Az akusztikus szózsák eljárás korpuszfüggetlenségének vizsgálata

Vetráb Mercedes[1], Gosztolya Gábor[1,2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
Szeged, Árpád tér 2.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
{ vetrabm, ggabor } @ inf.u-szeged.hu

Kivonat Cikkünkben egy jellemzőreprezentációs módszer, az akusztikus szózsák (Bag of Audio Words, BoAW) metódus szélesebb körű használhatóságát elemezzük. A BoAW eljárás lehetővé teszi a változó hosszúságú hangminták fix méretű jellemzővektorokként való kezelését. Ezáltal a különböző hangadatbázisok kezelhetővé és taníthatóvá válnak a hagyományos tanulóalgoritmusokkal is. A BoAW eljárás kezdeti lépésében klaszterközéppontokat (ún. kódszavakat) határozunk meg a keretszintű jellemzővektorok fölött valamilyen felügyelet nélküli módszerrel (pl. k-means klaszterezéssel, vagy akár csak véletlenszerű kiválasztással). Ezt a lépést hagyományosan az adott akusztikus adatbázis tanító halmazán szokás elvégezni. Ez azonban amellett, hogy minden adatbázison új kódszavak kiválasztását teszi szükségessé, így megnyújtva a jellemzőreprezentációk előállításának idejét, akár túlillesztést is okozhat. Jelen tanulmányunkban megvizsgáljuk, hogy mennyire korpuszfüggő az előálló kódszóhalmaz. Kísérleteinkben egy magyar nyelvű érzelemadatbázison mérünk osztályozási eredményeket, miközben a kódszavak kiválasztása vagy egy német nyelvű érzelemadatbázison, vagy egy magyar nyelvű, általános beszédadatbázison történik. Eredményeink szerint mindkét új típusú megközelítéssel elérhető, a korábban említett hagyományos megközelítéssel elérhető osztályozási pontosság, ami megkönnyítheti a BoAW eljárás gyakorlati alkalmazását.
Kulcsszavak: érzelemfelismerés, hangfeldolgozás, akusztikus szózsák reprezentáció

1. Bevezetés

Az emberi beszéd a közlendő szövegen túl egyéb információkat is magában hordoz. Árulkodhat akár a beszélő fizikai vagy emocionális állapotáról is. Napjainkban az automatikus érzelemdetektálás egy folyamatosan fejlődő ágazat. Technikai alkalmazási köre igen széles skálán mozog. Többek közt hasznos az ember-gép interakciók során (az ember kommunikációjának monitorozására) (James és mtsai, 2018), dialógusrendszereknél (Burkhardt és mtsai, 2009), az egészségi állapot

felméréseknél (Hossain és Muhammad, 2015; Norhafizah és mtsai, 2017), valamint a call-centerekben (Vidrascu és Devillers, 2005). Ezen terület fejlesztése és kutatási eredményeinek előrelépése akár hétköznapi rendszereink jelentős fejlődését is eredményezheti.

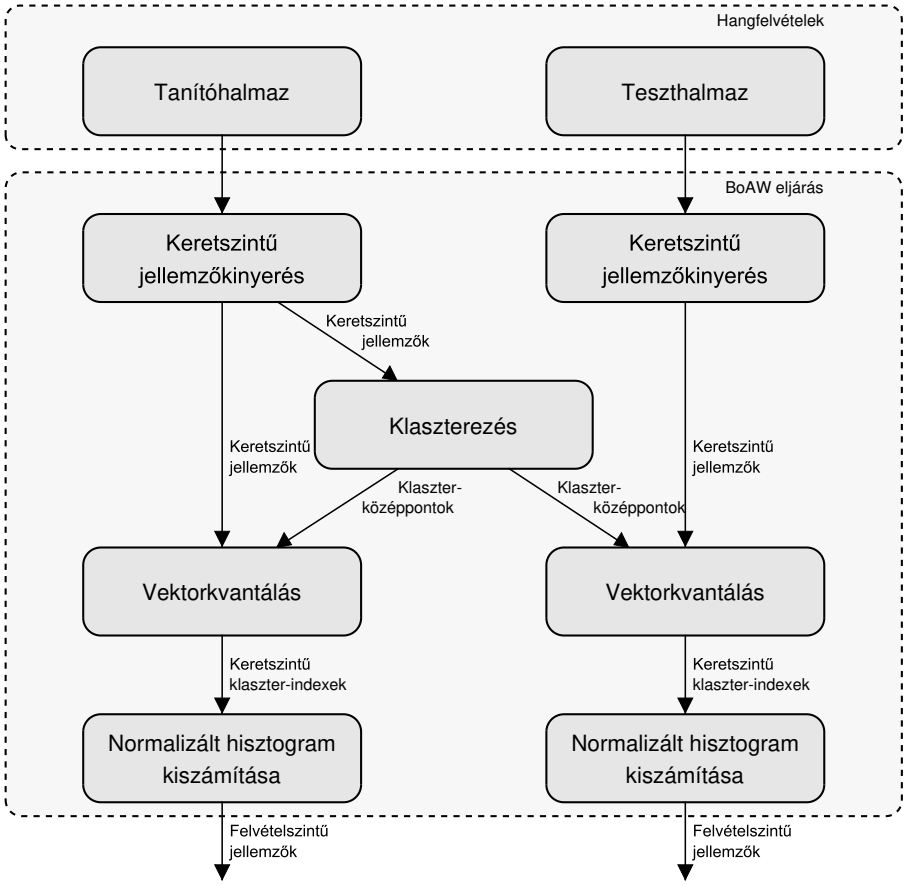
A terület kutatásának kezdete óta több módszert is kidolgoztak arra nézve, hogy a hangfelvételekből milyen módon érdemes jellemzőket kivonni, valamint arra, hogy melyek azok a tanulóalgoritmusok, amik a legoptimálisabb és legeffektívebb eredményeket szolgáltatják egy-egy mintahalmazon. Ezen cikk alapját egy korábbi tanulmányunk adta (Vetráb és Gosztolya, 2019), ahol az akusztikus szózsák (Bag-of-Audio-Words, BoAW (Pancoast és Akbacak, 2012)) technikát és annak sikerességét vizsgáltuk. Pozitív eredményeink alapján újabb kérdések merültek fel, melyeket az alábbi tesztek során igyekeztünk megválaszolni. Korábban a BoAW technika egyik legnagyobb hátrányának a jellemzők előállításához használatos kódhalmaz (codebook) előállításának hosszú idejét tapasztaltuk, melyre megoldást jelentene, ha lehetőség nyílna egy már előre meghatározott codebookot egy másik adatbázisban is felhasználni.

