Magyar Rendészet 2021/2. 47–65. DOI: 10.32577/mr.2021.2.3

Biometrikus beszélőazonosító rendszerek performanciája összehasonlításának elmélete és gyakorlata

FEJES Attila[1]

A biometrikus (automatikus) beszélőazonosítás széleskörűen alkalmazott mind a hazai, mind a nemzetközi kriminalisztikai gyakorlatban. A módszertan nagy sebességgel, kiválóan automatizálható adatfeldolgozási lehetőségekkel rendelkezik, pontos és valid eredményeket szolgáltat. A biometrikus azonosító rendszerek az összevetett hangfelvételeken beszélők azonosságának valószínűségét adják meg. Egy rendszer performanciájának meghatározásához azonosítási mátrixot kell előállítani, amely a valószínűségi értékeket tartalmazza. Tanulmányomban ismertetem a mátrixok előállításának folyamatát és szempontjait, az adatstruktúra felépítését. 136 beszélő személy hangmintáját használtam fel, amelyeket különböző időpontokban és eszközökkel rögzítettem. Az Oxford Wave Research Ltd. Vocalise és Phonexia biometrikus azonosító rendszerekkel létrehoztam a mátrixokat, illetve a match és nonmatch adatokat, amelyeket a Bio-Metrics performanciamérő szoftverrel értékeltem ki. Az eredmények értékelése megmutatta, hogy a teljesítőképesség meghatározásához több típusú kimenetet is fel kell használni, nem elegendő a leggyakrabban publikált Egyenlő Hibaarány (EER) közlése. A közel 40 ezer vizsgált valószínűségi érték elemzése alapján a megadott rendszerek megbízhatóan, megfelelő diszkriminatív erővel képesek azonosítani az egyező, és megkülönböztetni az eltérő személyeket.

Kulcsszavak: *beszélőazonosítás, hangbiometria, Likelihood Ratio (LR), performancia, hibaarányok*

1. A beszélőazonosítás alapjai

A beszéd alapján történő személyazonosítás alapja, hogy nincs két, teljesen megegyező fizikai és pszichológiai jellemzőkkel rendelkező ember, akik ugyanazon szociális és társadalmi környezetben nőttek fel, és akik beszédkészsége, nyelvhasználata,

[1] Fejes Attila nemzetbiztonsági őrnagy, hangtechnikai szakértő, Nemzetbiztonsági Szakszolgálat Szakértői Intézet; doktori hallgató, Nemzeti Közszolgálati Egyetem Rendészettudományi Doktori Iskola.
Attila Fejes National Security Major, Audio Forensic Expert, Institute for Expert Studies of the Special Service for National Security; PhD student, University of Public Service Doctoral School of Police Sciences and Law Enforcement.
E-mail: fejes.attila@nbsz.gov.hu, ORCID: https://orcid.org/0000-0003-4139-5718

kommunikációs képessége és szokása teljesen megegyezik. Mivel a beszédprodukciót fizikai és pszichikai adottságok, beszédképzési folyamatok, külső hatások, tanult szokások határozzák meg, mindezek egybevágósága lenne ahhoz minimálisan szükséges, hogy két különböző ember beszédprodukciója teljesen egyező legyen. Lehetséges, hogy a – remélhetőleg minél távolabbi – jövőben, amikor az emberi klónozás valóra válik, a beszéd egyediségét meghatározó jellemzőket majd újragondolják, de addig mondhatjuk, hogy még az azonos szociokulturális közegben felnőtt egypetéjű ikreknek a beszéde is jól mérhetően különbözik, ahogy Künzel tanulmányában[2] rámutatott az automatikus beszélőazonosítás alkalmazásával. Másrészt – csak elméleti síkon vizsgálódva – ha feltételezzük is a 100%-ban egyező anatómiai szerkezetet, a beszédet meghatározó további sajátosságok nem lehetnek ugyanazok: például általános műveltség, érdeklődési kör, szókincs, EQ- és IQ-hányados stb., tehát csupa olyan dolog, amelyet a beszélő szociokulturális közege és saját személyisége, képességei, nem utolsósorban érzelmei határoznak meg.

Az emberi beszéd variabilitása következtében nem vagyunk képesek egy hangot, hangsort pontosan ugyanúgy, mindenben egyező paraméterekkel kiejteni (vagy például egy betűt, írásproduktumot létrehozni) egy későbbi időpontban – eltekintve ennek statisztikai valószínűségétől. Ezzel együtt jelenleg még nem ismert, hogy a beszéd mely paraméterei reprezentálják pontosan az egyediséget, így az azonosítási eljárás során nem kulcsjellemzőket[3] keresünk, amelyek egyezősége alátámasztaná az azonosságot, hanem a beszédprodukciók összehasonlítása történik meg. Ez elvégezhető percepciós elemzéssel, akusztikai-fonetikai vizsgálatokkal és biometrikus (automatikus) módszertan alkalmazásával.

A percepciós elemzés során a szakértő észleléses úton detektálja a beszéd- és hangképzés egyéni jellemzőit (dallamvezetés, hangszínezet, beszédhibák, megakadásjelenségek stb.) és a beszédben fellelhető nyelvhasználati szokásokat. Az akusztikai-fonetikai vizsgálatok hangszínkép-összehasonlítás, a hangról különböző algoritmusok segítségével elkészített görbék és egyes jellemzők adatainak összehasonlítását tartalmazza. A biometrikus beszélőazonosítás során számítógépes rendszer határozza meg a beszélők azonosságának valószínűségét a bemenetre állított hangfelvételek felhasználásával.

Az első két részmódszertani elem rendkívül időigényes (két beszédprodukció összevetése és elemzése minimum 8-10 munkaórát igényel), és a végeredmény függ az eljáró szakértő tudásától, tapasztalatától, értékítéletétől – és a rendelkezésre álló eszközrendszerétől. A hangbiometria gyors (két személy esetében a valószínűségi értéket másodpercek alatt kiszámolja a rendszer), objektív módszer: az eredmény teljesen független a szakértőtől, bármikor reprodukálható. A hangbiometria alkalmazásával nagymennyiségű adatok feldolgozását is elvégezhetjük, amely akkor valósul meg, ha egy vagy több személy beszédét több személy hangmintájával vetjük

[2] Hermann J. Künzel: Automatic speaker recognition of identical twins. *The International Journal of Speech Language and the Law,* 17. (2010) 2. 251–277.

[3] Gósy Mária: *Fonetika, a beszéd tudománya.* Budapest, Osiris, 2004. 273.

össze, megvalósítva az 1:N, vagy az N:N metódust. A biometrikus azonosítás implementálható Big Data technológiai környezetbe, így a klasszikus, kriminalisztikai esetpéldától eltérően – amikor az 1:1 metódussal két beszédhang-mintát hasonlítunk össze – alkalmazható szűrő-kutató munkára, nagy rekordszámú adatbázisokban történő keresésre.

2. A hangbiometria

A biometria az ember egyes fizikai vagy viselkedésbeli jellemzőit használja fel személyazonosítás elvégzéséhez.[4] Ezek lehetnek például ujjnyomat, DNS, az írisz és a retina egyedi mintázata, arc, járás, gépelés, valamint a beszédhang. A különböző biometrikus jellemzők egyedi mintázatainak, tulajdonságainak leképezése változatos módszerekkel valósul meg. A biometrikus rendszerek leképezik a beszélő beszédképzésben részt vevő szerveinek (vokális traktus) karakterisztikáját és statisztikai modellt állítanak fel. Ez a technológia különbözik az emberi hallás és beszédfeldolgozás összetett eljárásától, ami egyik részről fiziológiai folyamat, másrészt idegi-ingerületi átvitelt követően az agyi működés eredménye.[5]

