Érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával

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

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
14vini24@gmail.com, ggabor@inf.u-szeged.hu

Kivonat Az érzelmek felismerése a beszédtechnika egy jelenleg is aktívan kutatott területe. A feladaton belül számos probléma fogalmazódott már meg; ezek egyike az egyes hangfelvételek leképezése jellemzőkre. Ennek különlegességét az adja, hogy a hangfelvételek várhatóan eltérő hosszúságúak, míg a következő lépésben alkalmazott osztályozó eljárás fix méretű jellemzővektorokat vár el. Jelen dolgozatban az érzelemfelismerés problémájára egy nemrégiben kifejlesztett jellemzőreprezentációs eljárást, az akusztikus szózsák (bag of audio words, BoAW) módszert alkalmazzuk, mely képes a változó hosszú bemondásokat fix jellemzőtérbe képezni. Kísérleti eredményeink alapján a BoAW eljárás versenyképes osztályozási teljesítményt tesz lehetővé, ugyanakkor a metódus számos paraméterrel rendelkezik, melyeket a megfelelő hatékonyság érdekében pontosan be kell hangolni.
Kulcsszavak: érzelemfelismerés, hangfeldolgozás, akusztikus szózsák reprezentáció

1. Bevezetés

Az emberi hang nem csupán a szöveg közlésének alapjául szolgál, hanem magában hordoz rejtett, mégis az adott beszélőre nézve fontos, fizikai és lelki jellemzőket is. Ezen indirekt hangi kifejeződések egyike a beszélő emocionális állapota. Napjainkban az automatikus érzelemdetektálás egy aktívan kutatott témakör. A gépek által használt érzelem-felismerő és -monitorozó rendszerek jelenleg is fejlődésben vannak. A technika alkalmazási köre elég széles skálán mozog. Többek közt hasznos az ember-gép interakciók során (az ember kommunikációjának monitorozására) [1], dialógusrendszereknél [2], az egészségi állapot felméréseknél [3,4], valamint a call-centerekben [5]. Az érzelemfelismerés fejlődésével jelenleg is létező munkák könnyíthetők meg, valamint a későbbiekben, mindennapjainkba is beszivárgó robotikai és informatikai rendszerek kiegészítéseképp, vagy akár alapjául is szolgálhat.

Ezen 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 az akusztikus szózsák (Bag-of-Audio-Words, BoAW) technikát és annak sikerességét vizsgálja, SVM tanulóalgoritmussal ötvözve. Kísérleteinket egy magyar nyelvű hangadatbázison végeztük; eredményeink azt mutatják, hogy a BoAW eljárás hatékony jellemzőreprezentációt tesz lehetővé érzelemfelismerés esetén is, mert a kapott pontosságmetrika-értékek (relatíve) magasaknak adódtak. Ugyanakkor azt a következtetést is levonhatjuk, hogy az eljárás érzékeny a paraméterbeállításokra, így azokra nagy figyelmet kell fordítani, hogy az osztályozás minősége megfelelően magasan alakuljon.

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 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.

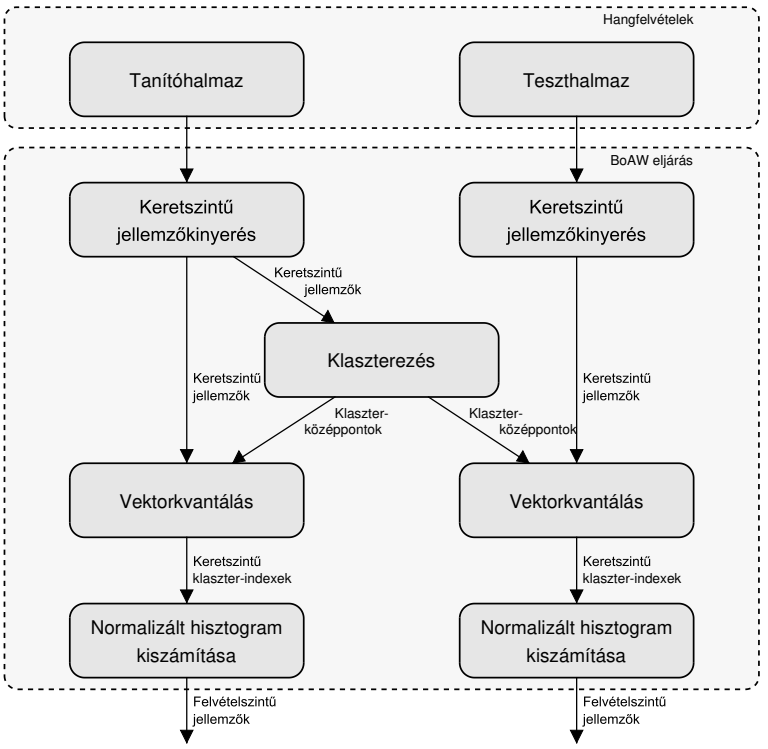
A következő lépés a vektorkvantálás, mely során az egyes felvételekhez tartozó keretszintű jellemzővektorokat kvantáljuk az előző lépésben generált kódszavaktól vett minimális euklideszi távolságuk alapján. Az eredeti jellemzővektorok helyettesítésre kerülnek a hozzájuk legközelebb lévő kódszó indexével. 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 a „*bag of audio words*" reprezentáció, ami majd a tanító algoritmusunk inputjául szolgál.