Jelen tanulmányunkban azt vizsgáljuk, hogy más adatbázisból előállított codebookkal lehet-e hasonló vagy jobb eredményeket elérni, mint az eredeti adatbázisból előállított használatával. Mivel eredményeink alapján pozitív választ kaptunk, így megfogalmaztuk következő lényeges felvetésünket, miszerint ha hasonló feladatra készült, de különböző akusztikájú adatbázissal teljesítménybeli javulást értünk el, akkor vajon lehet-e hasonlóan jó eredményeket elérni, egy eltérő feladatra készült adatbázis codebook-jának használatával.

2. Az akusztikus szózsák eljárás

Az általunk használt *akusztikus szózsák* technika, azaz a *bag of audio words* (vagy BoAW) hasonló a szövegfeldolgozásban ismert *bag of words* és a képfeldolgozásban alkalmazott *bag of visual words* (BoVW) módszerekhez. Az 1. ábrán látható, hogy a BOAW módszer egyes fázisaiban végrehajtott műveleteket mind a tanító, mind a teszt halmazon elvégezzük. Első lépésben a tanítóhalmaz hangfelvételeiből kinyerjük az előre meghatározott jellemzőket, melyekből minden kerethez egy-egy jellemzővektor áll elő (keretszintű jellemzők). Ezután a jellemzővektorokból klaszterezés (felügyelet nélküli módszer) segítségével elkészül a kódszavak halmaza (kódhalmaz, *codebook*). A folyamat során megadott számú csoportot hozunk létre, ahol a klaszterek középpontjai lesznek a kódszavak (codewords). A csoportok számát nevezzük a codebook méretének; ez a szózsák eljárás egyik paramétere is.

A következő lépés a vektorkvantálás, mely során az egyes felvételekhez tartozó keretszintű jellemzővektorokat kvantáljuk az előző lépésben generált kódszavaktól vett minimális euklideszi távolságuk alapján. Az eredeti jellemzővektorok helyettesítésre kerülnek a hozzájuk legközelebb lévő kódszó indexével. Végül egy hisztogramot készítünk a kódszavak és hozzájuk sorolt vektorok gyakoriságából. Ebből adódóan a hisztogram mérete megegyezik a codebook méretével, és függetlenné válik az adott hangfelvétel hosszától. Az így előállított vektorhalmaz lesz



1. ábra: Az akusztikus szózsák eljárás működési módja.

a „*bag of audio words*” reprezentáció, ami majd a tanító algoritmusunk inputjául szolgál.

Az 1. ábrán látható, hogy a teszthalmaz esetében a klaszterezés lépése kimarad. Ezt azért tehetjük meg, mert az BoaW eljárás lehetőséget nyújt arra, hogy a tanítóhalmazhoz alkalmazott paraméterbeállításokat és az elkészült codebookot lementsük, majd később ezt használjuk fel (akár más adatbázisokból származó) további hangfelvételek akusztikus szózsák jellemzőreprezentációjának előállításához. Hiszen attól függetlenül, hogy az új mintahalmaz felvételeit nem használtuk fel a klaszterezés során, a keretszintű jellemzővektoraikat még besorolhatjuk az egyes klaszterekbe a kódszavaktól vett távolságuk alapján.

3. Kísérleti körülmények

A következőkben bemutatjuk az elvégzett kísérletek technikai körülményeit: az alkalmazott adatbázisokat és azok szerepét, az osztályozási eljárást és paramétereit, a kiértékelésre használt metrikát, valamint a keretszintű jellemzőkészletet.

3.1. Adatbázisok

A kutatás során minden esetben a magyar érzelemadatbázis tanító és teszthalmazának akusztikus szózsák jellemzőreprezentációján végezzük az osztályozást és kiértékelést. A két további adatbázis a kódszavak előállításában kapott szerepet.