Az emberi beszédhang esetében az első lépés a jellemzőkinyerés, ezt követi a biometrikus modell felállítása, majd matematikai-statisztikai módszerekkel a személyazonosság valószínűségének meghatározása. A GMM-UBM,[6] az x-vektor és az i-vektor biometrikus motorral meghajtott rendszerek a jellemzőkinyerés és a modellezés során a hangot milliszekundum nagyságrendű részekre bontják, és a beszédspektrumából jellemzővektorokat állítanak fel, ezt követően *Score* értéket határoznak meg, amelyeket a valószínűségi számításokhoz használnak. Újabb, néhány éve megjelent technológia a hangbiometriában a mély neurális hálózatok alkalmazása, amely napjaink információtechnológiai világának egyik kulcsfogalmát, a mesterséges intelligenciát hívja segítségül. A neurális hálózatok tanító algoritmusokkal (eljárásokkal) végzik a valószínűségi érték meghatározását, a technológia alkalmazásával a biometrikus beszélőazonosítás teljesítőképessége várhatóan tovább fog növekedni. Egy gondolat erejéig említsük meg az automatikus beszédfelismerést, amely a biometrikus azonosításban alkalmazott fenti eljárásokat használja fel.[7] A beszédfelismerés egyre nagyobb jelentőséget kap napjainkban, gondoljunk csak a virtuális ügyfélszolgálatra, amikor az ügyfél a telefonos ügyintézés egy bizonyos pontjáig csak szintetizált beszéddel találkozik, vagy a rendészeti szervek hangfelvételfeldolgozási feladataira, amikor

---

4 Anil K. Jain – Arun A. Ross – Karthik Nandakumar: *Introduction to biometrics*. London, Springer, 2011. 3.

5 Homayoon Beigi: *Fundamentals of speaker recognition*. London, Springer, 2011. 54.

6 Gaussian Mixture Model – Universal Background Model: a Gauss-eloszlást (normáleloszlást) segítségül hívó módszertan, amely mellett az azonosító rendszer egy több száz órányi beszédhang felhasználásával készült általános háttérmodellt is tartalmaz. Ez utóbbi nem keverendő össze az egyes rendszerek által megkövetelt populációs adatbázissal, amelyet az elvégzendő azonosítási műveletben szereplő beszéd és a hangfájl átviteli csatornájának megfelelő hanganyagok felhasználásával kell konfigurálni.

7 Uday Kamath – John Liu – James Whitaker: *Deep learning for NLP and speech recognition*. Cham, Springer, 2019. 370.

emberi közreműködés nélkül kapjuk meg a hanganyag többé-kevésbé pontos szöveges leiratát.

A hangbiometriában az azonosság valószínűségének kiszámításához a *Bayes*-megközelítés keretrendszerét használják fel, amely a feltételes valószínűség tételeként ismert a matematikában. A *Bayes*-tételből származó *Likelihood Ratio* (LR) leegyszerűsítve két valószínűségi érték *(probabilistic)* hányadosa: $LR = p(E/H_0) / p(E/H_1)$. A képletben E a bizonyíték hangfelvételt *(Evidence)* jelöli, amely az ismeretlen beszélő hangját tartalmazza, $p(E/H_0)$ annak valószínűsége, hogy az ismeretlen személytől rögzített hangfelvétel az ismert személytől, $p(E/H_1)$ annak valószínűsége, hogy az ismeretlen személy hangfelvétele valaki mástól származik. A H_0 hipotézis az azonosságot, a H_1 a különbözőséget állítja. A Bayes-megközelítésnek előnye, hogy a bizonyíték (többnyire az a hangfelvétel, amelyen az ismeretlen beszélő hallható, és amelyen terhelő adatok hangzanak el) súlya és a két valószínűség (azonos-különböző) együtt értelmezhető.[8]

A kriminalisztikában az ismeretlen személytől származó hanganyag keletkezhet például telefonlehallgatás során, míg az ismert személytől rögzített hangmintát gyakran a kirendelt hangtechnikai szakértő készíti el az eljárásba bevont személy, gyanúsított közreműködésével.

A *Likelihood Ratio*[9] fogalma a szakértői bizonyítás során jól alkalmazható, nem igényel különösebb matematikai-statisztikai ismereteket, egyszerű, áttekinthető módon értelmezhetők vele a valószínűségi értékek. Az LR elméleti minimuma a nullát soha nem éri el, hiszen egy tört nevezőjében nem szerepelhet zéró érték. A maximuma a legtöbb biometrikus azonosító rendszer esetében 10 milliárd, nem függetlenül attól, hogy jelenleg közel 8 milliárd ember él a Földön. Az *LR* középértéke 1, amennyiben a képletben szereplő két valószínűségi érték megegyezik, ebben az esetben a beszélők azonossága nem bizonyítható, de nem is zárható ki. Megjegyzendő, hogy az azonosítási eljárások során nem ritka a 10 milliárd maximumérték, míg az $LR = 1$ adat, habár elméletileg lehetséges, a gyakorlatban nagyon kicsi gyakorisággal fordul csak elő.

A *Likelihood Ratio* értelmezése a következő, a Batvox 4.1 verziószámú biometrikus beszélőazonosító rendszerével készült ábrával szemléletesebbé tehető. A képen a 016 jelzésű női beszélő GSM-csatornán rögzített hangfelvétele (ismeretlen személy) és ugyanezen beszélő hangszakértői mintavételi munkaállomással rögzített hangmintája összevetésének eredményei láthatóak. Az Y-tengelyen a valószínűségi, X-tengelyen a *Score* érték van jelölve. Ahogy korábban említettük, a *Score* érték a kulcsa a biometrikus azonosító rendszer működésének, amelynek kiszámításához használt módszerek elmélete ismert, azonban az, hogy a jellemzőkinyerést és a modellalkotást pontosan milyen jellemzők felhasználásával és algoritmussal végzi a szoftver biometrikus motorja, a fejlesztők *Black Box*[10]-ként kezelik.

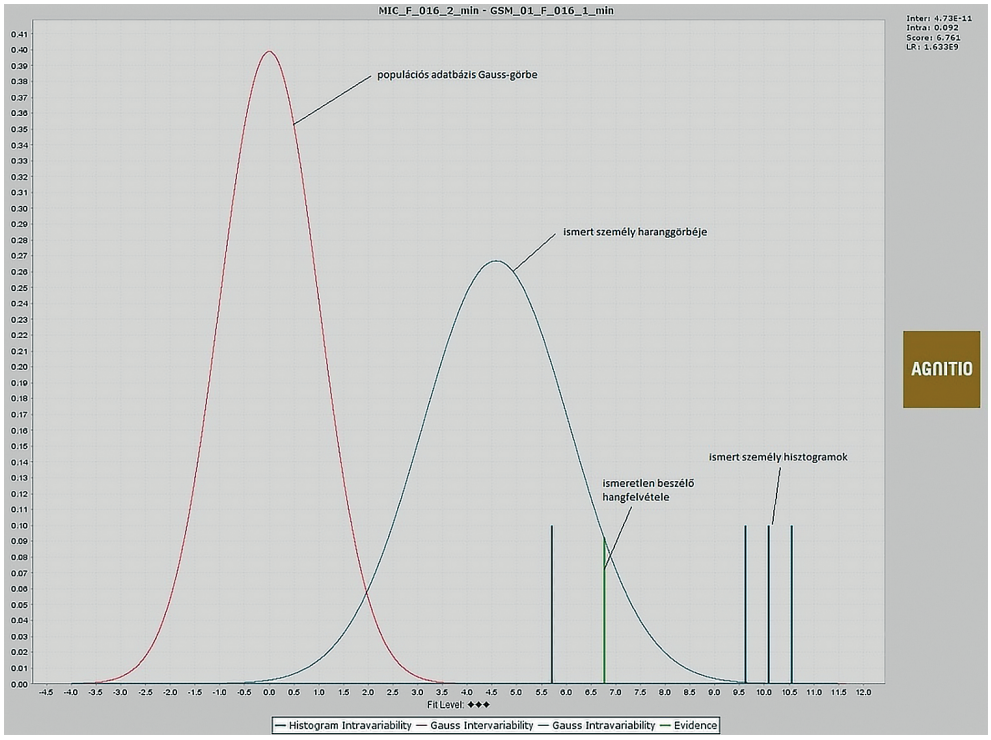
---

[8] Craig Adam: *Mathematics and statistics of forensic science.* Chichester, Wiley-Blackwell, 2010. 286.

[9] Ramos, Daniel – Juan Maroñas – Alicia Lozano-Diez: *Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems.* Madrid, 2017.