Látható, hogy a kvantálási lépést a teszthalmaz felvételeire is elvégezhetjük: bár a teszthalmaz felvételeit nem használtuk fel a klaszterezés során, a keretszintű jellemzővektoraikat ettől még besorolhatjuk az egyes klaszterekbe a kódszavaktól vett távolságuk alapján.

2.1. A szózsák eljárás paraméterei

A BoAW eljárásnak több olyan tulajdonsága, paramétere is van, mellyel befolyásolhatjuk a szózsák készítésének menetét. A most következő részben ismertetjük, hogy a tanulás sikerességére való befolyás szempontjából mely tényezők hatását vizsgáltuk.

Az egyik befolyásoló tényező a codebook készítése során használt klaszterező eljárás. Pancoast és Akbacak eredeti tanulmányukban k-means-t használtak [6];



1. ábra: Az akusztikus szózsák eljárás működési módja.

ugyanakkor a klaszterezendő keretek nagy száma miatt ennek a megközelítésnek igen magas a futási ideje. Rawat és mtsai. egyszerű véletlen mintavételezést javasoltak [7]; amellett, hogy ennek futási ideje nyilvánvalóan kedvezőbb, mint egy teljes klaszterezésnek, a tapasztalatok szerint (ld. [7,8]) ennek az eljárásnak a használata a teljesítményt is legfeljebb érintőlegesen rontja. Később Schmitt és mtsai. a k-means++ klaszterezési eljárás klaszterközéppont-inicializáló eljárását [9] alkalmazták a teljesen véletlenszerű mintavételezés helyett [10], így a klaszterközéppontok eloszlása kiegyensúlyozottabb lesz. Ezen kutatások alapján az utóbbi metódust választottuk, így kísérleteink során mindvégig azt fogjuk alkalmazni.

Egy másik szabályozható komponens a hisztogram előállításának módja. Pancoast és Akbacak azt javasolták, 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 [11]. Ha csupán a legközelebbi komponenst vesszük figyelembe, úgy gondoltuk, hogy túl nagy megszorítást eszközölünk, ezért kipróbáltuk a feltétel lazításának hatását is.

Az eddig taglalt módosításokon 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 normalizálhatjuk úgy, hogy a minimum és maximum értékekhez igazítva, 0 és 1 közötti skálára hozzuk az adatokat. Egy másik megoldás lehet, ha standardizálást hajtunk végre, tehát a mintákat úgy transzformáljuk, hogy szórásuk 1, átlaguk pedig 0 legyen.

3. Kísérletek

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

3.1. A magyar érzelemadatbázis

A kutatás során használt adatbázis 97 magyar anyanyelvű és magyarul beszélő személy hangját tartalmazza [12]. A beszédek televíziós műsorok során lettek felvéve. A szegmensek túlnyomó része érzelmekben gazdag, folyamatos, spontán beszédből lett kivágva. Kisebb részük improvizációs szórakoztató műsorból jön. Ebből fakadóan az elsőként említett kategóriába tartozó minták a színészi játék miatt, az érzelmek egy feljavított és egyértelműbb változatát tartalmazzák, míg a maradék improvizációs halmazban lévők közelebb állnak a hétköznapi, természetes érzelemkifejezéshez. Az adatbázis összesen 1111 mondatot tartalmaz, melyek egy 831 elemű tanító és 280 elemű teszt halmazra lettek osztva. Az osztályozás során négyféle érzelmet definiálunk a beszédekben: Harag, Öröm, Szomorúság és Semleges hangulat. Korábbi tanulmányok, melyek ugyanezzel az adatbázissal dolgoztak, 66-70%-os osztályozási pontosságot tudtak elérni [13].