Magyar érzelemadatbázis Az adatbázis 97 magyar anyanyelvű és magyarul beszélő személy hangját tartalmazza (Sztahó és mtsai, 2011). A beszédek televíziós műsorok során lettek felvéve. A szegmensek túlnyomó része érzelmekben gazdag, folyamatos, spontán beszédből lett kivágva. Kisebb részük improvizációs szórakoztató műsorból jön. Az adatbázis összesen 1111 mondatot tartalmaz, melyek egy 831 elemű tanító és 280 elemű teszt halmazra lettek osztva. A címkék négyféle érzelmet definiálnak a beszédekben: Harag, Öröm, Szomorúság és Semleges hangulat. A tanító adatbázis mintáinak eloszlása $\approx 57\%$ semleges, $\approx 6\%$ bánat, $\approx 9\%$ öröm és $\approx 27\%$ harag. A teszt adatbázis mintáinak eloszlása: $\approx 62\%$ semleges, $\approx 4\%$ bánat $\approx 7\%$ öröm és $\approx 27\%$ harag. A tanító adatbázis felvételeinek teljes hossza nagyjából 20 perc, míg a tesztfelvételeké nagyjából 7 perc.

Német érzelemadatbázis Az adatbázis 10 német anyanyelvű és németül beszélő személy hangját tartalmazza. A felvételeket 25 és 35 év közötti színészekkel készítették el. Minden alany 10 különböző mondatot mondott fel, mindegyiket több különböző érzelmi töltettel. A címkék hétféle érzelmet definiálnak a beszédekben: semleges, düh, unalom, undor, félelem, boldogság, szomorúság. Az adatbázis felvételeinek teljes hossza körülbelül 25 perc (Burkhardt és mtsai, 2005).

Magyar híradófelvételek adatbázisa Az adatbázis 8 különböző magyar nyelvű TV-s híradóműsor felvételit tartalmazza. Összesen 28 órányi felvételből áll. A felvételeken hivatásos televíziós műsorvezetők hallhatók, így érzelemdetektálás szempontjából, a hírek felolvasása érzelemmentesnek, azaz semlegesnek tekinthető. Az adatbázisban szereplő híradók felvételeit a felhasználás során először összekevertük, majd több különböző hosszúságú halmazra bontottuk: első 1 óra felvételei, első 2 óra felvételei, első 5 óra felvételei, első 10 óra felvételei. (Grósz és Tóth, 2013)

3.2. Osztályozás

Az osztályozást SVM-ek (Support Vector Machines (Schölkopf és mtsai, 2001)) használatával végeztük, lineáris kernellel, a libSVM implementációt használva (Chang és Lin, 2011). Az algoritmus komplexitás (complexity, C) paraméterét minden futtatás esetén többféle beállítással teszteltük. A lehetséges konfigurációk az alábbi 10 hatványok voltak: $0,00001$; $0,0001$; $0,001$; $0,01$; $0,1$ és 1. Egy adott modell „jóságának” mérésére az UAR metrikát (Unweighted Average

Recall: az adott osztályra helyesen osztályozott példák száma osztva az adott osztályba tartozó példák számával) alkalmaztuk.

A tanító halmazzal való munka során az osztályozó tanulását és kiértékelését 10-szeres keresztvalidálással (10-fold cross-validation, CV) hajtottuk végre. Tehát az aktuális mintahalmazt 10 közel egyenlő részre osztottuk úgy, hogy az azonos beszélőhöz tartozó minták, azonos fold-ba kerüljenek. Ezután minden lehetséges 9 tanító – 1 tesztelő halmaz kombinációra tanítottunk és kiértékeltünk egy SVM modellt. Később, a kereszvalidálás során a tanító halmazra kapott értékek alapján választottuk ki, hogy a teszteket milyen beállításokkal futtassuk.

A teszthalmazra adott predikcióinkat a teljes tanítóhalmazon tanított SVM modellek szolgáltatták.

3.3. Keretszintű jellemzőkészlet

Az akusztikus keretszintű jellemzők megválasztásának alapját a 2013-as INTERSPEECH Számítógépes Paralingvisztikai Versenyen kiadott cikk adta (Schuller és mtsai, 2013). Az ott publikált jellemzőkészlet 65 keretszintű jellemzőt, azaz LLD-t (low level descriptor) tartalmazott (4 darab energián alapuló LLD; 55 spektrális LLD; 6 hangosságon alapuló LLD), valamint ezek első fokú deriváltjait. Kutatásunk során ezen jellemzőket az openSMILE nevű program segítségével számoltuk le. A hangosság alapú leírókat 60 ms-os kerettel (Gaussian ablakfüggvény) és $0,4$ értékű szigmával, a másik két csoportot pedig 25 ms-os kerettel (Hamming ablakfüggvény) számítottuk. A Hamming-ablakokat a megszokott módon, átfedéssel, 10 ms-os eltolással helyeztük el.

3.4. Az akusztikus szózsák eljárás tesztelt paraméterei

Az akusztikus szózsák reprezentáció számítását megvalósító openXBOW (Schmitt és Schuller, 2017) program, a BoAW eljárás több, szabadon állítható paraméterének beállítási lehetőségével rendelkezik. Kísérleteink során teszteltük az adatok előfeldolgozásának, a codebook méretének valamint a kvantáláskor figyelembe vett legközelebbi szomszédok számának hatásait.