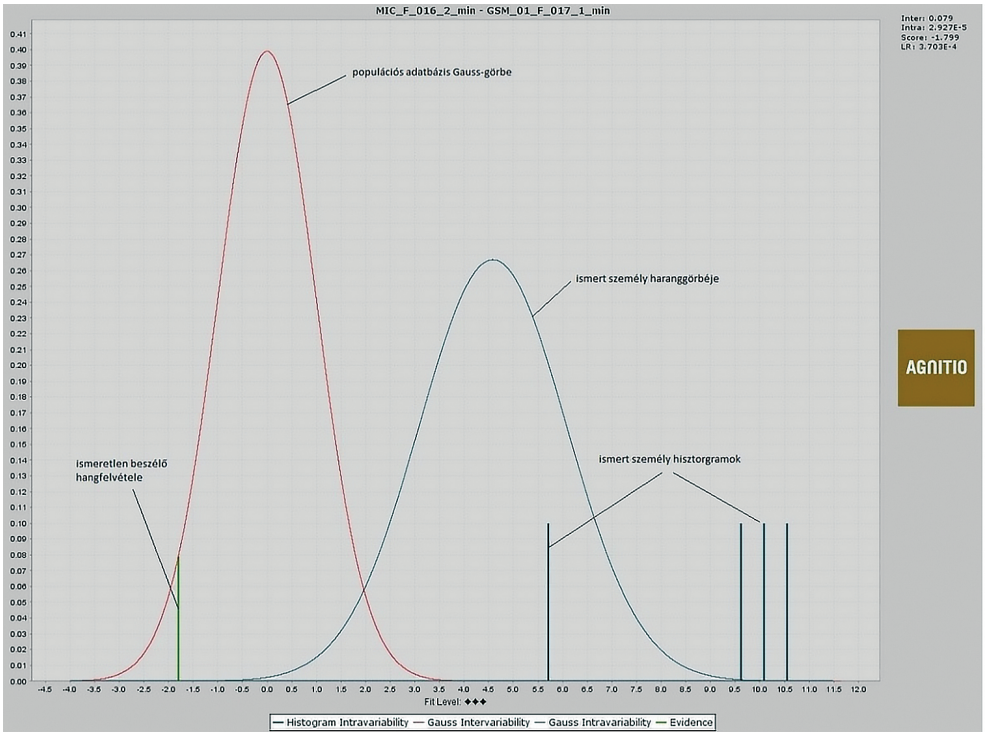
[10] *Black Box:* itt a fejlesztő által ipari titokként kezelt számítási eljárás.

Az 1. ábrán bal oldalon látható a populációs adatbázis haranggörbéje (Gauss-eloszlás), amely a „valaki más” beszélők csoportját reprezentálja. Az ismert személy hangmintája alapján generált Gauss-görbe és hisztogramok jobb oldalon helyezkednek el. Az ismeretlen beszélő hangfelvétele (bizonyíték) alapján készült egyenes és a két Gauss-görbe metszéspontjai és értékei határozzák meg az azonosság valószínűségét.



1. ábra: Beszélők azonosságának ábrázolása LR-grafikonon. Forrás: a szerző szerkesztése

A Batvox szoftver annak valószínűségét, hogy az ismeretlen beszélő hangfelvétele az ismerttől származik, Intravariabilitásnak *(Intravariability)* nevezi, a különbözőség valószínűségét pedig Intervariabilitásként *(Intervariability)* jelöli. A grafikonon az ismeretlen beszélő hangfelvétele alapján készült egyenes a populációs adatbázis haranggörbéjét nagyon kis számértéknél metszi, ennek megfelelően csak $4{,}73 \times 10^{-11}$ (0,0000000000473) annak a valószínűsége, hogy az ismeretlen hangfelvétel valaki mástól származik, míg az azonosság valószínűsége ennél kilenc nagyságrenddel nagyobb (0,092). A két érték hányadosa a *Likelihood Ratio,* amelynek értéke 1,633E9 ($1{,}633 \times 10^{9}$). Ez az *LR* adat azt jelenti, hogy egymilliárd-hatszázharmincmilliószor valószínűbb, hogy az ismeretlen és az ismert beszédhang azonos, mint hogy különböző személytől származik.



2. ábra: Beszélők különbözőségének ábrázolása LR-grafikonon. Forrás: a szerző szerkesztése

Az 1. ábra magyarázatának megfelelően látható, hogy a 2. ábrán a bizonyíték zöld színű egyenese nagyon kis értéknél metszi az ismert személy hangja alapján készített Gauss-eloszlást, míg a különbözőséget reprezentáló piros haranggörbét ehhez képest több nagyságrenddel nagyobb valószínűségnél keresztezi. Ez azt jelenti, hogy csak 0,0003703-szor ($3{,}703 \times 10^{-4}$) valószínűbb, hogy a két beszéd azonos, mint hogy különböző személyhez tartozik.

Az LR-eredmények tízes számrendszerben történő értelmezése aszimmetrikus skálát eredményez, hiszen a különbözőséget 0-nál nagyobb, de 1-nél kisebb számokkal reprezentálja, míg az azonosságot 1-től egészen 10 milliárdig tartó (10 nagyságrenddel nagyobb) értékekkel fejezi ki. A megoldás az adatok konverziója 10-es alapú logaritmus kiszámításával, ahol az *LR* = 1 középérték 0-át (LLR = 0) vesz fel így az azonosság 0–10 között, míg a különbözőség, ezzel szimmetrikusan, –10–0 intervallumban ábrázolható.

3. A performancia

Egy biometrikus beszélőazonosító rendszer performanciájának meghatározásához N:N metódus szerint szükséges összehasonlítani a hangmintákat. Ehhez egy beszélőtől két különböző időpontban rögzített mintákat használunk fel, és minden személy GSM-mintáját összevetünk minden beszélő stúdiómikrofonos hangfelvételével. Jelen tanulmányomban 136 női beszélő hangfelvételeit használtam fel, amelyek különböző időpontban és eltérő eszközökkel készültek. A mérés során 136 × 136 db (18 496) adat keletkezett, amelyekből 136 db az azonos beszélők (SS),[11] 18 360 a különböző beszélők (DS)[12] összevetésének eredményeit mutatja. A mintát adó személyekkel 10-15 perces spontán beszélgetést folytattam le, a hanganyagot GSM-csatornán és stúdiótechnikai eszközökkel rögzítettem. A GSM-felvételek modellezik az ismeretlen beszélő mintáit, míg a MIC jelzésűek a hangszakértő által felvett (ismert személytől származó) hangmintát jelképezik, ezeket 2 perces hosszúságúra editáltam. Az automatikus azonosító rendszerek kimenetén többféle módon jelennek meg a valószínűségi értékek, amelyekből kettőt a következő két ábra mutat be. A 3. ábrán a mobiltelefonos felvételek, a 4. ábrán a stúdiómikrofonos minták szerint vannak listázva az eredmények.

BIOMETRIC_SCORE	AUDIO_SEGMENT_ORIGIN	SPEAKER_ID
7,62934	GSM_01_F_001_1_min.wav	MIC_F_001_2_min
4,14141	GSM_01_F_001_1_min.wav	MIC_F_111_2_min
3,96426	GSM_01_F_001_1_min.wav	MIC_F_099_2_min
3,8505	GSM_01_F_001_1_min.wav	MIC_F_025_2_min
3,7608	GSM_01_F_001_1_min.wav	MIC_F_116_2_min
3,64059	GSM_01_F_001_1_min.wav	MIC_F_078_2_min
3,34572	GSM_01_F_001_1_min.wav	MIC_F_009_2_min
3,23954	GSM_01_F_001_1_min.wav	MIC_F_108_2_min
3,07506	GSM_01_F_001_1_min.wav	MIC_F_093_2_min
3,01125	GSM_01_F_001_1_min.wav	MIC_F_003_2_min
2,94585	GSM_01_F_001_1_min.wav	MIC_F_039_2_min
2,92762	GSM_01_F_001_1_min.wav	MIC_F_112_2_min
2,90069	GSM_01_F_001_1_min.wav	MIC_F_052_2_min
2,8877	GSM_01_F_001_1_min.wav	MIC_F_044_2_min
2,85289	GSM_01_F_001_1_min.wav	MIC_F_015_2_min
2,81896	GSM_01_F_001_1_min.wav	MIC_F_035_2_min
2,79769	GSM_01_F_001_1_min.wav	MIC_F_058_2_min
2,74169	GSM_01_F_001_1_min.wav	MIC_F_007_2_min
2,69781	GSM_01_F_001_1_min.wav	MIC_F_123_2_min
2,63162	GSM_01_F_001_1_min.wav	MIC_F_132_2_min
2,56246	GSM_01_F_001_1_min.wav	MIC_F_105_2_min
2,55688	GSM_01_F_001_1_min.wav	MIC_F_080_2_min
2,51681	GSM_01_F_001_1_min.wav	MIC_F_050_2_min
2,51359	GSM_01_F_001_1_min.wav	MIC_F_119_2_min
2,48804	GSM_01_F_001_1_min.wav	MIC_F_017_2_min

