Eötvös Loránd Tudományegyetem

Informatikai Kar

Programozáselmélet és Szoftvertechnológiai

Tanszék

Automatikus zenei hangszerfelismerés többszólamú zenében mély neuronhálók segítségével

Témavezető:
Gombos Gergő
Adjunktus, PhD

Szerző:
Hamrák János
programtervező informatikus MSc

Budapest, 2020

Az eredeti szakdolgozati / diplomamunka témabjelentő helye.

Tartalomjegyzék

1. fejezet

Bevezetés

Napjainkban a zenéhez legkönnyebben digitális formában, a világhálón keresztül férhetünk hozzá. Néhány kattintással olyan zenei tartalomszolgáltatókat érhetünk el, melyek széleskörű adatbázissal rendelkeznek. A folyamatosan bővülő adatmennyiség ellenére ezeknek az adatbázisoknak átláthatónak és könnyen kezelhetőnek kell maradniuk, hogy a felhasználókat a kívánt módon tudják kiszolgálni. Ennek érdekében nap mint nap új megoldások születnek zenei információk automatikus kinyerése és feldolgozása céljából. Ezek teszik lehetővé a digitálisan tárolt zenék körében például az osztályozást vagy keresést.

A zenei információk kinyerésének tudományába (Music Information Retrieval, a továbbiakban: "MIR") tartozik a továbbiakban taglalt probléma, az automatikus hangszerfelismerés. Ez egy osztályozási feladat. Célja, hogy a meglévő digitális hanganyag alapján az adott zenéről megállapítsuk, hogy milyen hangszerek szólalnak meg benne. Ezt az információt több célra is fel tudjuk használni, például:

- Későbbi feldolgozásra, további MIR feladatok inputjaként
- Statisztikák készítésére
- Adatbázisban való keresés szűrőfeltételeként
- Egy ajánlórendszer részeként, ahol az aktuális zeneszámot követően például egy hangszerelésében hasonló számot szeretnék ajánlani a felhasználónak.

Az automatikus hangszerfelismerés feladatot több aspektusból lehet megközelíteni, például a bemeneti adatok jellege, reprezentációja, a megvalósított architektúra,

az osztályozás módszere, vagy az osztályok száma alapján. A dolgozatom keretein belül felkutatok néhány létező megoldást, majd ezekre alapozva prezentálom saját megközelítésemet és ennek eredményeit. Az általam bemutatott megoldás egy multi-class multi-label osztályozást valósít meg mély neuronhálós rendszer segítségével többszólamú zenében.

1.1. Motiváció

Az ember kognitív képességei segítségével a zenében könnyedén fel tudja ismerni az egyes hangszereket. Ugyanez a feladat a számítógép számára azonban már sokkal kevésbé triviális. Ennek egyik oka, hogy egy hangszer megszólaltatásának digitális reprezentációja nagyon változatos lehet. Függ például a hangszíntől, hangmagasságtól, hangerőtől és előadásmódtól, de a felvétel minőségétől és az esetleges háttérzajtól is. További nehezítő körülmény a többszólamúság, amikor egy időben több hangszert is megszólaltatunk, ezzel összemosva az egyszólamú környezetben is sokváltozós képünket.

A MIR nagyban támaszkodik a mesterséges intelligenciára. A számítógépek számítási kapacitásának folyamatos növekedése és az elérhető adathalmazok gyarapodása által pedig egyre nagyobb figyelmet kap a mesterséges intelligencia egy kiemelten számításigényes részterülete: a mély tanulás. Ezt bizonyítja, hogy az évente megrendezésre kerülő ISMIR (International Society for Music Information Retrieval) konferencián 2010-ben még csak 2 ([1], [2]) mély tanulással kapcsolatos cikk jelent meg, de 2015-ben már 6, 2016-ban pedig már 16. [3]

A mély tanulás tehát egy ígéretes módszer lehet a MIR problémák megoldásában, ideértve az automatikus hangszerfelismerést is. Ezt kihasználva és megoldás életszerűségére törekedve döntöttem úgy, hogy elkezdek kísérletezni mély neuronhálókkal többszólamú zenében. Célom volt találni egy tanításra alkalmas adathalmazt, azon pedig tervezni egy olyan mély neuronhálós rendszert, amely a jelenlegi megoldások pontosságát meghaladja.

1.2. A dolgozat felépítése

Dolgozatomban tehát az eddigi kapcsolódó kutatásokat, illetve saját munkám eredményét dolgozom fel. A következő alfejezetben felsorolom az általam relevánsnak tartott, a State-of-the-Arthoz vezető kutatásokat.

