Nyelvészeti és számítástechnikai módszerek az igazságügyi nyelvészetben

Hunyadi László, Abari Kálmán, Tóth Enikő

Debreceni Egyetem

Általános és Alkalmazott Nyelvészeti Tanszék

4010, Debrecen, Pf. 24.

hunyadi@llab2.arts.klte.hu
abarik@pmail.arts.klte.hu
teniko@pmail.arts.klte.hu

Absztrakt. Cikkünk az igazságügyi nyelvészet ágazatába sorolható esettanulmány ismertetése, melynek során nyelvészeti és számítástechnikai eszközöket is alkalmazva arra a kérdésre kerestük a választ, vajon igazolható-e az, hogy egy adott digitális módon készült hangfelvételt digitálisan manipuláltak. Az elemzés célja annak kiderítése volt, hogy a hangfelvételen található-e vágásra, megszakításra utaló akusztikai jel. Interdiszciplináris elemzést végeztünk, három, egymástól jól elkülöníthető, szemantikai, kísérleti fonetikai, illetve számítástechnikai szempontból vizsgáltuk a fenti kérdést. Ezen módszerek együttes alkalmazása alapvetően sikeresnek bizonyult az eredeti kérdés megválaszolásában. A javasolt új módszerek a nyelvészet és a számítástudomány egyéb területein is hasznosak lehetnek.

1. Bevezetés

Az igazságügyi nyelvészet az alkalmazott nyelvészet ágazatai közé tartozik, eredményei elsősorban bírósági tárgyalásokon alkalmazhatók. Viszonylag rövid története ellenére egyre nagyobb jelentőséggel bír, elsősorban az új számítástechnikai technológiák megjelenésének köszönhetően. Két szempontból is figyelemreméltó a hangfelvételek készítésének lehetősége. Először, amikor technikai szempontból megvalósíthatóvá vált a hangfelvételek készítése, az igazságügyi nyelvészetben is megjelent a beszélő hangfelvétel alapján való azonosításának problémája. Másodszor, a hangfelvételek készítésének lehetősége együtt jár azok esetleges manipulálásával is. Míg a hagyományos hangszalagokkal való manipulálást viszonylag egyszerű kimutatni (a szalag fizikai károsodása vagy a szalagon lévő, manipulálásra utaló elektronikus jegyek révén), a digitális hangfelvételek esetén ugyanez a feladat igazi kihívást jelent. Mivel a digitális hangfelvételek számjegyek sorozataiként realizálódnak, feltehetjük, hogy a hangfelvétel manipulálása egyszerűen a számjegyek sorozatainak módosítását jelenti. A digitális hangszerkesztőkkel ez meg is valósítható,

viszont nyitva marad az a kérdés, vajon kimutatható-e a hangfájlok effajta módosítása.

A következőkben egy valós eseten alapuló vizsgálatot mutatunk be, amelynek célja annak megállapítása volt, hogy a rendelkezésünkre bocsátott digitális hanganyagot manipulálták-e.

2. A probléma

Az alábbiakban bemutatott vizsgálat nyelvészeti és számítástechnikai eszközök alkalmazásával válaszolja meg a következő kérdést: igazolható-e, hogy egy digitális módon készült hangfelvételt digitális módon manipuláltak. A kérdésfelvetés és annak vizsgálata valós eseten alapul.[1]

A hatóságoktól egy audio CD-t kaptunk, amely a kérdéses értekezlet jegyzőkönyvét tartalmazta wav formátumban. Továbbá rendelkezésünkre bocsátották azt a merevlemezt, amelyre a felvételt eredetileg rögzítették, azonban információik szerint a hangállományt korábban letörölték. Így ki kellett dolgoznunk egy olyan eljárást, amely segítségével megállapítható, hogy a merevlemezről letörölt fájlok egyes részei valamilyen mértékben helyreállíthatók-e.

Ennek megfelelően először megvizsgáltuk, hogy a CD-n lévő hanganyagon találunk-e manipulálásra utaló jeleket, másodszor megpróbáltuk a letörölt hangfájlokat, a lehetőség szerint, azonosítani. A hanganyagot először nyelvészeti elemzésnek vetettük alá, majd figyelembe vettük, hogy digitális formában lévő anyagot elemzünk, ezért elektronikus reprezentációjuk formáját is vizsgáltuk.