3. ábra: Nuance Forensic rendszer kimenete. Forrás: a szerző szerkesztése

MIC_001			
	File name	Speaker LLR	Quality
	GSM_01_F_001_1_min.wav	6.22	100,00%
	GSM_01_F_009_1_min.wav	1.84	85,00%
	GSM_01_F_099_1_min.wav	0.16	100,00%
	GSM_01_F_116_1_min.wav	0.13	100,00%
	GSM_01_F_123_1_min.wav	-0.29	100,00%
	GSM_01_F_093_1_min.wav	-0.34	100,00%
	GSM_01_F_111_1_min.wav	-0.52	100,00%
	GSM_01_F_015_1_min.wav	-0.58	85,00%
	GSM_01_F_078_1_min.wav	-0.63	88,00%
	GSM_01_F_018_1_min.wav	-0.76	88,00%
	GSM_01_F_052_1_min.wav	-1.03	87,00%
	GSM_01_F_089_1_min.wav	-1.28	100,00%
	GSM_01_F_108_1_min.wav	-1.45	100,00%
	GSM_01_F_132_1_min.wav	-1.48	95,00%
	GSM_01_F_006_1_min.wav	-1.57	89,00%
	GSM_01_F_056_1_min.wav	-1.59	86,00%
	GSM_01_F_031_1_min.wav	-1.66	85,00%
	GSM_01_F_014_1_min.wav	-1.85	86,00%
	GSM_01_F_098_1_min.wav	-1.95	100,00%
	GSM_01_F_083_1_min.wav	-2.02	100,00%
	GSM_01_F_105_1_min.wav	-2.49	100,00%

4. ábra: Phonexia rendszer kimenete. Forrás: a szerző szerkesztése

[11] SS: Same Source.
[12] DS: Different Source.

A performancia meghatározásához két út kínálkozik: az egyik, hogy külön adatállományba válogatjuk le a *Same Source* és a *Different Source* eredményeket, a másik pedig, hogy azonosítási mátrixot hozunk létre. A későbbiekben láthatjuk, hogy mindkét módszer együttes alkalmazása lehet a célravezető különböző rendszerek esetében. Az 5. ábra táblázatának az első sora és az első oszlopa egyaránt emelkedő sorba rendezve mutatja az eredményeket, így az *SS* értékei jól látható módon különíthetőek el a *DS* adataitól.

	GSM_F_001	GSM_F_002	GSM_F_003	GSM_F_004	GSM_F_005	GSM_F_006	GSM_F_007	GSM_F_008	GSM_F_009	GSM_F_010	GSM_F_011	GSM_F_012
MIC_F_001	3,8894894	-8,188838	-6,404308	-7,504419	-9,13741	-4,305273	-7,772248	-7,233298	-4,94654	-9,455144	-7,27875	-8,39539
MIC_F_002	-1,941102	5,0590302	-2,893701	-2,08267	-1,365276	-1,780133	-0,981464	0,6605574	-3,030735	-2,957029	-2,320178	-2,932884
MIC_F_003	-0,914121	-4,234293	4,9716365	-2,104439	-3,244066	-2,999837	-2,632228	-4,96184	-5,0665	-2,113422	-3,038802	-3,787259
MIC_F_004	-2,521695	-2,104469	-2,669408	4,0707419	-0,568945	-2,490925	-1,041065	-2,151039	-2,421919	-3,495229	-2,414598	-3,37573
MIC_F_005	-1,71472	-2,472524	-2,84716	-2,433587	1,6126291	-2,057142	-2,328566	-3,361143	-1,383492	0,8471985	-3,065756	-0,479022
MIC_F_006	-1,101845	-0,956672	0,8355022	-2,933733	-2,73073	1,7795604	-1,259907	-1,655297	-0,903596	-0,077421	-2,32727	-0,659066
MIC_F_007	-0,530777	-0,477965	-0,5919	-2,706162	-1,855861	0,9230963	4,4930429	-1,386486	-2,445868	0,3516481	-1,468388	-1,389002
MIC_F_008	-2,88746	-1,392678	-2,653148	-3,0169	-3,963893	-2,074154	-2,973053	5,6165904	-2,671768	-2,866815	-2,005771	-2,966912
MIC_F_009	-0,467795	-2,347728	-2,51172	-3,42797	-1,51391	-2,575618	-1,76014	-3,325982	3,4826461	-3,603775	-3,069445	-3,911537
MIC_F_010	-2,413283	-2,113661	-0,437073	-1,751477	-2,053038	-1,736337	-2,228704	-2,964136	-2,328125	3,0114221	-2,577815	-2,644377
MIC_F_011	-1,446306	-2,214704	-1,506418	-0,79294	-0,522184	-1,281596	-1,258904	-2,292001	-2,047616	-2,478049	1,8487724	-1,800614
MIC_F_012	-0,668222	-1,701535	-2,36222	-2,064566	-1,829699	-1,045754	-1,515759	-2,162917	-2,371695	-0,216522	-2,735535	3,0663532

5. ábra: Azonosításimátrix-részlet. Forrás: a szerző szerkesztése

Megjegyzendő, hogy egyes rendszerek képesek a teljes azonosítási mátrixot előállítani (például Vocalise), illetve ezek közül egyes típusok (például Batvox) a táblázatot csak részenként képesek generálni – kisebb számítási kapacitásuk miatt –, így ezekből egyenként szükséges a teljes mátrixot létrehozni. Egy táblázatkezelő programban célszerűen alkalmazott feltételes formázás eszköztárral megjelenített azonosítási mátrix további előnye, hogy a több, mint 18 ezer darab valószínűségi adat eltérései, az *SS*- és *DS*-eredmények struktúrája megjeleníti a téves elfogadások vagy a hibás elutasítások előfordulását és mértékét.

A biometrikus beszélőazonosító rendszerek – ahogy a fentiekben ismertettük – az eredményeket nem bináris formában (azonos-különböző), hanem numerikus módon fejezik ki, amely adatok a beszélők azonosságának valószínűségét mutatják meg. A hibaarány értelmezéséhez négy alapfogalmat kell bevezetnünk.

Küszöbszint: az a valószínűségi érték, amely alatt különbözőnek, illetve amely felett azonosnak valószínűsítjük a beszélőket, akiktől a hangminta származik.

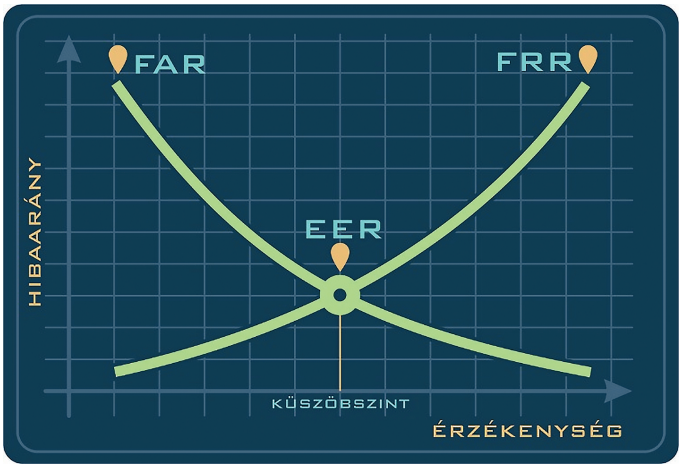
False Accept Rate (FAR): téves elfogadás aránya, amikor különböző személyek valószínűségi értéke a küszöbszint feletti adatot vesz fel, így a beszélőket – hibásan – azonos személynek minősíthetjük.

False Reject Rate (FRR): téves elutasítás aránya, amikor azonos személyek valószínűségi értéke a küszöbszint alatti adatot vesz fel, így a beszélőket – hibásan – különböző személynek minősíthetjük.

Equal Error Rate (EER): az a pont, ahol az FAR és az FRR megegyezik.

