# FORvoice 120+: Statisztikai vizsgálatok és automatikus beszélő verifikációs kísérletek időben eltérő felvételek és különböző beszéd feladatok szerint

Sztahó Dávid, Beke András, Szaszák György

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

**Kivonat:** A jelen tanulmányban a FORvoice120+ adatbázison végzett akusztikai-fonetikai elemzéseket és automatikus beszélő azonosítási kísérleteket mutatjuk be, a jelenleg elkészült 60 beszélő felvételeivel. Személyfüggő akusztikai jellemzők statisztikai vizsgálatait és automatikus beszélő verifikációs teszteket végeztünk különböző időbeli és beszéd típusbeli eltérések elemzésére. A statisztikai elemzéseknél alaphanghoz, formánsokhoz és beszéd tempóhoz kapcsolódó akusztikai-fonetikai jellemzőket vizsgálatunk. Az eredmények azt mutatták, hogy az eltérő időben történő hangrögzítések alig befolyásolták a jellemzők statisztikai értékeit, míg az eltérő beszédfeladatoknál jelentős eltérés volt tapasztalható. Automatikus beszélő azonosítási (verifikációs) kísérleteket is végeztünk i-vektor és x-vektor implementációkkal. A tesztek alapján elmondható, hogy minél hosszabb beszéd szegmenseket alkalmazunk, annál pontosabb lesz a felismerési eredmény.

## 1 Bevezetés

Az igazságügyi hangszakértői gyakorlatban az utóbbi időszakban megjelentek és kezdenek elterjedni azok az automatikus módszerek, amelyek egy paradigmaváltás következtében jöttek létre (Morrison, 2011; Saks & Koehler, 2005). Ez a paradigmaváltás igyekszik feloldani azt a kérdést, hogy a mért értékek mennyire tipikusak az egyénre, illetve a populációra nézve. Ez az eljárásmód a kriminalisztika egyéb azonosító technológiáinak módszertanában (pl. DNS azonosítás) is megjelent, és egy egységes összehasonlító rendszert tesz lehetővé, amelybe minen egyéni jellegzetesség mérése beilleszthető valószínűségi értékekkel. Az új paradigma a valószínűségi-arány keretrendszer (likelihood-ratio framework, LR) mennyiségi megvalósítását eredményezi, amely során két hipotézist kell vizsgálni: „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?”. Ezek aránya fejezi ki a bizonyítékok erősségét:

$$LR = \frac{p(E|H_{azonos\ személy})}{p(E|H_{eltérő\ személy})}$$

Az LR rendszerén belül a hang alapú beszélőazonosítási kísérletek elvégzéséhez egy olyan adatbázisra van szükség, amely megfelel az új paradigma alapfeltevéseinek (Beke és mtsai., 2020; Morrison és mtsai., 2012):

1) több alkalommal kell minden beszélőtől mintákat rögzíteni (hasonlóság modellezése),
2) sok beszélőt kell tartalmaznia lehetőleg a populációra reprezentatíven (a tipikusság modellezéséhez),
3) különböző módon rögzített hangmintákat kell felvenni (ún. channel mismatch kompenzálására, pl. telefonos vagy stúdió minőségű),
4) egy beszélőtől különböző beszédtípusokat kell rögzíteni a beszédstílus különbségeiből fakadó beszélőn belül is megjelenő eltérések kompenzálására (speech style mismatch compensation).

A jelen cikkben bemutatott kísérletek egy ilyen, most készülő adatbázison valósultak meg. A FORVoice 120+ beszédadatbázis 120 beszélő felvételeit fogja tartalmazni. Ebből jelenleg 60 beszélő felvételei készültek el, amelyeken az eredményeket bemutatjuk. Az adatbázis lehetővé teszi automatikus beszélő azonosítási és -verifikációs kísérletek futtatását, amelyek során eltérő időbeli felvételek és eltérő beszéd feladatok összehasonlítását lehet elvégezni. Ezzel hozzájárul a kriminalisztikai célú hang összehasonlítások módszertanához.