3.2. Osztályozás

Az osztályozást SVM-ek (Support Vector Machines [14]) használatával végeztük, lineáris kernellel, a libSVM implementációt használva [15]. Az algoritmus komplexitás (complexity, C) paraméterét minden minta esetén többféle beállítással teszteltük. A lehetséges konfigurációk az alábbi 10 hatványok voltak: 0.00001; 0.0001; 0.001; 0.01; 0.1; 1 és 10. Az algoritmus tanulását és kiértékelését 10-szeres keresztvalidálással (10-fold cross-validation, CV) végeztük el. Tehát az aktuális mintahalmazt 10 egyenlő részre osztottuk, és minden lehetséges 9 tanító – 1 tesztelő halmaz kombinációra tanítottunk és kiértékeltünk egy SVM modellt. A teszthalmazra adott predikcióinkat a teljes tanítóhalmazon tanított SVM modellek szolgáltatták. 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 kereszvalidálás során a tanító halmazra kapott értékek alapján választottuk ki, hogy a kutatás egyes fázisaiban mely paraméterértékekkel dolgozzunk tovább.

Megközelítés	Maximális UAR	Codebook méret
Változatlan jellemzők	44.34%	32 768
Normalizált jellemzők	70.77%	8 192
Standardizált jellemzők	68.29%	8 192

1. táblázat. A keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció során.

3.3. Keretszintű jellemzőkészlet

Az akusztikus keretszintű jellemzők megválasztásának alapját a 2013-2014-es INTERSPEECH Számítógépes Paralingvisztikai Versenyen kiadott cikk adta. 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 window function) és 0.4 értékű szigmával, a másik két csoportot pedig 25 ms-os kerettel (Hamming window function) számítottuk. A Hamming-ablakokat a megszokott módon, átfedéssel, 10 ms-os eltolással helyeztük el.

frekvenciát valósítottak meg.

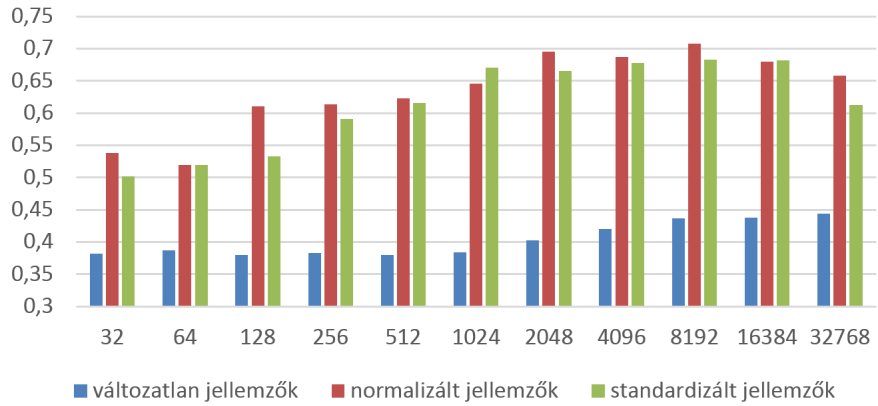
3.4. Az akusztikus szózsák eljárás paraméterei

Az egyik olyan paraméter, melyet minden esetben megadott értékekkel vizsgáltunk, az a codebook mérete volt. Az eljárás első lépése során megadhatjuk, hogy hány klasztert hozzunk létre, tehát hány kódszavunk legyen. Az általunk vizsgált értékek minden esetben az alábbi skálára terjedtek ki: 32, 64, 128, 256, 512, 1 024, 2 048, 4 096, 8 192, 16 384, 32 768. Az eljárás végén a számolt hisztogramot minden esetben normalizáltuk, azaz a kapott gyakoriságokat elosztottuk a hangfelvétel kereteinek számával.