Egyik, említett beállítási lehetőségünkkel az elkészítendő codebook méretét adhatjuk meg. Itt határozzuk meg, hogy a klaszterezés során, hány klasztert állítsunk elő, tehát hány kódszót határozzunk meg.

Egy másik szabályozható komponens a hisztogram előállításának módja. Korábbi cikkünkben (Vetráb és Gosztolya, 2019) született megerősítő eredményeink alapján egyértelmű volt számunkra, hogy minden kerethez a legközelebbi klaszter helyett a legközelebbi a db. klasztert rendeljük hozzá, mivel így azonos méretű jellemzővektor mellett pontosabban írhatjuk le az adott felvételt.

Az előző módosításon túl, a kezdeti keretszintű jellemzőkészleten is hajthatunk végre előfeldolgozást. Előfordulhat, hogy az eredeti adatok túlságosan szétszórva helyezkednek el a térben, valamint vannak köztük olyan minták, melyek kiugró értékeikkel fals irányba mozdíthatják a tanulást. Ennek kiküszöbölésére a jellemzővektorokat előfeldolgozásnak vetettük alá.

Jellemző-transzformáció	**a**	**UAR**		**Codebook méret**
		CV	**Teszt**	
Normalizálás	5	58,08%	48,13%	512
	10	57,48%	50,27%	512
Standardizálás	5	55,43%	53,54%	512
	10	56,57%	64,32%	256

1. táblázat. Baseline: a keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció során és a teszthalmazon.

A tesztelt értékek az alábbiként alakultak:

- Codebook méretek: 32, 64, 128, 256, 512, 1 024, 2 048
- Kvantáláskor figyelembe vett legközelebbi szomszédok száma: 5, 10
- Előfeldolgozási opciók: standardizálás, normalizálás

4. Tesztek és eredmények

A következőkben ismertetésre kerül a kísérletek pontos menete, valamint az egyes fázisokban kapott eredmények kiértékelése. Mivel minden hangadatbázisból 2×65 darab jellemzőt vontunk ki, hogy figyelembe vehessük a delta értékeket is, így párhuzamosan két külön codebook is készült az alap és a delta jellemzőkhöz az elemzések során. Ebből adódóan, az itt feltüntetett codebook méreteket 2-vel szorozva kapjuk meg az aktuálisan felhasznált jellemzők számát.

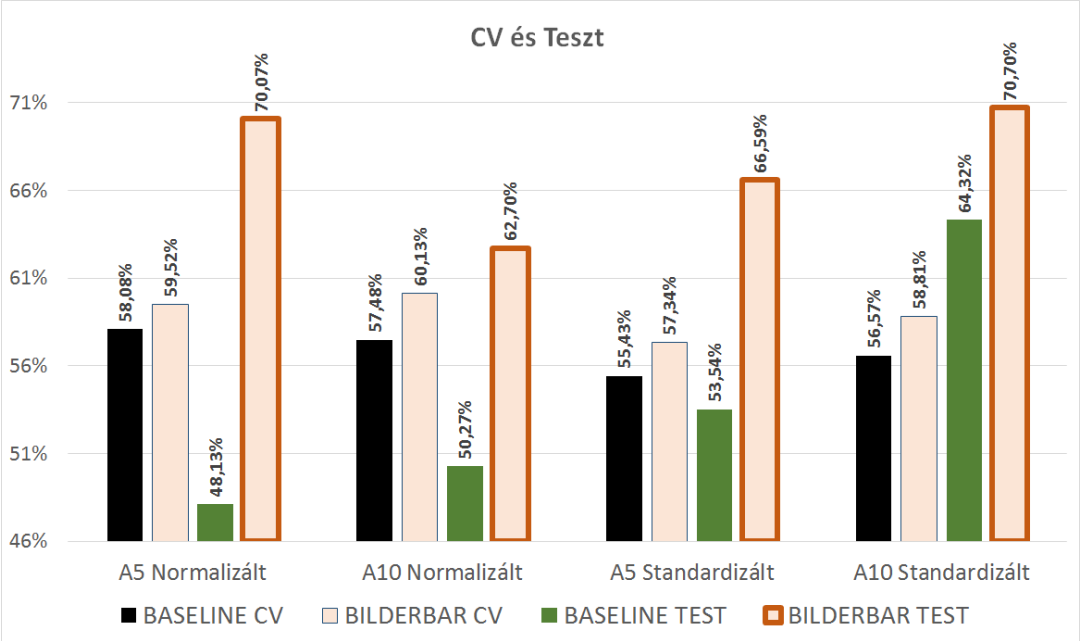
Az egyes tesztesetek közötti különbséget a BoAW reprezentáció előállításához használt különböző codebookok adják. Három esetet vizsgálunk:

- A felhasznált codebook a magyar érzelemadatbázis tanító halmazából készül
- A felhasznált codebook a német érzelemadatbázisból készül
- A felhasznált codebook a magyar általános beszédadatbázisból készül

Első feltevésünk alapját korábbi cikkünk (Vetráb és Gosztolya, 2019) szolgáltatta. Az ott kapott eredmények alapján az akusztikus szózsák jellemzőreprezentációs technika az érzelemdetektálás feladatában versenyképesnek bizonyult. Legnagyobb hátránya a jellemzők előállításához használatos codebook előállításának hosszú ideje volt, így felmerült a kérdés, miszerint egy adott adatbázis jellemzőkészletének kinyeréséhez alkalmasan felhasználható-e másik adatbázisból kapott codebook.