Az automatikus beszélő azonosítás és verifikáció jelenlegi baseline rendszerének, amely illeszkedik a LR-ratio keretrendszerbe, az x-vektorokat használó megvalósítás számít (Snyder és mtsai., 2018). Ez a korábbi i-vektoros megoldást váltotta fel (Dehak és mtsai., 2009, 2010) a deep learning elterjedése által létrejött mély neurális hálózatokat megvalósító technikákkal. A felvételeken hallható személyek azonosítása során több eltérő összehasonlítási módot tudunk megkülönböztetni. A beszélők azonosítása (*speaker identification*) során a felismerendő személyazonosság már egy meglévő zárt halmazból kerül ki, tehát előre tudjuk, hogy kik azok a beszélők, akik közül fel kell ismernünk a felvételen hallottat. Ezzel szemben a beszélő verifikáció (*speaker verification*) két beszédminta hasonlóságának mértékét hivatott megállapítani. Ilyen szituációval találkozhatunk tipikusan akkor, amikor egy célszemély azonosságát szeretnénk verifikálni, megerősíteni, hogy tényleg ő hallható a felvételen. Ekkor rendelkezésünkre áll a célszemélytől valamennyi hanganyag, ami alapján egy beszédlenyomatot képzünk, és az igazolni kívánt felvételből nyert lenyomat ehhez való hasonlóságát szeretnénk mérni.

A jelen tanulmány személyfüggő akusztikai jellemzők statisztikai vizsgálatait és automatikus beszélő verifikációs tesztek eredményeit mutatja be különböző időbeli és beszéd típusbeli eltérések elemzésére. A statisztikai tesztek során megvizsgáltuk, hogy az időben eltérő felvételek és a beszédfeladat típusa befolyásolja-e a mérhető akusztikai-fonetikai paramétereket. Az automatikus beszélő verifikációs kísérletek során pedig az eltérő időtartamú egységek hatását vizsgáltuk. A 2. fejezetben bemutatjuk az adatbázist, utána pedig az elemzéshez alkalmazott eljárásokat írjuk le. Ezután következnek az elért eredmények a 4. fejezetben, majd az ezekből levonható konklúziós és összefoglalás.

# 2 Adatbázis

A bevezetőben ismertetett igazságügyi hangszakértői kísérletekhez készítendő adatbázis tervezett 120 beszélőjéből 60 beszélő felvétele készült el eddig. A felvételek stúdió minőségű fejmikrofonokkal készültek csendes szobában. A felvételi paraméterek: 44.1kHz mintavételi frekvencia, 16 bites kvantálás, PCM lineáris kódolás. A beszélők (beszélgető partnerek) egy szobában tartózkodtak, egymástól 2-3 méter távolságban. Az adatbázis jelenleg 31 férfi (életkor: 24.2±4.6) és 29 női (életkor: 24.4±5.2) beszélőt tartalmaz. Egy beszélőtől két, eltérő időben készült felvétel került rögzítésre. A két felvétel között két hét telt el minden esetben. Ezeket jelöljük a továbbiakban session 1 és session 2 kifejezésekkel. Minden felvétel három beszédfeladatot tartalmaz (jelölése: task 1-3):

1. Szabad párbeszéd (10 perc).
2. Irányított információcsere (~8 perc): hibás terméklistákon található olvashatatlan információk beszerzése a beszélgető partnertől.
3. A megelőző nap tárgyilagos elmesélése (~3 perc).

A kísérletekhez a felvételeket feldaraboltuk az 500 ms-ot meghaladó időtartamú szünetek mentén. Az 1 másodpercnél rövidebb szakaszokat elhagytuk az így kapott közlések közül. Ezután a közléseket három csoportba osztottuk időtartam szerint: (i) 1-2 mp közöttiek, (ii) 2-5 mp közöttiek, valamint (iii) 5 mp felettiek.

# 3 Eljárások

## 3.1 Statisztikai tesztek

A munka során különböző statisztikai teszteket végeztünk, amelyekben különböző akusztikai jellemzők eltérését vizsgáltuk bizonyos szempontok szerint. Az alkalmazott statisztikai eljárás a Generalized Linear Mixed Models (GLMM) volt (McCulloch & Neuhaus, 2014). A teszteket az SPSS 22.0 (Corp, 2013) verziójával valósítottuk meg. A GLMM során az alkalmazott *id*-k a beszélők azonosítói voltak, a *fixed effects* változók pedig a *session* és *task* azonosítók. Ilyen módon mérhetők voltak, hogy a felvétel időpontja és a feladat típusa szerint kimutatható-e statisztikai eltérés a mért akusztika-fonetikai értékek között. Ez információt ad arról, hogy mennyire alkalmazhatók az időben, illetve a feladat típusok szerint eltérő felvételek a személyek azonosítása során.