4. Tesztek és eredmények

A következőkben ismertetésre kerül a kísérletek pontos menete, valamint az egyes fázisokban kapott eredmények kiértékelése.

Az első összehasonlítandó tényező a keretszintű jellemzővektorok kezelése volt a klaszterezés megkezdése előtt. Három esetet vizsgáltunk: 1) a jellemzővektorokat érintetlenül hagytuk, 2) a jellemzővektorokat normalizáltuk, 3) a jellemzővektorokat standardizáltuk. Az eredmények alapján (ld. 2. ábra és 1. táblázat) a normalizálás és a standardizálás közel azonos teljesítményjavulást nyújt ahhoz képest, ha az adatokon semmilyen további módosítást nem hajtunk végre. A bemeneti jellemzők normalizálásának vagy standardizálásának további előnye,



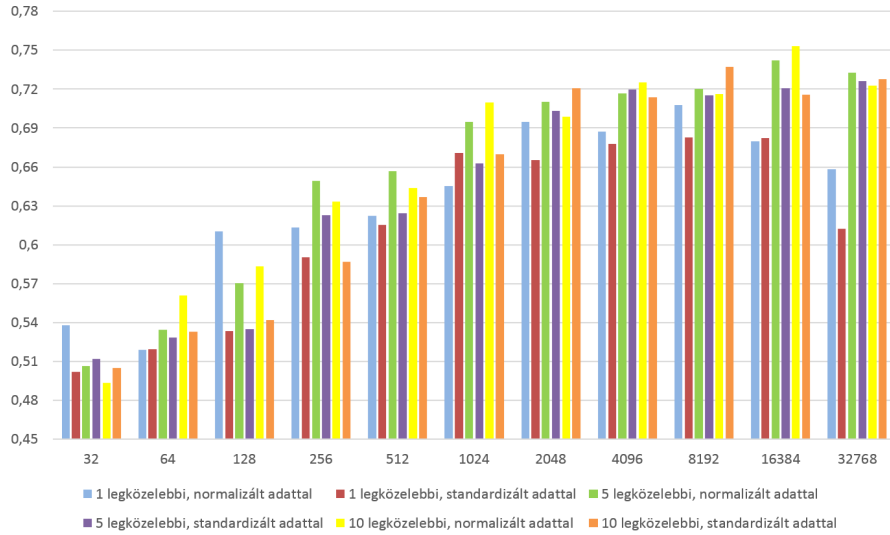
2. ábra: A keretszintű jellemzők normalizálásával és standardizálásával elért eredmények, különböző szózsák méreteknél.

hogy lényegesen kevesebb klaszter (mindkét esetben 8192) szükséges az optimális teljesítményhez, mint ha a jellemzőket változatlanul hagynánk (32768), így a felvételszintű SVM tanítása is kisebb jellemzőtérben történik. Ezen eredmények alapján a további teszteket párhuzamosan, normalizációval és standardizációval is elvégeztük.

A következő összehasonlításban azt vizsgáltuk, hogy a hisztogram létrehozásakor egy-egy keretszintű jellemzővektort hány legközelebbi kódszóhoz érdemes hozzárendelnünk. Most az $a = 1$, $a = 5$ és $a = 10$ eseteket vizsgáltuk; a három lehetőséget mind normalizált, mind standardizált adatokon kiértékeltük. A 3. ábrán szereplő legjobb eredményekből, levonható az a következtetés, hogy az $a = 5$ és $a = 10$ értékek a legtöbb esetben jobb felvételreprezentációt eredményeznek, mivel az osztályozás során kapott UAR értékek magasabbnak adódtak, mint az $a = 1$ beállítás esetén. Ez mind normalizálás, mind standardizálás esetén fennállt. A két hozzárendelés-érték segítségével elért teljesítményértékek között azonban nem tapasztaltunk lényeges különbséget, és az a paraméter az akusztikus szózsák eljárás által eredményül adott jellemzővektor méretét sem befolyásolja, csak annak kiszámítását befolyásolja csekély mértékben.

A 2. táblázatból leolvashatóak a legjobb eredmények a különböző a értékek esetén; ezek jól mutatják, hogy az 5 és 10 legközelebbi kódszót nézve közel azonos mértékű javulást kapunk, az 1 szavas változathoz képest; az $a = 10$ eset mindkét jellemzőnormalizálási eljárás esetén valamivel jobbnak adódott, de a különbség valószínűleg nem szignifikáns.