Az LR = 1 (LLR = 0) középértékek elméletiek abban az értelemben, hogy a gyakorlatban küszöbszintként más értéket határozunk meg. Ez abból adódik, hogy minden automatikus azonosító rendszer hibaaránnyal dolgozik, így előfordul, hogy különböző

személyeknél az LR = 1 értéknél magasabb, vagy azonos személyeknél alacsonyabb valószínűségi adatot kapunk. A biometrikus beszélőazonosító rendszerekkel meglévő több mint 10 éves gyakorlati tapasztalataim és kutatásaim azt mutatják, hogy a téves elfogadás vagy elutasítás döntő részben akkor következik be, ha a vizsgálati anyag minősége vagy a nettó beszédhossz alacsony (de még megfelel a kritériumoknak). A beszédérthetőséget csökkentő jelenségeket (zaj, torzítás, egyéb, például alacsony dinamika, jelszint) nem tartalmazó hanganyagokon, amelyeken a nettó beszédhossz legalább több percnyi hosszúságú és a beszélő érthetően, a természetes humán beszédprodukciónak megfelelően kommunikál, csak elvétve fordul elő a két hiba valamelyike. Ráadásul a hibás előfordulások esetében a valószínűségi érték a téves elfogadás és az elutasítás esetén is a nulla LR-értékhez közelít, maximum 2 nagyságrendű szórással. A hibaarányok grafikus ábrázolása és az EER értelmezése a 6. ábrán látható.



6. ábra: A különböző hibaarányok értelmezése. Forrás: a szerző szerkesztése

A grafikonról leolvasható, hogy ha növeljük a küszöbszintet, csökken a téves elfogadás aránya, hiszen minél nagyobb a valószínűségi eredmény (és a küszöbszint), annál biztosabb, hogy a rendszer helyesen azonosította az egyező személyeket. Ugyanígy, ha csökkentjük a küszöbszintet, a téves elutasítás aránya is csökken, mert az eredmények alapján egyre több beszélőt fogunk helyesen különbözőnek minősíteni. Mindezekből következik, hogy a küszöbszint módosításával az FAR és az FRR egymással ellentétes irányban változik, így a két hibaarány közlése önmagában nem ad teljes képet a rendszer pontosságáról, ehhez az EER meghatározása szükséges.

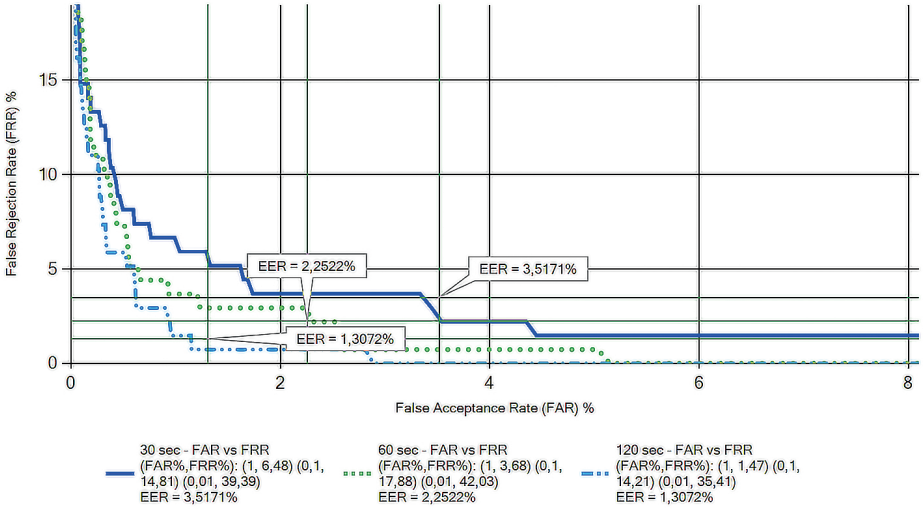
4. Mérési eredmények

A biometrikus rendszerek performanciájának meghatározásához 136 női és férfi beszélő két különböző időpontban és csatornán rögzített, spontán beszélgetést

tartalmazó hangmintáit használtam fel. A mintákat 30, 60 és 120 másodperces hosszúságúakra vágtam azért, hogy az eredmények a beszédhossz függvényében is elemezhetők legyenek.

Első lépésként az FAR (X-tengely) és az FRR (Y-tengely) hibaarányok együttes ábrázolását mutató Detection Error Trade-off (DET) görbéket vettem fel, amelyekről leolvasható az EER értéke. A DET-grafikonon a nagyobb pontosságú (kisebb hibaaránnyal dolgozó) rendszer görbéje a zéró ponthoz közelít, ettől távolodva a pontosság csökken. A pontokkal és egyenesekkel ábrázolt görbe a 120, a csak pontokkal megjelenített a 60 és a folytonos egyenes a 30 másodperces hangfelvételek összevetésének eredményeit ábrázolja. Az EER-eredmények a következőek:

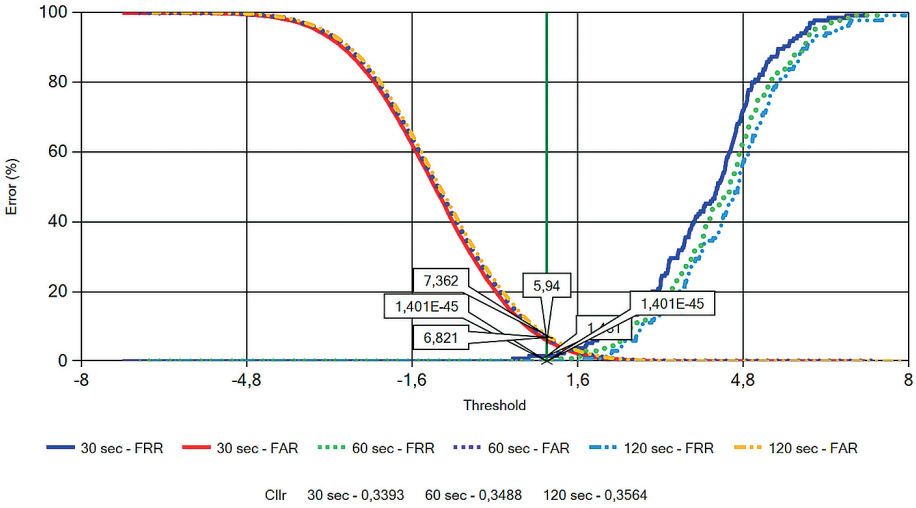
30 másodperces hangfelvételek: 3,5171%,
60 másodperces hangfelvételek: 2,2522%,
120 másodperces hangfelvételek: 1,3072%.



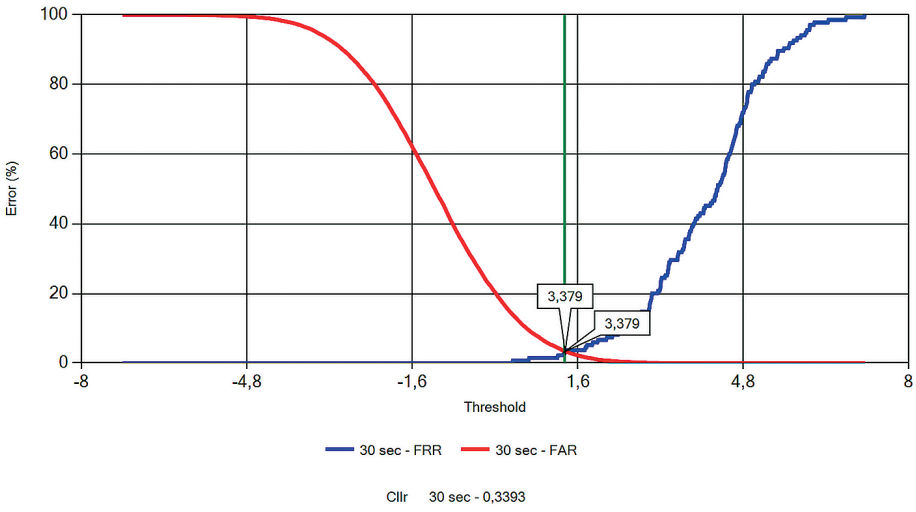
7. ábra: DET-görbék különböző hosszúságú felvételek esetén. Forrás: a szerző szerkesztése

A DET-grafikon jól vizualizálja az egyező beszélők különféle módon módosított (hossz, jelszint, zajszint stb.) felvételeinek, vagy eltérő biometrikus azonosító rendszerek összehasonlítási eredményeit a hibaarány függvényében. Ugyanakkor az EER meghatározása ez esetben nem pontos, mivel fix rögzítésű FAR-értékeknél értelmezi az FRR hibaarányt. A DET-grafikonon az FAR három értéken van rögzítve, ezek: 1, 0,1, 0,01. A 30 másodperces felvételeknél így, amikor az FAR = 1, akkor az FRR = 6,48, ha az FAR = 0,1, az FRR = 14,81 és az FAR = 0,01 esetén az FRR = 39,39. A DET-görbe kiválóan használható különböző típusú rendszerek összehasonlítására, vagy jelen tanulmány eredményeinek elemzésére, de nem ad teljes képet.