A második fejezetben betekintést adok a téma elméleti hátterébe. Először kifejtem a zenével kapcsolatos főbb fogalmakat, bemutatom fontosabb tulajdonságait, reprezentációit. Kitérek a MIR bemutatására is. Ezután bevezetem a gépi tanulás és a mély tanulás fogalmát.

A harmadik fejezet az adathalmazokról fog szólni. Itt előbb felsorolom az adathalmazok kiválasztásának szempontjait, majd minden felhasznált adathalmaznak ismertetem a főbb jellemzőit.

A negyedik fejezetben a módszertanról ejtek szót. Itt kifejtésre kerülnek az adatok előfeldolgozási módszerei, az általam bemutaott mély tanulási architektúrák, illetve ezek megvalósításai.

Az ötödik fejezetben részletezem az általam végzett kisérleteket és ezek eredményeit. Ezeket összevetem egymással, illetve a releváns State-of-the-Art kutatásokkal.

A hatodik fejezetben összegzem a leírtakat, valamint továbbgondolom a kutatásomat, felvázolok néhány ötletet annak jövőjéről.

1.3. Kapcsolódó munkák

Az automatikus hangszerfelismerés témában a korábbi kutatások túlnyomó része a monofónikus, azaz egyhangszeres zenékkel foglalkozik. Martin és Kim [4] mintafelismerési statisztikai technikája 1023 izolált hangjegy és 15 különböző hangszer között a hangszercsaládok felismerésében 90%-os, egyéni hangszerek felismerésében pedig 70%-os pontosságot produkált. Brown [5] a kepsztrális együtthatókat használta fel K-közép klaszterezési módszeréhez. Eronen és Klapuri [6] széleskörű, spektrális és időbeli feature-halmaz segítségével - összesen 43 különböző feature felhasználásával - 81%-os hangszer és 95%-os hangszercsalád pontosságról számolt be. Deng [7] klasszikus zenei hangszerek tekintetében elemezte a különböző, gépi tanulási módszerekben használatos feature összeállításokat. Bhalke [8] tört Fourier-transzformáción alapuló MFCC feature-ök segítségével tervezett CPNN osztályozót mutatott be, amellyel

hangszercsaládok tekintetében 96.5%-os, hangszerek tekintetében pedig 91.84%-os pontosságot ért el.

Többszólamú környezetbe való átültetéssel foglalkozott Burred tanulmánya [9], aki a többszólamúságot két kísérlettel közelítette meg. Először csak egy-egy hangjegyet kombinált össze többszólamú hangjeggyé. Itt két szimultán hangjegy esetén 73.15%-os, három hangjegyre 55.56%-os, négy hangjegy kombinációjára pedig 55.18%-os pontosságot sikerült elérni. Másik kísérletként hosszabb szekvenciákat kombináltak össze, ekkor két hang esetén 63.68%-os, három hang esetén pedig 56.40%-os pontosságot kaptak.

Eggink és Brown [10] a polifónikus zenékben a hiányzó adat elméletükkel próbálták feltárni az egyes hangszereket. Ennek lényege, hogy felderítették azon idő- és frekvenciabeli részeket a zenén belül, ahol szeparáltan egy hangszer tulajdonságait vélték felfedezni és ezt dolgozták fel. Erre a módszerre épített Giannoulis és Klapuri [11] kutatása is, és hasonló megközelítést alkalmazott Garcia [12] is.

Jiang [13] egy többlépcsős megoldást mutat be. Első lépésben a hangszercsaládot határozták meg, ezzel szűkítve a a lehetséges hangszerek halmazát és a változók számát. A pontos hangszer-meghatározás csak ezután következett.

Az előbbi kutatások többnyire hagyományos gépi tanulási megoldásokat alkalmaztak, amelyekhez maguk nyerték ki a különböző bemeneti feature-öket. Humphrey [14] írásában a mély tanulási architektúrákat ismerteti a MIR terület korszerű irányzataként. A témában gyakorlati segítségként szolgál Choi [15] írása, amiben konkrét adatreprezentációkat, mély neuronhálós rétegeket, és mély tanulási technikákat mutat be.

Li [16] a nyers hanganyagot inputként felhasználva egy konvolúciós mély neuronhálós rendszert mutatott be a polifónikus zenében való automatikus hangszerfelismerés kapcsán. Ezt a megoldást aztán összevetette hagyományos gépi tanulási módszerekkel is. A mély neuronhálós rendszer teljesített legjobban. 75.60%-os pontossággal, 68.88%-os felidézéssel, 72.08 mikro F értékkel és 64.33 makro F értékkel. Han [17] szintén egy mély konvolúciós hálót használt, azonban az osztályozás szempontjából máshogyan járt el: a zenékben egy darab domináns hangszert keresett. Bemenetként a zenék spektogramját használta fel, 0.602-es mikro és 0.503-as makro F értéket ért el.