Az eddig ismertetett döntéseket a tanító halmazon végzett keresztvalidációs technika által adott százalékok alapján hoztuk meg. Ezzel határoztuk meg azt a szűkebb paraméterhalmazt, melyre a teszt mintákon is kiértékeljük az SVM algoritmust. A 2. és 3. ábrán is látható, hogy 1 024-es codebook méretig konzisztensen növekvő tendenciát mutat minden próba. Ezen számok alapján úgy döntöttünk,



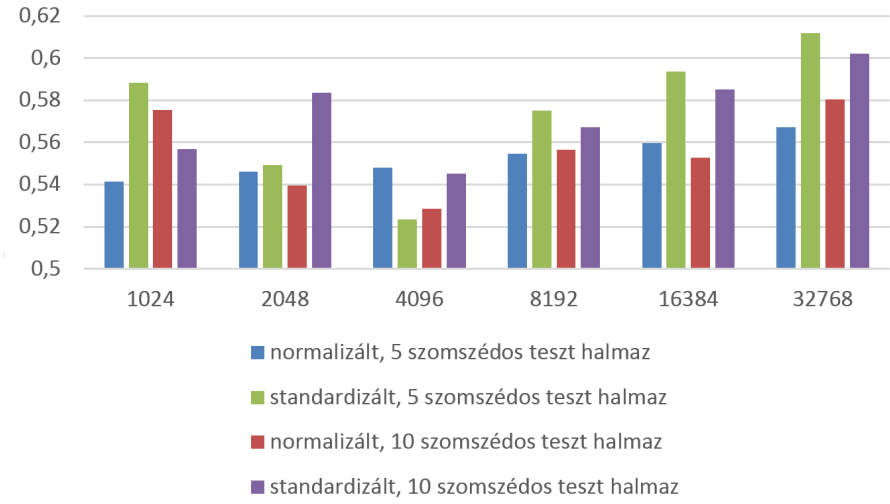
3. ábra: 1/5/10 legközelebbi kódszóhoz való társításnál kapott eredmények az adatok normalizálásának és standardizálásának függvényében.

Jellemző-transzformáció	**a**	**Maximális UAR**	**Codebook méret**
Normalizálás	1	70,77%	8 192
	5	74,20%	16 384
	10	75,31%	16 384
Standardizálás	1	68,29%	8 192
	5	72,63%	32 768
	10	73,73%	8 192

2. táblázat. Az 5 és 10 legközelebbi kódszóhoz való igazításnál kapott legjobb eredmények a normalizálás, valamint a standardizálás függvényében a keresztvalidáció során.

hogy a teszthalmazon való kiértékelést elegendő csupán az 1 024 és afeletti méretekre elvégezni, ugyanis ezalatt a tanulás állandó jelleggel rosszabbnak bizonyult. A 4. ábrán láthatók azon eredményeink, melyeket a teszt halmazokon való UAR-ok kiszámítása után kaptunk; a keresztvalidálás során kapott legjobb paraméterekhez tartozó, teszthalmazon mért UAR értékek pedig a 3. táblázatban találhatóak.

Az itt látható értékek lényegesen alacsonyabbak ugyan, mint amiket a keresztvalidáció során kaptunk, de a két halmazon kapott értékek természetesen nem hasonlíthatók össze direktben. Ugyanígy, a korábban erre az adatbázisra elért eredmények sem vethetőek össze közvetlenül az általunk kapott értékekkel, hiszen ott (egyszeres) keresztvalidációt végeztek, és az eredményeket standard osztályozási pontosságban adták meg, míg jelen tanulmányunkban, a kiegyensú-



4. ábra: A teszthalmazon való kiértékelés eredményei.

Jellemző-transzformáció	**a**	**Maximális UAR**	**Codebook méret**
Normalizálás	5	55,97%	16 384
	10	55,27%	16 384
Standardizálás	5	61,19%	32 768
	10	56,83%	8 192

3. táblázat. A teszthalmazon való kiértékelés eredményei.

lyozatlan osztályeloszlást ellensúlyozandó, UAR-t használtunk. A 4. ábrán látható a codebook méretének pozitív hatása, de a tendencia korántsem egyértelmű; például az 1 024 klasztert használó modellek jobbnak adódtak, mint a 4 096 klasztert használók. Elmondható az is, hogy minél nagyobb méretet választunk, annál nagyobb annak is az esélye, hogy túltanulási hibába futunk vele és az osztályozónk elveszíti általánosító képességét.