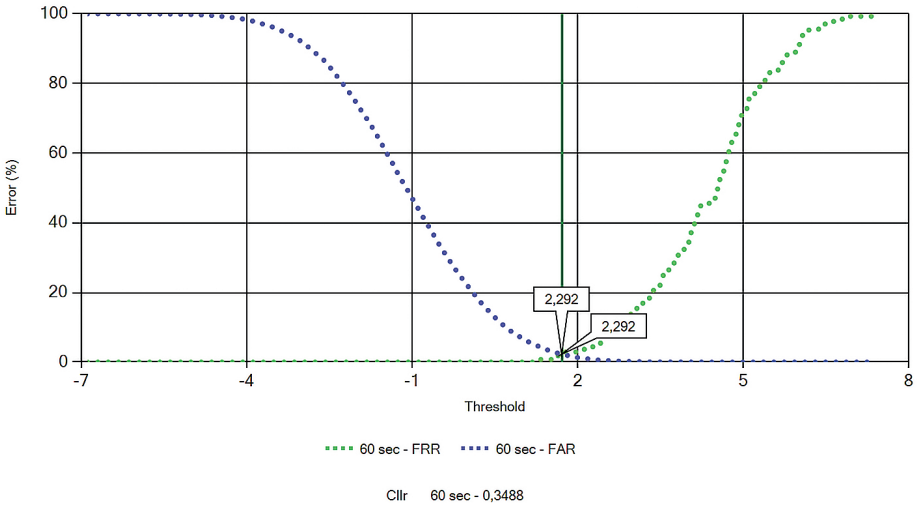
Az EER precíz meghatározásához az FAR és FRR hibaarányok görbéinek felvétele (és folytatólagos ábrázolása) szükséges, amelyek metszéspontja megmutatja a pontos értéket, amely a következő ábrákon látható, sorrendben a 30, 60 és 120 másodpercnyi hosszú fájlok esetében.



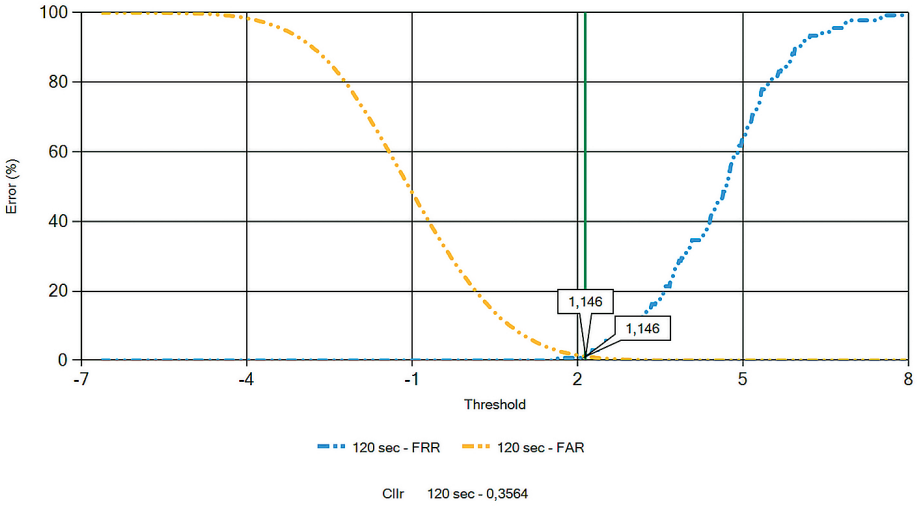
8. ábra: EER összesített, pontos adatai 30, 60 és 120 másodperces felvételek esetén. Forrás: a szerző szerkesztése



9. ábra: EER pontos adatai 30 másodperces felvételek esetén. Forrás: a szerző szerkesztése

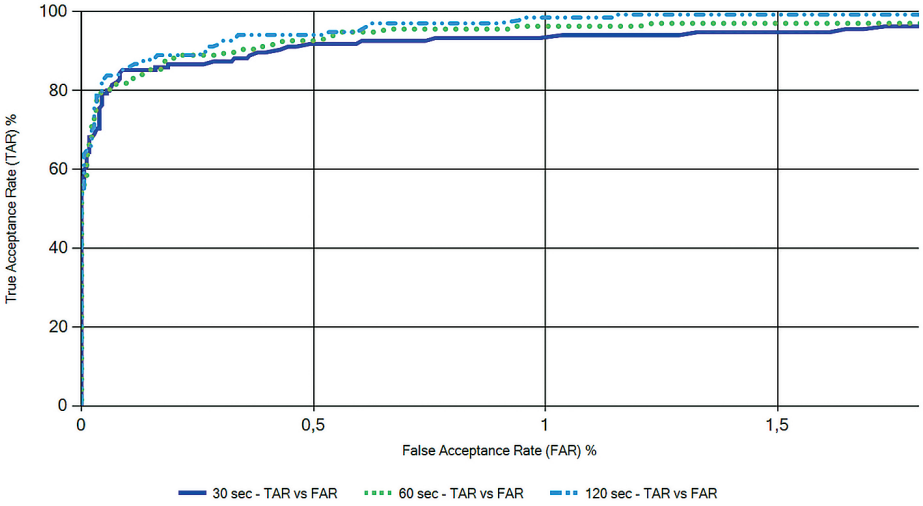


10. ábra: EER pontos adatai 60 másodperces felvételek esetén. Forrás: a szerző szerkesztése



11. ábra: EER pontos adatai 120 másodperces felvételek esetén. Forrás: a szerző szerkesztése

A performancia meghatározásához további információkat szolgáltat a 12. ábrán látható Receiver Operating Charasteristic (ROC) Plot ábrázolás, amely a téves elfogadás és a helyes elfogadás arányát (*True Acceptance Rate* – TAR) ábrázolja a koordinátarendszerben.



12. ábra: ROC-görbe 30, 60 és 120 másodperces felvételek esetén. Forrás: a szerző szerkesztése

Itt az X-tengelyen ábrázolt FAR lesz a vivő *(Receiver)* adatsor – amely tipikusan logaritmikus, így skálázása különbözik az Y-tengelyétől –, és amelynek értékeinél a helyes elfogadás arányát (TAR) ábrázoljuk. Minél közelebb van a ROC-görbe a zéróhoz, annál pontosabbnak minősíthető a mérési menet vagy az azonosító rendszer (amennyiben ezek összehasonlítása a cél).

A fenti méréseket a Vocalise rendszerrel végeztem el, amely nem LR valószínűségi értékkel, hanem Score pontszámmal mutatja az összevetett beszélők azonosságának valószínűségét. A performancia meghatározásához használt Biometrics szoftver alkalmas Score és LR-adatok elemzésére egyaránt, azonban figyelembe kell venni, hogy a *Cost Log Likelihood Ratio* (Cllr) csak LR-eredmények esetében ad pontos értéket. A Cllr értéke a rendszer pontosságát mutatja meg, általánosan elfogadott, hogy egy jól működő szoftver esetében a Cllr < 0,2.[13] Látható, hogy a fentiekben ez a kritérium nem teljesül, mivel nem LR-adatokkal dolgoztunk, ugyanakkor a Cllr-min egy alkalmazható mérőszám a diszkriminatív erő jelzésére.[14] Ez a jellemző megmutatja, hogy a rendszer a két hipotézis (H0-azonosság, H1-különbözőség) között milyen erővel tud különbséget tenni. Alacsony diszkriminatív erő esetén például 1:N összehasonlításban az első helyen álló Same Source értéke és a többi Different Source adat között kicsi a különbség. Minél kisebb a Cllr-min, annál nagyobb diszkriminatív erővel

[13] Daniel Ramos – Rudolf Haraksim – Didier Meuwly: Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data in Brief,* 10. (2017), 2. 75–92.