## 3.2 Akusztikai-fonetikai jellemzők

Az akusztikai-fonetikai jellemzőket a Praat (Boersma, 2001) segítségével számítottuk ki. A következő jellemzőket vizsgáltuk (a későbbiekben használt jelölésüket zárójelben tüntettük fel).

(1) A felvételenként számított *alaphang értékek átlaga és szórása (f0.avg, f0.std)*. A számítási ablakméret 50 ms volt, 10 ms-os lépésközzel. Minden felvételre kiszámítottuk az összes alaphang értéket (ahol zöngés hangok fordultak elő), és ezekenek vettük az átlagát és szórását.

(2) *Artikulációs sebesség (art_tempo)*. Minden felvételre kiszámítottuk az artikulációs sebességet.

(3) *Pairwise variability indices (rPVIc, rPVIv, nPVIc, nPVIv)*. Felvételenként kiszámítottuk külön a magánhangzó és mássalhangzó időtartamok időtartamának változásának mérőszámát (Grabe & Low, 2002). Két változatot alkalmaztunk: nyers (*raw*) és normalizált, amelyeket a következő módon számítottunk:

$$rPVI = \left[\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{(m-1)}\right] \text{ és}$$

$$nPVI = 100 * \left[\sum_{k=1}^{m-1} \left|\frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}}\right| / (m-1)\right],$$

ahol a $d_k$ a *k.* fonémát jelöli, az $m$ pedig a fonémák teljes számát. Mindkét jellemzőt külön kiszámítottuk a mássalhangzókra és a magánhangzókra is (jelölésük: *rPVIc, rPVIv, nPVIc, nPVIv*).

(4) Felvételenkét az /e/ és /a/ hangokon számolt *első három formáns és sávszélességük átlaga és szórása (E.f1.avg, E.f1bw.avg, E.f1.std, E.f1bw.std*, a további formánsok és az /a/ hang hasonlóan jelölve*)*. A számításnál 25 ms-os ablakméretet és 10 ms-os lépésközt alkalmaztunk.

A jellemzőket beszédfeladatonként és *session*-önként számoltuk ki, tehát pl. egy alaphang átlagértéket számoltunk a *session 1* és *task 1* felvételen, egyet a *session 1 task 2* felvételen, és így tovább. A jellemzők kiszámításánál nem vettük figyelembe a szünetek mentén való darabolást.

### 3.3 Beszélő verifikáció

A statisztikai vizsgálatok után automatikus beszélő verifikációs kísérleteket végeztünk az eddig elkészült adatbázison annak érdekében, hogy megvizsgáljuk, hogy az eltérő időtartamú felvételek mennyire befolyásolják a gépi modellekkel kapott eredményeket.

Az automatikus beszélő verifikációs kísérletekhez i-vektor (Dehak és mtsai., 2009) és x-vektor (Snyder és mtsai., 2018) alapú megoldásokat használtunk fel. A megvalósítások KALDI keretrendszerben készültek David Snyder receptjei alapján, amelyek jelenleg state-of-the-art *baseline* megoldásoknak számítanak (*i-vector and x-vector KALDI recipe*, 2018).

Az i-vektor implementálása során 512 keverékszámú GMM-UBM modellt használtunk, az i-vektorok mérete pedig 100 volt. Az x-vektor esetén a tanított TDNN 400 dimenziós volt. Akusztikai jellemzőként mindkét megoldásnál 12 MFCC-t alkalmaztunk. Szintén mindkét eljárás esetén PLDA-val (Ioffe, 2006) történt a tesztesetek kiértékelése. Az eredményeket *equal error rate* (EER) szerint mértük.