2. fejezet

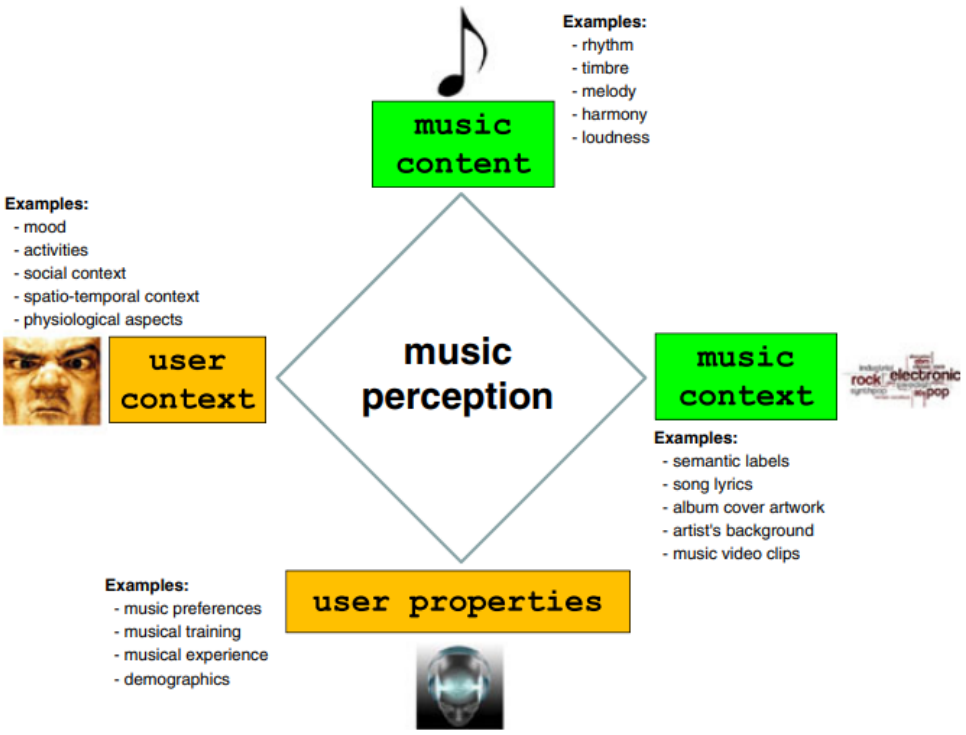
Elméleti háttér

Ebben a fejezetben a dolgozathoz kapcsolódó fogalmakat és elméleti alapokat mutatom be. Először magának a zenének a releváns tulajdonságairól ejtek szót. Ezután ismertetem a MIR kutatási területet, amelybe dolgozatom is tartozik. Majd végül a mesterséges intelligencián alapuló megoldásokról nyújtok elméleti bevezetőt, érintve a hagyományos gépi tanulás és a mély tanulás módszereit is.

2.1. Zene és reprezentációi

2.1.1. Zene fogalma, tulajdonságai

A zene egy meglehetősen összetett fogalom. Az ember számára a zene megjelenhet például hang formájában, leírhatjuk őket szimbólumok segítségével egy kottában, előfordulhat szöveges formában dalszövegként, képi formában egy albumborító, vagy egy zenész képében, illetve mozdulatokban egy zenei előadás keretében. A teljes zenei élményt ezek kombinációja nyújtja. A zene észlelését befolyásoló tényezőket Schedl [18] a következő kategóriákba sorolta: a zene tartalma (music content), a zene kontextusa (music context), a hallgató kontextusa (user context) és a hallgató tulajdonságai (user properties). [19]



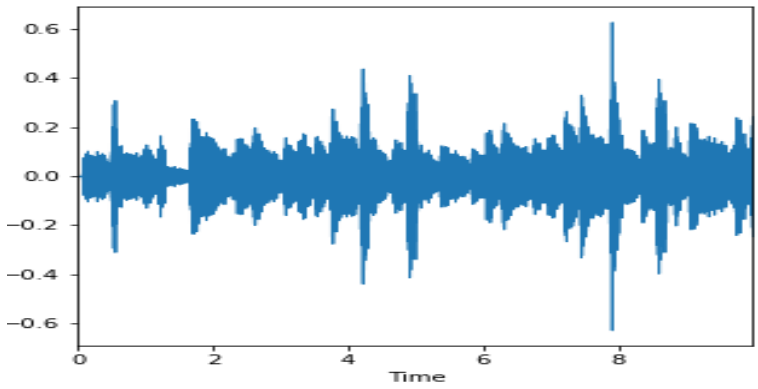
2.1. ábra. A zene észlelését meghatározó tényezők, forrás: [18]