4.1. Baseline

Baseline-ként a magyar nyelvű érzelemorientált felvételeket tartalmazó mintahalmazt használtuk a codebook-ok meghatározásához. A kapott értékek az 1. táblázatban láthatók.



2. ábra: A baseline, valamint az eltérő emocionális adatbázisból származó codebookkal futtatott keresztvalidáció és teszt eredményei

A keresztvalidáció során kapott legjobb eredmény, normalizálással, 5 szomszéd figyelembe vételével, 512 méretű codebookkal és adódott, 58, 08%-ra. A teszt során kapott legjobb eredmény, standardizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 64, 32%-ra. Ezen kívül megfigyelhető, hogy 4-ből 3 alkalommal a tesztadatbázison kapott eredmények alacsonyabbak voltak, mint a keresztvalidálás eredményei.

4.2. Eltérő, emocionális adatbázisból származó codebook

Első kísérletünkben vizsgáltuk, hogy más adatbázisból előállított codebookkal való munka, képes-e hasonló vagy jobb eredményeket produkálni, mint az eredeti adatbázisból előállított codebook. A codebookokat a korábban már leírt, német nyelvű, érzelem orientált felvételeket tartalmazó, általunk *"bilderbar"*-ként hivatkozott adatbázisból készítettük el. Ezután, ezen codebook-okat felhasználva elkészítettük az akusztikus szózsák jellemzőreprezentációt a magyar nyelvű, érzelem orientált adatbázishoz. A tesztek során kapott értékek az 2. táblázatban láthatók.

A keresztvalidáció során kapott legjobb eredmény, normalizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 60, 13%-ra. A teszt során kapott legjobb eredmény, standardizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 70, 70%-ra.

A 2. ábrán látható, hogy mind a négy tesztesetnél szignifikáns javulást értünk el a baseline-hoz képest. Normalizált adat és 10 szomszéd figyelembevételénél,

Jellemző-transzformáció	**a**	**UAR**		**Codebook méret**
		CV	**Teszt**	
Normalizálás	5	59,52%	70,07%	1 024
	10	60,13%	62,70%	256
Standardizálás	5	57,34%	66,59%	128
	10	58,81%	70,70%	256

2. táblázat. BilderbarDB: A keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció és a teszt során.

valamint standardizált adat és 5 szomszéd figyelembevételénél a szükséges codebook méretének csökkenését is tapasztaltuk, melynek hatására a jellemzőreprezentáció előállításához szükséges idő is csökkent. Bár a keresztvalidálás során 4-ből 3 esetben rosszabbul teljesített az eltérő adatbázisú codebookkal dolgozó modell, a tesztek során mégis minden esetben jobb eredményeket kaptunk, mint a baseline. Ez arra enged következtetni, hogy a saját adatbázisból készített codebook túltanulásra hajlamosítja a modellünket, míg az eltérő adatbázisú codebook használata kiegyensúlyozza ezt.

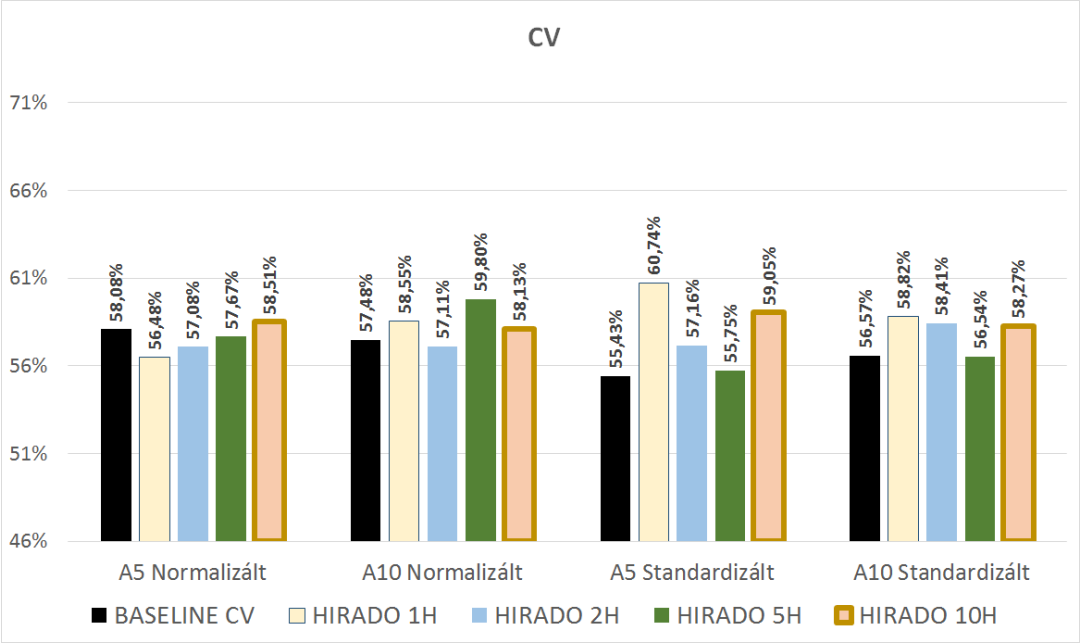
4.3. Eltérő, semleges adatbázisból származó codebook