[14] Massimo Tistarelli – Christophe Champod: *Handbook of biometrics for forensic science*. Cham, Springer, 2017.

rendelkezik a rendszer. A következő táblázatban a fenti mérés további eredményeit találjuk.

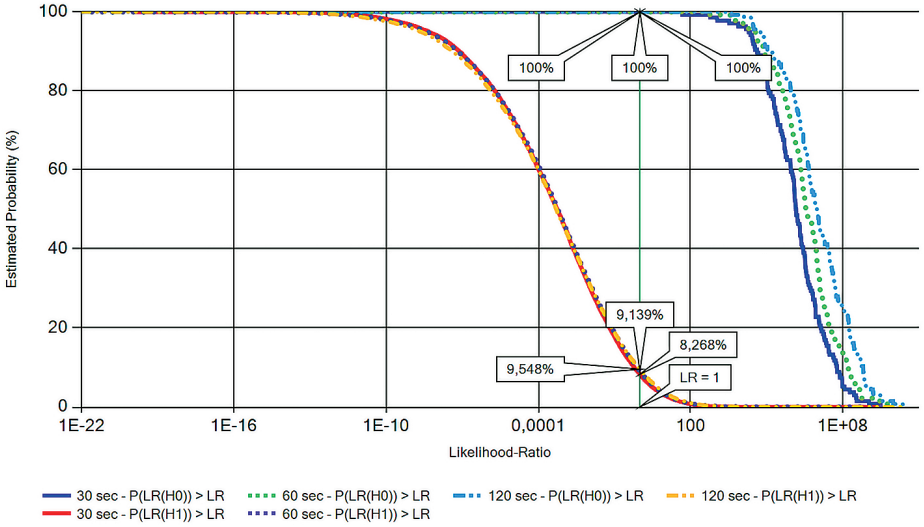
1. táblázat: A Vocalise szoftver eredményei alapján számított jellemzők különböző hosszúságú felvételek esetében. Forrás: a szerző szerkesztése

	H0 átlag	**H1 átlag**	**H0 szórás**	**H1 szórás**	**Cllr-min**
30 sec	4,086518	-1,170386	1,245846	1,388363	0,090828
60 sec	4,37253	-1,091934	1,205438	1,389762	0,056511
120 sec	4,578468	-1,043887	1,198233	1,395037	0,037457

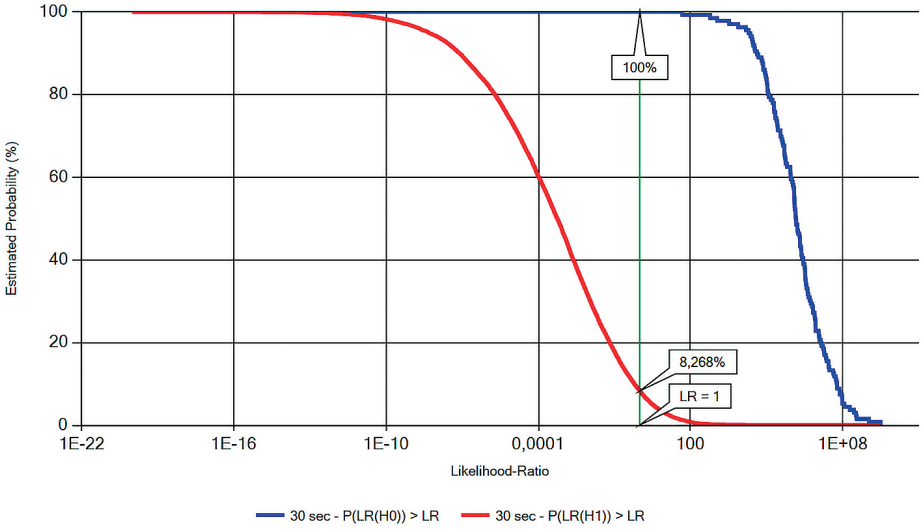
A táblázatban szereplő H0 és H1 hipotézisek Score átlaga közötti különbségek nem mondhatóak jelentősnek, ráadásul a három különböző hosszúságú vizsgálati anyagok esetén az értékek nagyon közel vannak egymáshoz. Ez azt jelenti, hogy a rendszer átlagos eredményeire nincs nagy hatással a vizsgálati anyagok hossza (amennyiben azok 30 és 120 másodperc között helyezkednek el), ugyanakkor ez az átlag nem nyújt részletes képet arra vonatkozóan, hogy egy-egy Score érték önmagában az azonosságot vagy a beszélők különbözőségét valószínűsíti-e. Ezt támasztják alá a hipotézisenkénti szórásértékek,[15] amelyek szintén egymáshoz közel helyezkednek el. A táblázat Cllr-min értékei ugyanakkor megmutatják, hogy minél hosszabb a vizsgálati anyag, annál jobb a rendszer diszkriminatív képessége, így a mérési menetenként 18 496 db Score adat áttekintése nélkül mondhatjuk, hogy a különbség a H0 és H1 hipotézisek adatai között jelentősnek mondható.

A 13. ábrán a Phonexia azonosító rendszer LR-értékei alapján készült Tippet Plot grafikonokat láthatjuk különböző hosszúságú felvételek összevetése esetében. A Phonexia szoftvere mély neurális hálózatú biometrikus motorral dolgozik, LLR-adatokkal is reprezentálja a beszélők azonosságának valószínűségét (Az LLR-ból egyszerű hatványozással képeztem az LR-adatokat). A Tippet Plot egy kumulatív valószínűségi eloszlás diagram, amely a H0 és H1 hipotézisek eloszlását ábrázolja meghatározott (itt LR = 1) értéknél. A két hipotézisgörbe közötti távolság a rendszer teljesítményét jeleníti meg, látható, hogy mindhárom hossznál a H0 100%, ez azt jelenti, hogy minden Same Source adat nagyobb volt, mint LR = 1, tehát a rendszer helyesen azonosította be az egyező beszélőket. Ugyanakkor a H1 görbék az LR = 1 érték egyenesét nem nullpontban metszi, ez pedig azt jelenti, hogy a jelölt százalékértékekben előfordult LR = 1-nél nagyobb adat különböző beszélők esetében, amely tény a téves elfogadást reprezentálja. Megfigyelhető a diagramon, hogy a téves elfogadások adatai alacsonyak, és a zöld egyenestől kis távolságra már eléri a zérót mindhárom H1 görbe, így a téves adatok közel helyezkednek el az LR = 1 értékhez. A 14–16. ábrákon a különböző hosszúságú vizsgálati anyagok alapján elkészített Tippet Plot görbéket láthatjuk hossz szerint egyenkénti ábrázolásban.

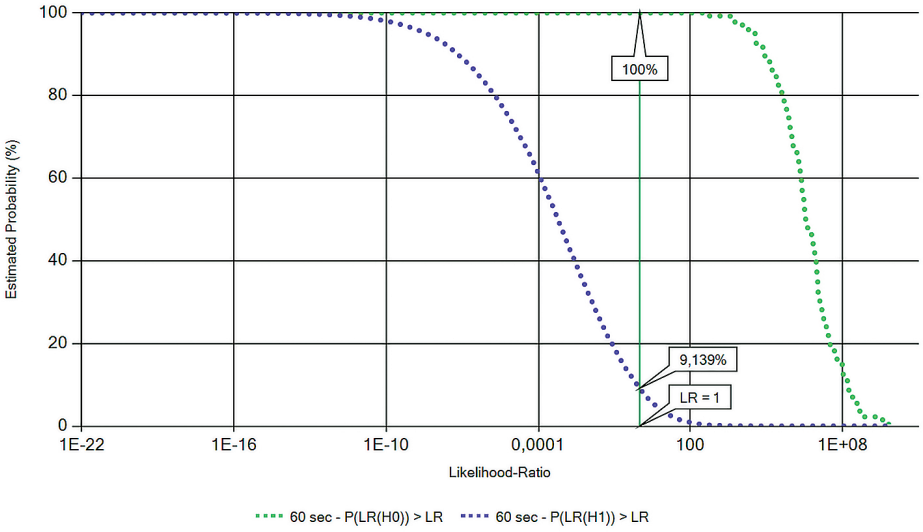
[15] Szórás: az értékek átlagos eltérése az átlagtól.