A felvételeket két halmazra osztottuk. A 60 beszélő közül 40-et használtunk a i-vektor kinyerő, az x-vektor TDNN valamint a PLDA tanítására. A maradék 20 beszélő

1. Táblázat: Az akusztikai-fonetikai jellemzőkre kapott, GLMM-el mért *p* értékek a *task* és *session* változókra. A 99%-on szignifikáns** és 95%-on szignifikáns* eltéréseket külön jelöltük.

| jellemző | *task* | *session* |
|---|---|---|
| art_tempo | 0.000** | 0.068 |
| rPVIc | 0.000** | 0.105 |
| rPVIv | 0.000** | 0.185 |
| nPVIc | 0.000** | 0.153 |
| nPVIv | 0.042* | 0.286 |
| E.f1.avg | 0.023* | 0.769 |
| E.f2.avg | 0.000** | 0.213 |
| E.f3.avg | 0.000** | 0.578 |
| E.f1.std | 0.000** | 0.864 |
| E.f2.std | 0.000** | 0.225 |
| E.f3.std | 0.000** | 0.402 |
| E.f1bw.avg | 0.169 | 0.698 |
| E.f2bw.avg | 0.001** | 0.715 |
| E.f3bw.avg | 0.534 | 0.171 |
| E.f1bw.std | 0.103 | 0.484 |
| E.f2bw.std | 0.021* | 0.654 |
| E.f3bw.std | 0.605 | 0.274 |
| O.f1.avg | 0.103 | 0.971 |
| O.f2.avg | 0.000** | 0.970 |
| O.f3.avg | 0.177 | 0.700 |
| O.f1.std | 0.000** | 0.620 |
| O.f2.std | 0.000** | 0.582 |
| O.f3.std | 0.001** | 0.786 |
| O.f1bw.avg | 0.000** | 0.593 |
| O.f2bw.avg | 0.032* | 0.564 |
| O.f3bw.avg | 0.048* | 0.825 |
| O.f1bw.std | 0.160 | 0.788 |
| O.f2bw.std | 0.086 | 0.172 |
| O.f3bw.std | 0.674 | 0.200 |
| F0 avg | 0.001** | 0.579 |
| F0 std | 0.814 | 0.032* |

felvételeit alkalmaztuk tesztelésre a következő módon. A *session 1*-be tartozó felvételek kerültek az *enrollment* halmazba, vagyis ezek voltak azok a felvételek, amelyeken a beszélők átlagvektorait számítottuk ki. A *session 2*-be tartozó felvételek voltak a konkrét tesztesetek (*target*), amelyekben a személyazonosságot meg kellett állapítani.

Az előzetes kísérletek azt mutatták, hogy a 40 beszélő hanganyaga nem elegendő a TDNN és a PLDA tanítására, ezért ehhez még felhasználtuk a BABEL (Roach és mtsai., 1996) és az MRBA (Vicsi & Vig, 1998) adatbázisokat is, amelyekben összes 388 beszélő szerepelt, összesen 120 percnyi beszédidővel.

# 4 Eredmények

## 4.1 GLMM eredmények

A GLMM-el végzett elemzések statisztikai eredményeit az 1. táblázat tartalmazza. Minden jellemzőhöz megadtuk a feladat (*task*) és felvételi időpont (*session*) szerinti p értéket. A 95%-os, valamint a 99%-os szignifikancia szintű eltéréseket külön kiemeltük (* és ** jelölések). A jellemzőket a 3.2 fejezetben leírt jelölésekkel láttuk el.

A *session* változó esetén csupán egyetlen esetben tértek el a mért értékek statisztikailag egymástól, az alaphang szórásakor. Minden más esetben azt mondhatjuk, hogy a felvételek időben eltérő rögzítése nem volt hatással arra, hogy a mért értékek jelentősen eltérnek-e egymástól.

A beszédfeladatok esetén pont az ellenkező jelenséget tapasztaltuk. Csupán néhány olyan jellemző van, amelynél nincs szignifikáns különbség a *task* változó értékeinek függvényében.