3. Módszerek

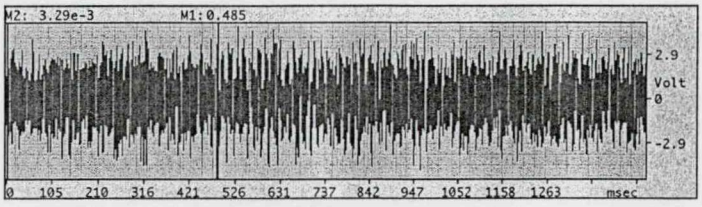
3.1 A lehetséges manipulálására utaló digitális jelek azonosítása

Az igazságügyi nyelvészetben korábban tudomásunk szerint csak analóg felvételek manipulálásával foglalkoztak (ld. Gruber et al, 1993, Gruber – Posa, 1995). Esetünkben azonban a digitális adatreprezentálás természete miatt a már kidolgozott módszert nem alkalmazhattuk. Így először egy előzetes kísérletet végeztünk annak meghatározására, hogy a szándékos manipulálás vajon hagy-e azonosítható jelet az anyagban.

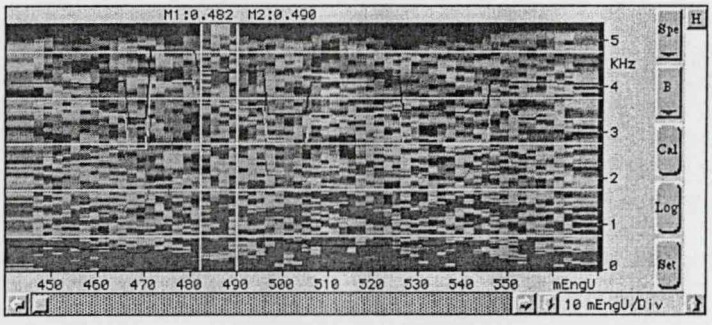
---

[1] Egy kft. elnökségi értekezletéről kétféle jegyzőkönyvet nyújtottak be. Az egyik jegyzőkönyvet a már régóta hatályban lévő ügyvezető igazgató írta alá, a másik jegyzőkönyvet pedig az ezen az ülésen állítólagosan megválasztott igazgató. A korábbi igazgató az értekezletről digitális hangfelvételt készített. Az újonnan megválasztott igazgató viszont amellett érvelt, hogy a hangfelvétel nem hiteles, valószínűleg manipulálták, és a releváns részeket, amelyek az általa benyújtott jegyzőkönyvben találhatók meg, kivágták. Feladatunk tehát az volt, hogy eldöntsük, történt-e manipulálás, melyik jegyzőkönyv hiteles.

Az elemzést a hanghullám spektrografikus reprezentációjára alapoztuk, azt feltételezve, hogy a manipulálás szóhatároknál a legvalószínűbb, különösen a környezeti zajt, de beszédjelet nem tartalmazó szakaszokban (továbbiakban zajszegmensekben) jellemző. Ennek modellezésére eltávolítottuk egy adott zajszegmens egy részét, majd a hátramaradt két részt konkatenáltuk. A szokásos felbontást nagymértékben növelve a spektrogramon láthatóvá vált a konkatenálás helye, egy digitális jel, amely a tekintett hangfájl minden más szegmensétől egyértelműen különbözött. Az 1. ábrán látható a konkatenálás hanghulláma, míg a 2. ábra ugyanazon hanghullám nagy felbontású spektrális képe.



1. ábra Az 1., 3. zajszegmens összefűzése 0,484 ms-nál, miután a 2. zajszegmenst eltávolítottuk közülük



2. ábra A 1. ábrán látható összefűzött hanghullám spektrogramja 0,485 ms körül nagyítva

Ahogy a 2. ábrán látható, a digitális manipulálás helyén egyértelmű nyom látható, amely legalább három szignifikáns jellemzővel bír:

(1) a manipulálás helyén a jelek szabályos függőleges eloszlást mutatnak, vagyis minden frekvenciánál megnövekedett az intenzitás

(2) ez a növekedés szabályos horizontális (idő szerinti) eloszlást is mutat, szimmetrikus struktúra jelenik meg a kb. 0,008 ms szélességű ablakban: az intenzívebb centrális eloszlás mindkét oldalán egy-egy kevésbé intenzív 0,002 ms-os szegmens helyezkedik el

(3) ezen intenzitás-megoszlás maximális értéke jóval meghaladja a felvett szokásos zaj mért maximumát (5 kHz)

Következő feladatunk az volt, hogy megmutassuk, hogy a kísérlet során talált digitális jel (ld. 2. ábra) megbízható eszközt jelent a hangfelvétel digitális

manipulálásának kimutatására. A bizonyítást az alábbi négy lépésben végeztük: igazolni kívántuk, hogy: (1) az elemzendő anyag zajszegmenseinek vágása hasonló nyomot hagy hátra, (2) beltéri körülmények között rögzített zaj tetszőleges szegmensének eltávolítása hasonló nyomot hagy hátra, (3) a vágás nem szoftverfüggő, (4) a vágás nem hardverfüggő.