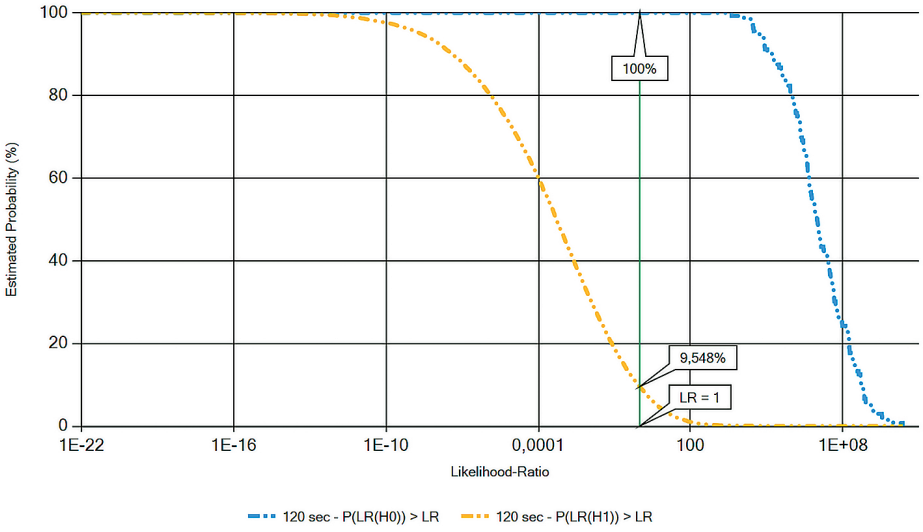
13. ábra: Tippet Plot adatai 30, 60 és 120 másodperces felvételek esetén. Forrás: a szerző szerkesztése



14. ábra: Tippet Plot adatai 30 másodperces felvételek esetén. Forrás: a szerző szerkesztése



15. ábra: Tippet Plot adatai 60 másodperces felvételek esetén. Forrás: a szerző szerkesztése



16. ábra: Tippet Plot adatai 120 másodperces felvételek esetén. Forrás: a szerző szerkesztése

A Tippet Plot görbéje láthatóvá teszi a H0 hipotézis LR-értékeinek nagyságrend szerinti megoszlását (jobb oldali görbék): minél nagyobb a hossz, annál nagyobb a magasabb LR-értékek előfordulásának aránya. Az LR-eredmények alapján számított jellemzők a 2. táblázatban láthatók.

2. táblázat: A Phonexia szoftver eredményei alapján számított jellemzők különböző hosszúságú felvételek esetében. Forrás: a szerző szerkesztése

	H0 átlag	**H1 átlag**	**H0 szórás**	**H1 szórás**	**Cllr**
30 sec	47 820 210	14,15365	277 158 500	840,8036	0,15749
60 sec	132 472 500	16,39306	681 993 900	872,3867	0,17407
120 sec	405 481 400	37,13468	2 159 151 000	2 655,7100	0,18677

A 2. táblázatban láthatóan már más az adatok struktúrája a Score eredmények alapján számítotthoz képest, amelyet az 1. táblázatban láthattunk. Itt a H0 átlag közel egy nagyságrenddel több a 2 perces felvételek esetén a fél percesekhez viszonyítva, amit a H0 szórása is jól demonstrál. Nincs ilyen nagyságú különbség a H1 adatai esetében, azonban itt is látható, hogy jelentősen nő a szórás és az átlag is a hanganyag hosszának növekedésével. Ugyanakkor a Cllr értéke között nincs nagy különbség, ami azt mutatja, hogy a rendszer „biztos kézzel" (jelentős diszkriminatív erővel) képes az azonos beszélők eredményeit a különbözőktől elválasztani.

5. Konklúzió

A biometrikus azonosító rendszerek teljesítményének kifejezésére sokszor csak a hibaarányt jelenítik meg a szűk szakértői közösségen kívül. Ez leegyszerűsíti a hangbiometria technológiájával szemben támasztott minőségi követelményeket, és semmiképp sem fejezi ki a performancia széleskörűen értelmezett jellemzőit. Az FAR-, FRR-, EER-hibaarányok informatívak és jól értelmezhetők, ugyanakkor ezek mellett szükséges vizsgálni az átlag, szórás, Cllr, Cllr-min eredményeket. A hibaarányok grafikus ábrázolása szintén jól demonstrálja a téves elfogadás/elutasítás eloszlását, ugyanakkor szükséges ezek mellett a ROC- és a Tippet Plot-görbék felvétele is.

A tanulmányban a 136 beszélő hangmintáit felhasználva megállapítható, hogy az EER-hibaarány a vizsgálati anyag hosszával fordítottan arányos. Ez további kutatásokat tesz szükségessé annak meghatározásához, hogy a magyar nyelven beszélők hangfelvételei vonatkozásában mi a minimális és maximális hossz, ami egy összehasonlító vizsgálat esetén elfogadható. Az átlag és a szórásadatok szintén visszatükrözték azt, hogy hosszabb hanganyagok esetén a biometrikus rendszer pontosabb eredményeket szolgáltat, azonban a Score és LR-adatok matematikai-statisztikai eszközökkel történő elemzése még jelentős kutatási potenciált hordoz magában. A mérési eredményekben ellentmondás látszik a Cllr adatokra vonatkozóan. Általános szabály alapján minél kisebb a Cllr, annál nagyobb a diszkriminatív erő, ugyanakkor a Phonexia szoftver esetében a hanganyagok hossza és a Cllr egyenes arányosságot mutatott. Ez rávilágít arra, hogy a beszédkutatás, a beszélő személy azonosításának módszertana még számos izgalmas kutatás alapjául szolgál a jövőben.

A cikk az Innovációs és Technológiai Minisztérium Kooperatív Doktori Program Doktori Hallgatói Ösztöndíj Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

IRODALOMJEGYZÉK

Adam, Craig: *Mathematics and statistics of forensic science.* Chichester, Wiley-Blackwell, 2010.

Beigi, Homayoon: *Fundamentals of speaker recognition.* London, Springer, 2011. Online: https://doi.org/10.1007/978-0-387-77592-0

Jain, Anil K. – Arun A. Ross – Karthik Nandakumar: *Introduction to biometrics.* London, Springer, 2011. Online: https://doi.org/10.1007/978-0-387-77326-1

Kamath, Uday – John Liu – James Whitaker: *Deep learning for NLP and speech recognition.* Cham, Springer, 2019. Online: https://doi.org/10.1007/978-3-030-14596-5

Gósy Mária: *Fonetika, a beszéd tudománya.* Budapest, Osiris, 2004.

Künzel, Hermann J.: Automatic speaker recognition of identical twins. *The International Journal of Speech Language and the Law,* 17. (2010), 2. 251–277. Online: https://doi.org/10.1558/ijsll.v17i2.251

Ramos, Daniel – Juan Maroñas – Alicia Lozano-Diez: *Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems.* Madrid, 2017.

Ramos, Daniel – Rudolf Haraksim – Didier Meuwly: Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data in Brief,* 10. (2017), 2. 75–92. Online: https://doi.org/10.1016/j.dib.2016.11.008

Tistarelli, Massimo – Christophe Champod: *Handbook of biometrics for forensic science.* Cham, Springer, 2017. Online: https://doi.org/10.1007/978-3-319-50673-9

ABSTRACT

Theory and Practice of Comparing the Performance of Biometric Speech Recognition Systems

Attila FEJES

The biometric (automatic) speaker recognition method has been widely used in both domestic and international forensic practice. The methodology has high-speed, excellently automated data processing capabilities and it provides accurate and valid results. Biometric speaker recognition systems give the probability of the identity of those speaking on the compared audio recordings. To determine the performance of a system, an identification matrix should be generated, which contains the probability scores. In my study I describe the process and aspects of the production of matrixes and the data structure. I used audio samples of 136 speakers recorded at various times and with various devices. I created the matrix and the match and non-match scores using the Vocalise biometric identification system of Oxford Wave Research Ltd. and the Phonexia software. I evaluated the results with the Bio-Metrics performance measurement software. The evaluation of the results shows that to determine the

performance, several types of output should be used; it is not sufficient to report the most frequently published Equal Error Rate (EER). Based on the analysis of the approximately forty thousand probability results examined, the given system is able to identify the same speakers reliably and with adequate discriminative power and differentiate among different speakers.

Keywords: *speaker recognition, voice biometrics, Likelihood Ratio (LR), performance, error rates*