A beszéd tempóra vonatkozó jellemzők (artikulációs tempó és PVI jellemzők) mind szignifikáns eltéréseket mutatnak az eltérő beszédfeladatoknál, míg az időbeli eltérés nem volt számottevő hatással a mért értékekre. A formánsok esetén vegyes szignifikanciájú eltéréseket mértünk a beszédfeladatok között, ám az időben eltérő felvételek itt sem mutattak sehol sem eltéréseket. Az átlagos alaphang értékek szignifikánsan megváltoztak, ha eltérő beszédfeladatról volt szó, ám nem voltak eltérő populációból tekinthetőnek, ha az időbeli eltéréseket nézzük. Ezzel szemben az alaphangok felvételenkénti szórása pont a *session* változó szerint volt eltérő, míg a *task* változó nem volt rá jelentős hatással. Az összes eredményt tekintetbe véve azt mondhatjuk, hogy az időben eltérő felvételek kevésbé (alig) befolyásolják a beszélőkre jellemző értékeket, az eltérő beszédfeladatok viszont jelentős hatással vannak, ezért azokat majd érdemes figyelembe venni a későbbiekben a beszélő azonosítási kísérletek során.

## 4.2 Automatikus beszélő verifikáció

Az automatikus beszélő verifikációs kísérletek során azt vizsgáltuk meg, hogy a különböző időtartamú beszédszakaszok hogyan befolyásolják az beszélők azonosításának pontosságát. Ehhez a következő teszteseteket végeztük el:

(a) minden beszédszakasz felhasználása (*all*),

(b) a 1-2 mp közötti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) (*1-2v1-2*),
(c) a 2-5 mp közötti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) (*2-5v2-5*),
(d) az 5 mp feletti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) (*5v5*),
(e) 1-2 mp időtartamú felvételekkel történő tanítás, és 5 mp feletti időtartamú felvételekkel való tesztelés (*1-2v5*),
(f) 5 mp feletti időtartamú felvételekkel való tanítás, 1-2 mp időtartamú felvételekkel való tesztelés (*5v1-2*). A kapott eredményeket a 2. táblázat tartalmazza, ahol a teszteseteket az előzőekben leírt módon jelöltük. A táblázatban a PLDA pontozással kapott EER% értékeket tüntettük fel. Mivel jelenleg még nincs elegendő hanganyag a beszédfeladatok külön alkalmazására, ezért az összes beszédfeladatot felhasználtuk a kísérletek során.

Az eredmények azt mutatják, hogy az i-vektoros megvalósítás alacsonyabb tévesztési százalékokat produkál annak ellenére, hogy az x-vektoros rendszer elvileg újabb technológiának számít. Ez azért lehet, mert az x-vektor kinyerés mély tanuláson alapul, így tanításukhoz sokkal több hanganyag szükséges, mint az i-vektorhoz. A jelenlegi adatbázis nem éri el az a méretet, amivel a TDNN háló tanítható (az MRBA és BABEL kiegészítéssel együtt sem). A nemzetközi irodalomban sem egyértelmű az x-vektor alapú megközelítés felsőbbrendű helyzete (Kanagasundaram és mtsai., 2011; Sarkar és mtsai., 2012).

Az összes hanganyaggal elvégzett kísérletek (*a* eset) eredménye (5.4% EER) összevethető a nemzetközi irodalommal (Snyder és mtsai., 2018). Ahogy azt várni lehetett, az 5 mp-nél hosszabb felvételekkel kaptuk a legjobb eredményt (a hosszabb minták jobban leírják a beszélőt). Ennek megfelelően a legrövidebb minták (1-2 mp) adták a legrosszabb azonosítást (7.727% a 3.193%-hoz képest).

2. Táblázat: Az automatikus beszélő verifikáció eredményei. PLDA-val kapott EER %-ok az i-vektor és x-vektor implementációkra.

| Teszteset | i-vektor | x-vektor |
|---|---|---|
| *all* | 5.405 | 9.276 |
| *1-2v1-2* | 6.605 | 11.38 |
| *2-5v2-5* | 3.957 | 6.345 |
| *5v5* | 3.193 | 1.739 |
| *1-2v5* | 3.193 | 1.91 |
| *5v1-2* | 7.727 | 10.56 |

## 5 Konklúzió

Az akusztikai-fonetikai paraméterek elemzése alapján elmondható, hogy a felvételek időbeli eltérése nem mutatott jelentős eltérést a mérésekben. Tehát ez a változó nem

okoz zavart akkor, amikor beszélő verifikációt, azonosítást végzünk. Csupán az, hogy egy adott személytől különböző időben rögzítünk hanganyagot, nem befolyásolja az azonosítást (ha egyéb beszédképzést befolyásoló tényező, például megfázás, nem jelentkezik).