A zenei tartalom fogalma utal azokra a tulajdonságokra, melyek a hangok fizikai jelként való leírása definiál. Ilyen például a ritmus, a hangszín, a dallam, a harmónia, a hangerő, vagy a dalszöveg. Ezzel szemben a zene kontextusa alatt azokat a tényezőket értjük, melyeket nem tudunk közvetlenül a zenéből kinyerni, mégis szorosan kapcsolódik hozzá. Ide tartozik például az előadó hírneve, az albumborító, a művész kulturális- vagy politikai háttértörténete, vagy a zeneéhez készített videoklip. Ami a hallgatóval kapcsolatos aspektusokat illeti, a hallgató kontextusa alatt értjük a dinamikus, gyorsan változó tényezőket. Ide sorolható a hallgató aktuális hangulata, tevékenysége, társadalmi helyzete, tér- és időbeli helye, illetve pszichológiai állapota. Ezzel ellentétben a hallgató tulajdonságai az állandó, vagy csak lassan változó jellemzőit takarja. Ilyen az egyén zenei ízlése, zeneelméleti képzettsége, demográfiai adatai, a hallgatott előadóval kapcsolatos véleménye, vagy a barátai zenei ízlése, véleménye. [19]

Dolgozatomban az észlelést meghatározó tulajdonságok közül a zenei tartalmat fogom felhasználni a hangszerek kinyerése érdekében. A zenei tartalom számítógépes felhasználásához pedig elengedhetetlen, hogy a hangokat megfelelően tudjuk számítógépen ábrázolni.

2.1.2. Hang reprezentációk

A hangok fizikai mivoltukban rezgésekként jelennek meg. A rezgéseket matematikailag olyan folytonos függvényekkel tudjuk leírni, melyek értelmezési tartománya az idő, értékkészlete pedig a nyugalmi állapothoz viszonyított pillanatnyi kitérés. Ilyen lehet például egy szinuszgörbe. Ahhoz, hogy a hangokat számítógépen tudjuk tárolni és feldolgozni, ezeket a függvényeket kell ábrázolnunk. Mivel azonban a számítógép számábrázolása véges, ezért a hangokat először digitalizálni kell. Ez azt jelenti, hogy a folytonos függvényeket diszkrét, azaz véges helyen vett és véges értékekkel rendelkező fügvényekké alakítjuk. Ez úgy történik, hogy ez eredeti függvényünkből megadott időközönként mintát veszünk, diszkrét értékre kerekítjük, és ezeket az értékeket összefűzzük. Az így kapott függvény lesz a hangnak az idő függvényében ábrázolt digitális reprezentációja.

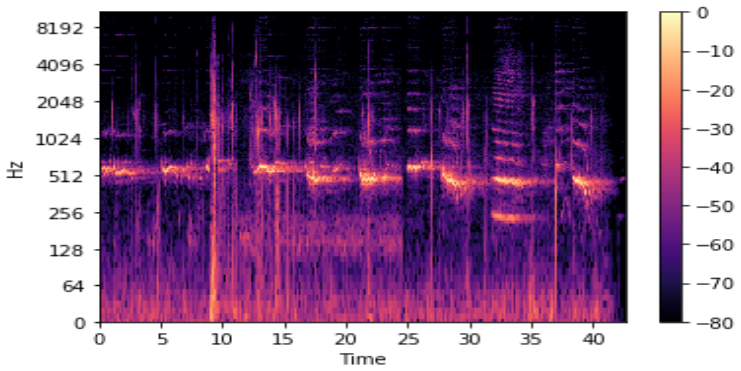


2.2. ábra. Tíz másodperces hanganyag hanghullám reprezentációja

A hangok idő függvényében ábrázolt digitális reprezentációja tehát egydimenziós, mivel egy függvénygörbének tekinthetjük. Ezt szokták hívni hullámforma reprezentációnak, illetve nyers hangnak (raw audio) is, ugyanis további reprezentációkká tudjuk transzformálni. A MIR területen megjelenő mély tanulási megoldások jelentős része ezen nyers hangábrázolás helyett inkább a kétdimenziós reprezentációkat alkalmazza bemenetként. Ezt azzal indokolják, hogy a nyers bemeneten való tanítás sikeréhez nagyobb adathalmaz szükséges, mint a kétdimenziós reprezentációkéhoz. [15]

Az említett kétdimenziós reprezentációk Fourier-transzformáción alapszanak. A Fourier-transzformáció nagyon leegyszerűsítve egy olyan függvény, amely bemenetként kap egy idő függvényében ábrázolt jelet és ezt felbontja frekvenciákra [20].

