming paradigm shift in forensic identification science. Science Magazine, 309, 892-895 (2005)

Vicsi, K., Kocsor, A., Teleki, Cs., Tóth, L.: Beszédadatbázis irodai számítógép-felhasználói környezetben, Second Conference on Hungarian Computational Linguistics (MSZNY 2004), Szeged, 2004. po. 315 (2004)

Vicsi, K., Vig, A.: First Hungarian Speech Database. Beszédkutatás ’98. pp. 163–177. (1998)

Zhang, C., Woodland, P. C.: Standalone training of context-dependent deep neural network acoustic models. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5597-5601). IEEE. (2014)