Az előző tesztek alapján jól látszódott, hogy az eltérő adatbázisból előállított codebook-ok szignifikáns javulást hoztak. Mivel a codebook előállításának fő lépése a felügyelet nélküli klaszterezés, így felmerült a kérdés, hogy vajon befolyásolja-e az osztályozás sikerességét, ha a codebook készítéséhez használt adatbázis eltérő célra készült, mint a codebook-ot felhasználó adatbázis osztályozási feladata?

Az új codebook-okat a korábban már említett, nem érzelemdetektálási célra készült, magyar nyelvű, televíziós híradás felvételeit tartalmazó adatbázis részhalmazaiból állítottuk elő, majd ezek segítségével készítettük el a magyar emocionális adatbázis jellemzőreprezentációját.

Négyféle esetet vizsgáltunk annak érdekében, hogy kiderítsük, vajon a codebook elkészítéséhez használt adatbázis hossza befolyásolja-e az osztályozó teljesítményét: 1 órányi, 2 órányi, 5 órányi és 10 órányi felvételre készítettünk elemzést. A kapott értékek a 4.3. táblázatban láthatók.

A 3. ábrán látható keresztvalidáció során kapott eredményeken semmilyen irányú általános konvergenciát nem tapasztaltunk, így nem tudtunk összefüggést vonni az adatbázis hossza és az osztályozás sikeressége között. Ugyanez mondható el az előfeldolgozási módszer típusáról és a legközelebbi szomszédok számáról is. Minden eredmény 55,75% és 60,74% között mozgott. A legjobb eredményeket adó codebook méretek jelentős része 1 024 és 2 048 volt, ami igen nagy jellemző bázist jelent, ami könnyen eredményezhet túlillesztést, így ilyen szempontból a kisebb codebook-okat igénylő német adatbázissal való munka jobbnak bizonyul.



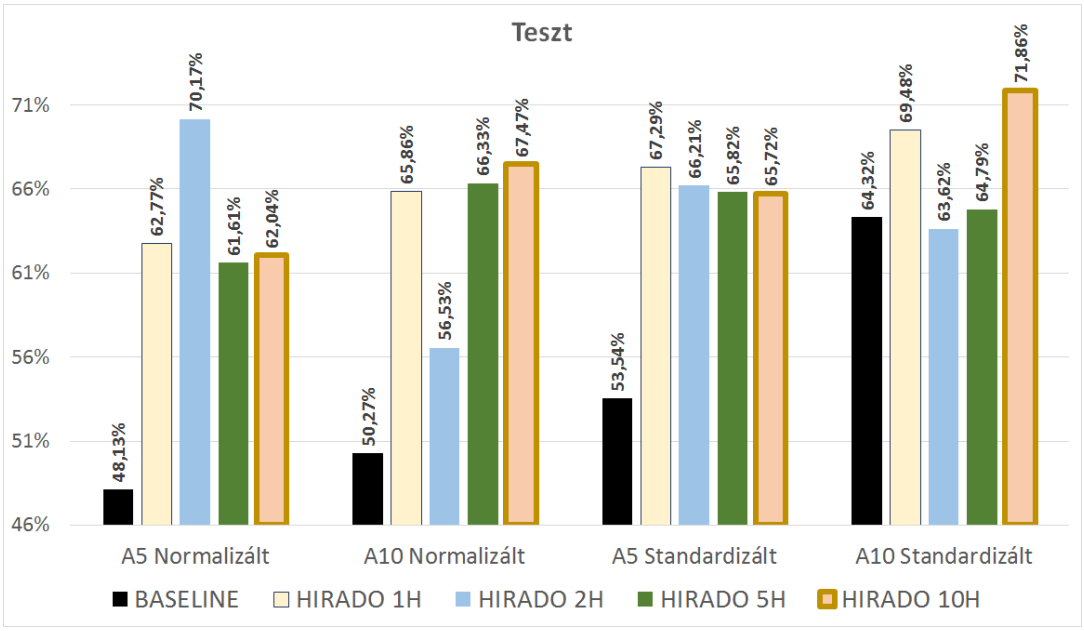
3. ábra: A baseline, valamint az általános magyar adatbázisból származó codebookkal futtatott keresztvalidáció eredményei

Keresztvalidálás során a legjobb eredményt, azaz 60,74%-ot az 1 órás felvételekkel értük el, standardizálással, 5 szomszéd figyelembe vételével és 1 024-es codebook mérettel. Az eredmények a felvételek hosszától függően, szignifikáns eltérést nem mutattak, így ezzel kapcsolatban nem fedeztünk fel hosszútávú összefüggéseket. Ezen kívül nem kaptunk szignifikánsan se jobb se rosszabb eredményt, mint a kifejezetten érzelemdetektálás céljához készített, német adatbázisból előállított codebook felhasználásakor.