Ezzel ellentétben, a beszéd stílusát meghatározó változó (jelen esetben a beszéd feladat) jelentős hatással volt a mérhető eltérésekre. A monológok és a szabad párbeszédek során megfigyelhető volt olyan eltérés, amely szignifikánsnak mutatkozott. Célszerű tehát egy adott személytől sokféle beszéd helyzetet rögzíteni, ha személyazonosítást megvalósító feladatról van szó.

Az automatikus beszélő azonosítást célzó kísérletek során az megnyilatkozások időtartama (*utterances*) hatással vannak a beszélő azonosítás pontosságára. Minél hosszabb felvétel áll rendelkezésre, annál jobb eredményt lehet elérni az általánosan elterjedt i-vektor és x-vektor alapú rendszerrel. 5 másodpercnél hosszabb felvételek esetén 1.739% EER-t lehet elérni.

A jelenleg rendelkezésre álló hanganyag 60 beszélőt tartalmaz. A végső tervezett 120 beszélővel már robosztusabb eredményeket és elemzéseket lehet majd elkészíteni. Ezen kívül ez már elegendő lesz ahhoz is, hogy a beszéd feladatok közötti eltéréseket automatikus verifikációs kísérletekkel vizsgáljuk.

## 6 Összefoglalás

A jelen tanulmányban a FORvoice120+ adatbázison végzett akusztikai-fonetikai elemzéseket és automatikus beszélő azonosítási kísérleteket mutattuk be, a jelenleg elkészült 60 beszélő felvételeivel.

A statisztikai mérésekhez alaphangból, formánsokból és beszéd tempóhoz kapcsolódó akusztikai-fonetikai jellemzőket alkalmaztuk. Az eredmények azt mutatták, hogy az eltérő időben történő hangrögzítések alig befolyásolták a jellemzők statisztikai értékeit, míg az eltérő beszédfeladatoknál jelentős eltérés volt tapasztalható.

Automatikus beszélő azonosítási (verifikációs) kísérleteket is végeztünk i-vektor és x-vektor implementációkkal. A tesztek alapján elmondható, hogy minél hosszabb beszéd szegmenseket alkalmazunk, annál pontosabb lesz a felismerési eredmény.

## Köszönetnyilvánítás

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

## Hivatkozások

Beke, A., Szaszák, G., & Sztahó, D. (2020). FORvoice 120+: Magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra. In G. Berend, G. Gosztolya, & V. Vincze

(Szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (o. 95–101). Szegedi Tudományegyetem, Informatikai Intézet; MTMT. https://m2.mtmt.hu/api/publication/31148107

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, *5*(9), 341–345.

Corp, I. B. M. (2013). IBM SPSS statistics for windows, version 22.0. *Armonk, NY: IBM Corp*.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *Tenth Annual conference of the international speech communication association*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(515–546).

Ioffe, S. (2006). Probabilistic linear discriminant analysis. *European Conference on Computer Vision*, 531–542.

*I-vector and x-vector KALDI recipe*. (2018). https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16

Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S., & Mason, M. W. (2011). I-vector based speaker recognition on short utterances. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2341–2344.

McCulloch, C. E., & Neuhaus, J. M. (2014). Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*(3), 91–98.

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

Roach, P., Arnfield, S., Barry, W., Baltova, J., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meister, E., & Vicsi, K. (1996). BABEL: An Eastern European multi-language database. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1892–1893 köt.3. https://doi.org/10.1109/ICSLP.1996.608002

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895.

Sarkar, A. K., Matrouf, D., Bousquet, P. M., & Bonastre, J.-F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. *Thirteenth Annual Conference of the International Speech Communication Association*.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333.

Vicsi, K., & Vig, A. (1998). First Hungarian speech database. *Beszédkutatás*, *98*, 163–177.