5. Összegzés

Jelen cikkünkben az akusztikus szózsák (Bag-of-Audio-Words, BoAW) jellemzőreprezentációs eljárást alkalmaztuk egy magyar nyelvű érzelemfelismerési feladaton. Az eljárásnak számos paramétere van, így számos gépi tanulási modellt kellett tanítanunk a különböző paraméter-kombinációkra. Mért eredményeink alapján a bemeneti jellemzőket mindenképpen érdemes azonos skálára hoznunk normalizálás vagy standardizálás segítségével, és az alkalmazott kódszavak szá-

mát is érdemes magasnak választanunk (8 192-32 768). Az egyes kereteket is érdemes párhuzamosan több klaszterbe sorolnunk.

Annak kapcsán, hogy az itt elért eredményeink által merre haladhatunk tovább a későbbiekben, több lehetőség is felmerül. Jelenleg a keretszintű jellemzők közül, csupán az első 65-öt vettünk figyelembe; a későbbiekben használhatjuk az LLD-k elsőrendű deriváltjait is. Másrészről a codebook generálás során alkalmazott klaszterező eljárást jelenleg korábbi kutatásokra hivatkozva választottuk ki. Ezen metódusok eredményességét a tanulóadatbázisunkon mi magunk is letesztelhetnénk. Emellett lehetőségünk van más keretszintű jellemzőkészleteket is letesztelni. További érdekes kísérletek végezhetők több adatbázis használatával is; hasonló jellegű korpuszok esetén jogos kérdés a BoAW eljárás paramétereinek stabilitása, de akár a kódszavak átvitele is.

Hivatkozások

1. James, J., Tian, L., Inez Watson, C.: An open source emotional speech corpus for human robot interaction applications. In: Interspeech, Hyderabad, India (2018) 2768–2772
2. Burkhardt, F., van Ballegooy, M., Engelbrecht, K.P., Polzehl, T., Stegmann, J.: Emotion detection in dialog systems: Applications, strategies and challenges. In: ACII, Amsterdam, Hollandia (2009) 985–989
3. Hossain, M.S., Muhammad, G.: Cloud-assisted speech and face recognition framework for health monitoring. Mobile Networks and Applications **20**(3) (2015) 391–399
4. Norhafizah, D., Pg, B., Muhammad, H., Lim, T.H., Binti, N.S., Arifin, M.: Detection of real-life emotions in call centers. In: ICIEA, Siem Reap, Kambodzsa (2017) 985–989
5. Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Interspeech, Lisszabon, Portugália (2005) 1841–1844
6. Pancoast, S., Akbacak, M.: Bag-of-Audio-Words approach for multimedia event classification. In: Interspeech, Portland, USA (2012) 2105–2108
7. Rawat, S., Schulam, P.F., Burger, S., Ding, D., Wang, Y., Metze, F.: Robust audio-codebooks for large-scale event detection in consumer videos. In: Interspeech, Lyon, Franciaország (2013) 2929–2933
8. Schuller, B., Steidl, S., Batliner, A., Hantke, S., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A.S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., Tzirakis, P., Zafeiriou, S.: The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring. In: Interspeech. (2017) 3442–3446
9. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: SODA, New Orleans, Louisiana, USA (2007) 1027–1035
10. Schmitt, M., Ringeval, F., Schuller, B.: At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. In: Interspeech, San Francisco, USA (2016) 495–499
11. Pancoast, S., Akbacak, M.: Softening quantization in bag-of-audio-words. In: ICASSP, Florence, Olaszország (2014) 1370–1374

12. Sztahó, D., Imre, V., Vicsi, K.: Automatic classification of emotions in spontaneous speech. In: COST 2102, Budapest (2011) 229–239
13. Vicsi, K., Sztahó, D.: Recognition of emotions on the basis of different levels of speech segments. Journal of Advanced Computational Intelligence and Intelligent Informatics **16**(2) (2012) 335–340
14. 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) (2001) 1443–1471
15. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 1–27