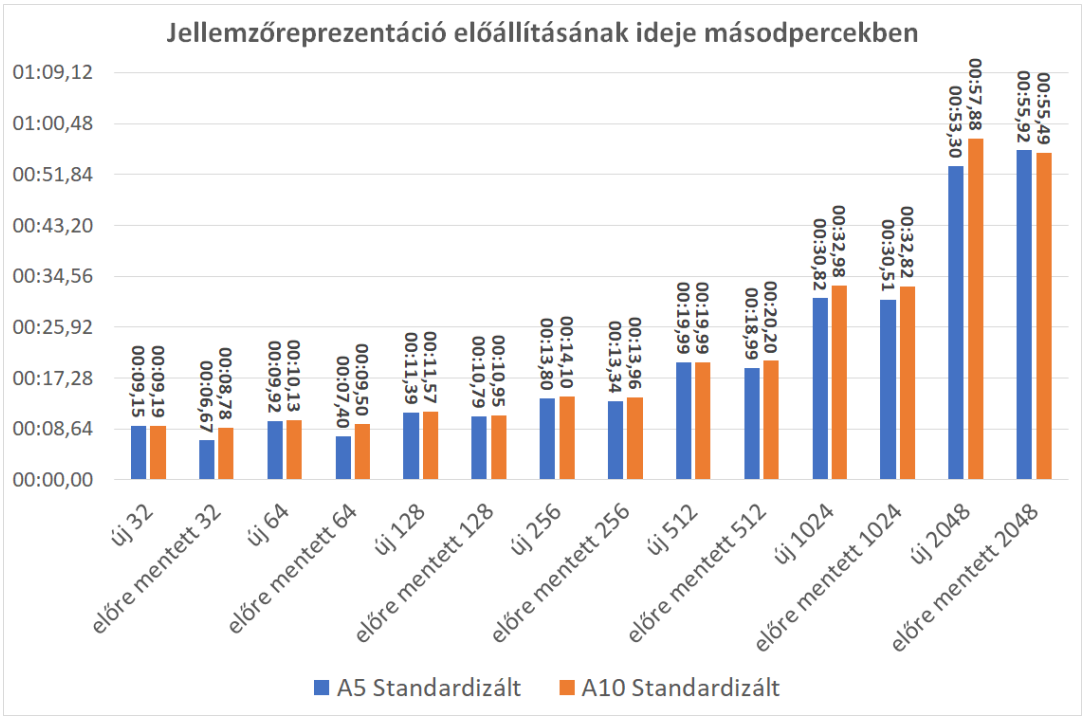
A 4. ábrán feltüntetett, teszthalmazon kapott eredmények nagyon hasonlónak adódtak az előző, német adatbázis segítségével végzett tesztekhez. Itt is elmondható, hogy az eltérő adatbázisból származó codebook használata minden esetben javított a keresztvalidálás eredményein (egy 0,58%-os visszaesést kivéve a 2 órás adatbázis - 10 legközelebbi szomszéd - normalizálás esetben).

Tesztelés során a legjobb eredményt, azaz 71,86%-ot, a 10 órás felvételekkel értük el, standardizálással, 10 szomszéd figyelembe vételével és 1 024-es codebook mérettel. Ám a szükséges codebook méretek növekedésén túl további konzekvenciákat itt sem vonhatunk le.

Mivel eredményeink számításához használt adathalmazunk kisméretű, valamint a cikkünkben szereplő statisztikák nem mutatnak teljesen sztochasztikus viselkedést, így felmerülhet a kérdés, miszerint eredményeink mennyire lehetnek véletlenszerűek? Ez azonban egyértelműen kizárható. Egy korábbi tanulmány (Vetráb és Gosztolya, 2019) ugyan ezen adatbázissal végzett tesztjei minden codebook méretre vonatkozóan tartalmazzák az eredményeket, melyek így átfogóbb képet mutatva egyértelműsítik a teljesítmények közötti korrelációkat.



4. ábra: A baseline, valamint az általános magyar adatbázisból származó codebookkal futtatott teszt eredményei



5. ábra: A BoAW jellemzőreprezentáció előállításához szükséges idők a újonnan létrehozott és már meglévő codebook felhasználásának esetében, codebook méret (x tengely) és a használt beállítások függvényében.

Adatbázis	Jellemző-transzformáció	a	UAR CV	UAR Teszt	Codebook méret
1 órás híradó	Normalizálás	5	56,48%	62,77%	512
		10	58,55%	65,86%	1 024
	Standardizálás	5	60,74%	67,29%	1 024
		10	58,82%	69,48%	1 024
2 órás híradó	Normalizálás	5	57,08%	70,17%	1 024
		10	57,11%	56,53%	32
	Standardizálás	5	57,16%	66,21%	2 048
		10	58,41%	63,62%	2 048
5 órás híradó	Normalizálás	5	57,67%	61,61%	2 048
		10	59,80%	66,33%	1 024
	Standardizálás	5	55,75%	65,82%	128
		10	56,54%	64,79%	2 048
10 órás híradó	Normalizálás	5	58,51%	62,04%	2 048
		10	58,13%	67,47%	1 024
	Standardizálás	5	59,05%	65,72%	1 024
		10	58,27%	71,86%	1 024

3. táblázat. Híradó: A keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció során.