Ennek érdekében az adott zajból és más zajt tartalmazó felvételekből statisztikailag megfelelő számú vágást hajtottunk végre, majd ezeket megismételtük különböző szoftver és hardver platformokat használva.[2] A tapasztalatokat összegezve megállapítottuk, hogy a feltevésünk helytálló.

3. 2 Szemantikai elemzés

A hanganyagot szemantikai elemzésnek is alávetettük, mely a rendelkezésre álló szöveg tartalmi integritását vizsgálta, abból a feltevésből kiindulva, hogy a csonkítás általában a tartalmi összefüggések sérülésével jár. A szemantikai elemzés során a szövegben olyan részeket kerestünk, amelyek szemantikailag inkoherensek, vagyis vágásra utalhatnak, majd ellenőriztük, vajon tartalmaznak-e ezek a részek manipulálásra utaló digitális jeleket.

A szemantikai elemzés célja tehát annak kiderítése volt, hogy a beszélgetés struktúrájában található-e valamilyen információs hézag, hiányzó összefüggés, így a szöveg strukturális koherenciáját vizsgáltuk (ld. Brown – Yule, 1993). A szöveget abból a szempontból tanulmányoztuk, hogy megállapíthassuk, vajon sérült-e a szövegkohézió, illetve koherencia, vagyis a szöveg kohéziós összefüggéseit vizsgáltuk: a névmások használatát és antecedensükkel való anaforikus kapcsolatukat és a strukturális viszonyokat (pl. elliptikus szerkezetek, szintaktikai ismétlés, igeidők használata és ezek sorozatainak konzekvenciája). Szükségesnek tartottuk a beszélgetés információ-struktúrájának, a beszélők által közvetíteni szándékozott feltételezett üzenetek elemzését is. Azt a vizsgálati szempontot tartottuk szem előtt, hogy tartalmaz-e az anyag olyan hézagokat, amelyek nem rekonstruálhatók a háttér-információ alapján (vagyis a szöveg lehetséges manipulálására utaló jeleket kerestünk).

3.3 A merevlemezen lévő fájlok manipulálásának statisztikai azonosítása

Elemzésre ugyan megkaptuk azt a merevlemezt, amely feltételezhetően az eredeti hangfelvételt tartalmazta, azonban azt később letörölték, így nem állt módunkban a kérdéses értekezlet átiratát egybevetni az eredeti felvétellel. Feltételeztük, hogy az eredeti felvétel néhány töredéke még megtalálható a merevlemezen, annak ellenére, hogy hagyományosan ezeket a fájlokat nem tekintik rekonstruálhatónak. Ezen hangfájlok/hangfájltöredékek azonosítására a nullátmenet (ZC) kiszámítását választottuk, az alábbi megfontolások alapján.

---

2 Kísérleteink nagy részét Macintosh számítógépen, SoundScope/16 felhasználásával végeztük, de egy Intel-alapú PC-n Cool Edit programmal végzett kísérletek ugyanezt az eredményt hozták

A beszéd legfontosabb jellemzője, hogy zöngés, zöngétlen és csendes szegmensek sorozataiból áll. Általában véve a zöngétlen hangoknak magasabb a frekvenciája (így ZC értékük is magasabb), mint a zöngés hangoké, míg a csend oszcillációja (és ZC értéke) gyakorlatilag 0.

Húsz különböző fájltípust vizsgáltunk, azokat, amelyek a legnagyobb valószínűséggel fordulnak elő egy átlagos számítógépen (ld. az 1. táblázatot)[3], és minden fájltípusra húsz fájlt tekintettünk, egyenként 10 Mb körüli mérettel.

1. táblázat Nullátmenetek számának átalaga és szórása

	Átlag	Szórás
avi	123.15	122.84
bmp	96.60	155.21
chm	480.90	12.51
dat	157.30	210.22
dll	333.35	97.11
doc	169.90	110.63
exe	370.20	87.91
gif	510.40	1.50
hlp	277.90	49.05
html	0.70	3.05
hpg	501.85	7.47
mp3	481.20	59.91
pdf	310.70	122.35
rtf	7.30	30.47
sys	381.05	87.30
txt	3.45	10.52
txt2	151.00	19.10
wav	98.10	18.54
wav2	73.20	11.85
xls	197.60	63.04
zip	496.75	11.03

Az összesítő táblázat első két oszlopa az egyes fájltípusok 1 Kb-jára eső nullátmenetek számának átlagát és azok szórását tartalmazza. A hangfájlokra (wav és wav2) ezek az értékek 98,1 és 73,2, a szórás viszonylag kicsi (18,54, valamint 11,85). Az adatok elemzésekor jelentős eltérést találtunk a chm, dll, exe, gif, hlp, jpg, mp3, pdf, sys, txt2 és zip fájltípusok valamint a html, rtf és txt fájlok között, az előbbiek magasabb, az utóbbiak alacsonyabb értékekkel rendelkeznek. Az avi, bmp, dat, doc és xls fájlokhoz tartozó átlagértékek közel voltak a hangfájlok értékeihez, de szórásaikban jelentősen eltértek azoktól.