Az 5. ábrán láthatók a jellemzőreprezentációk előállításához szükséges idők. Szinte minden esetben több, mint 1 másodperc javulást értünk el, mikor előre elkészített codebook-okat használtunk. Először csekélynek tűnhet ez az érték, de képzeljünk el egy valós idejű rendszert. Egy számítógépes program vagy akár telefonos applikáció esetén, ahol a felhasználói élmény fontos részét képezi a rendszer válaszadási ideje, és ahol sok esetben tized másodpercekben mérik ezt, ott már ez az 1 másodperc is javítja szoftverünk minőségét.

5. Összegzés

Jelen cikkünkben az akusztikus szózsák *(Bag-of-Audio-Words, BoAW)* jellemzőreprezentációs eljárást egy időben, többféle adatbázison alkalmaztuk, érzelemfelismerési feladatra. A különböző esetek kombinációi miatt számos gépi tanulási modellt kellett tanítanunk a különböző paraméter-kombinációkra, így a tesztek futási ideje fontos szempont volt. Mért eredményeink alapján a bemeneti jellemzőket továbbra is érdemes azonos skálára hoznunk normalizálás vagy standardizálás segítségével, ám továbbra sem bizonyult egyértelműen jobbnak egyik a másiknál. Hasonlóan a legközelebbi szomszédok számában sem véltünk felfedezni korrelációt, így továbbra is mind az 5, mind pedig a 10 szomszéd használata hasonlóan jónak bizonyul. A már említett, korábbi (BoAW módszertant vizs-

gáló és a magyar érzelemadatbázissal dolgozó) cikkhez képest az egyes ajánlott codebook méretek idegen adatbázisokból történő codebook kivonás segítségével csökkenthetők lettek. Ebben a tekintetben az idegen, de azonos típusú osztályozási feladatra készült adatbázisból kinyert codebook-ok jobban teljesítettek, mint a más típusú adatbázis codebook-ai.

Tesztjeink alapján egyértelműen kijelenthető, hogy egy-egy előállított codebook eredményesen használható más adatbázisok jellemzőreprezentációjának előállításakor. Azzal kapcsolatban, hogy célszerű-e hasonló célból készített adatbázisok között átvinni a codebookot, vagy bármely két adatbázis között átjárható-e az út, nem találtunk egyértelmű választ. Eredményeink mindkét esetben hasonló skálán mozogtak, szignifikáns különbséget nem adva, csupán a codebook méretekben tértek el, így ez a terület további kutatásokat igényel.

Annak kapcsán, hogy elért eredményeink által merre haladhatunk tovább a későbbiekben, több lehetőség is felmerül. Figyelemre méltó kérdés, hogy a codebook átvitel milyen adatbázis típusok között alakulhat a legeredményesebben. Található-e ebben erősebb összefüggés. Emellett lehetőségünk van más keretszintű jellemzőkészleteket is letesztelni.

Köszönetnyilvánítás

Jelen kutatás eredményei az „Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein” című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatásával készültek. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal is támogatta (FK 124413). Gosztolya Gábor kutatásait az MTA Bolyai János ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-19-4-SZTE-51) támogatta.

Hivatkozások

Burkhardt, F., van Ballegooy, M., Engelbrecht, K.P., Polzehl, T., Stegmann, J.: Emotion detection in dialog systems: Applications, strategies and challenges. In: ACII. pp. 985–989. Amszterdam, Hollandia (Sep 2009)

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Interspeech. pp. 1517–1520 (2005)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)

Grósz, T., Tóth, L.: A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition. In: TSD. pp. 36–43. Pilsen, Csehország (2013)

Hossain, M.S., Muhammad, G.: Cloud-assisted speech and face recognition framework for health monitoring. Mobile Networks and Applications 20(3), 391–399 (2015)

James, J., Tian, L., Inez Watson, C.: An open source emotional speech corpus for human robot interaction applications. In: Interspeech. pp. 2768–2772. Hyderabad, India (Sep 2018)

Norhafizah, D., Pg, B., Muhammad, H., Lim, T.H., Binti, N.S., Arifin, M.: Detection of real-life emotions in call centers. In: ICIEA. pp. 985–989. Siem Reap, Kambodzsa (June 2017)

Pancoast, S., Akbacak, M.: Bag-of-Audio-Words approach for multimedia event classification. In: Interspeech. pp. 2105–2108. Portland, USA (Sep 2012)

Schmitt, M., Schuller, B.: openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit. Journal of Machine Learning Research 18(96), 1–5 (2017), http://jmlr.org/papers/v18/17-113.html

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. pp. 148–152 (08 2013)

Sztahó, D., Imre, V., Vicsi, K.: Automatic classification of emotions in spontaneous speech. In: COST 2102. pp. 229–239. Budapest (2011)

Vetráb, M., Gosztolya, G.: Érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával. In: MSZNY. pp. 265–274. Szeged (2019)

Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Interspeech. pp. 1841–1844. Lisszabon, Portugália (Sep 2005)