A fenti kísérleti adatok azt mutatják, hogy egy sztenderd adatrekonstrukciós szempontból nézve reménytelennek tűnő esetben a bitszintű töredékek statisztikai elemzése bizonyos lehetőséget nyújthat az adatok helyreállítására.[4]

3 A fájltípusok között a txt kiterjesztés az ékezet nélküli ASCII kódú fájlokat reprezentálja, míg a txt2 kiterjesztés az ékezeteseket. A wav és a wav2 kiterjesztés 22,050 Hz-es 16 bites magyar beszélt adat reprezentálása, de a wav2 nem tartalmaz zajokat.

4 Következtetések

Az előző részben ismertetett módszereket alkalmaztuk annak érdekében, hogy választ kapjunk az eredeti kérdésre, vajon manipulálták-e a digitális hangfájlt. Összesítve az adott esetre vonatkozó vizsgálati eredményeket arra a következtetésre jutottunk, hogy (1) a hangfelvételben nem találtunk digitális manipulálásra utaló nyomot és (2) feltéve, hogy az általunk azonosított nyom az egyetlen minta, amellyel a digitális manipulálás azonosítható, a fájlon digitális manipulálást nem hajtottak végre.

Cikkünkben egy adott igazságügyi eset kapcsán, nyelvészeti és számítástechnikai eszközöket is alkalmazva arra a kérdésre kerestük a választ, vajon igazolható-e az, hogy egy adott digitális módon készült hangfelvételt digitálisan manipuláltak. Megmutattuk, hogy a hagyományos spektrografikus elemzés során a beszédhang új dimenziókra, szinte mikroszkopikus szegmensekre való szűkítésével egy olyan tartomány tárható fel, amelyben a hangfájl digitális manipulálásának nyomai kimutathatók. Ezt az eljárást hagyományos szemantikai tartalmi elemzéssel kombináltuk, amely során az egyik megközelítés kimenete bemenetként szolgálhatott egy másik eljáráshoz. A hangfájlokra alkalmazott digitális manipulálás jellemző nyomainak azonosítása mellett egy statisztikai eljárást is kidolgoztunk a különböző típusú adatfájlok azonosítására, amivel elősegíthetjük a merevlemezeken lévő már letörölt, header nélküli fájltöredékek azonosítását.

Irodalom

1. Borbándi J., Csáki, E., Nemes L. (szerk): A nyelvész szerepe a kriminalisztikában: 7. Országos Kriminalisztikai Tanácskozás. BM, Budapest, 1977
2. Brown, G., Yule, G.: *Discourse analysis. (Diskurzuselemzés)* Cambridge University Press, Cambridge, 1993
3. Gruber, J. S., Posa F. T.: Voicegram identification evidence. ('Spektrogramok' azonosítása) *American Jurisprudence*, 54., 1995
4. Gruber, J. S., Posa F. T., Pellicano A. J.: Audiotape recordings: evidence, experts and technology. (Magnófelvételek: bizonyíték, szakértők, és technológia) *American Jurisprudence*, 48., 1993
5. Hunyadi, L., Abari, K., Tóth, E.: *Forensic Linguistics: its Contribution to Humanities Computing.* Literary and Linguistic Computing, Vol. 18, No. 1, 2003
6. Kontra, M.: Nyelv és jog. In: Kiefer, F. (szerk) A magyar nyelv kézikönyve. Akadémiai Kiadó, Budapest, 2003
7. Schanze, H. A. European perspective for the PC age. (A PC korszak európai perspektívái) *Literary and linguistic computing*, 5(2), 1980: 171-3

[4] A történethez hozzátartozik, hogy annak ellenére, hogy a merevlemezen lévő fájlok manipulálását kimutató statisztikai eljárás jól működött kísérleti körülmények között, a kidolgozott eljárást nem alkalmazhattuk a kérdéses esetben, ugyanis a merevlemezt, amely vélhetőleg az eredeti hangfelvételt tartalmazta, újraformázták. Ennek következtében nem állt rendelkezésünkre adat, amin a statisztikai elemzést elvégezhettük volna.