- **Ablakozott Fourier transzformált (STFT)**: Gábor Dénes magyar fizikus nevéhez kötődik. A bemeneti jelet egyenlő méretű időszeletekre (ablakokra) osztjuk, majd ezeken alkalmazzuk a Fourier-transzformációt. Ezáltal kapjuk meg a hang spektrumát (spectogram). [20], [15]



2.3. ábra. Spektrum [21]

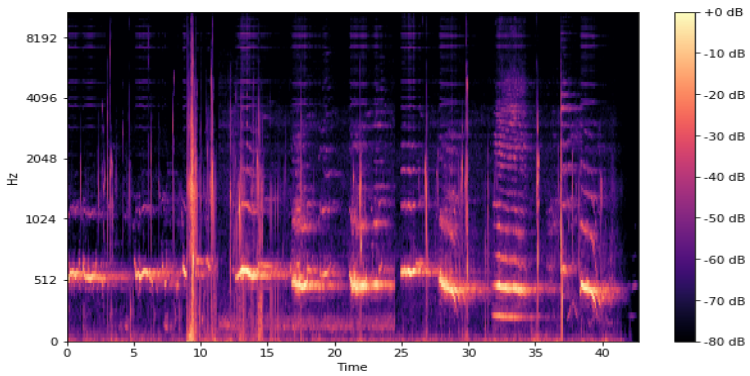
- **Mel-frekvenciára transzformált spektrum (Mel-spectogram)**: A mel-spectogram az előbb említett spektrum Mel skálára való transzformáltja. A Mel skálát egy nemlineáris függvény segítségével kapjuk a frekvenciaskálából. Célja, hogy a skála jobban reprezentálja az ember hallásának tartományait. Tehát az értékek közti különbség a Mel skálán megfeleltethető legyen annak, hogy az ember mennyire különböző magasságúnak hallja ezeket. A frekvencia skálán például sokkal nagyobbnak érezzük az 500Hz és 1000Hz közti különbséget, mint a 7500Hz és 8000Hz közti különbséget. A mel-frekvenicára való konverzió képlete a következő:

$$Mel(f) = 2595 * ln(1 + \frac{f}{700}) \tag{2.1}$$

Ahol Mel(f) az adott frekvenciaérték mel-skálán való értéke, f pedig az adott frekvencia érték. A képlet segítségével egymást átfedő frekvencia sávokat alakítunk ki, melyek a mel-skálát tekintve egyenlő távolságra helyezkednek el egymástól, majd a frekvenciasávok energia értékeit egyenként leképzzük mel-skálára. [20], [15]

Léteznek egyéb, hasonló skálák is, mint pl. Bark skála, vagy az hallás pszichológián alapúló ERB. Ezek MIR környezetben még nem kerültek összevetésre,

de beszédfelismerés környezetben nem mutatnak szignifikáns eltérést az egyes skálák eredményei. A Mel skála, illetve a mel-spectogram használata azonban egy gyakran használt és jól felhasználható reprezentációnak bizonyul MIR környezetben. [15]

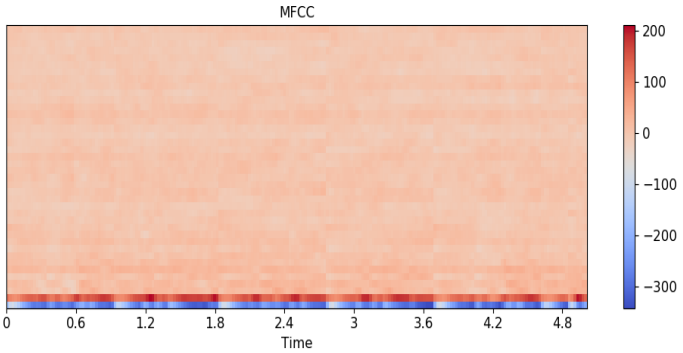


2.4. ábra. Melspectogram reprezentáció [21]

- **Mel-frekvencián vett kepsztrum együtthatók (MFCC)**: Az MFCC együtthatókat a melspectogramon való diszkrét koszinusz-transzformációval kapjuk meg. A képlet a következő:

$$C_i(m) = \sum_{n=0}^{M-1} S_i(n) cos[\pi m \frac{n - 0.5}{M}], 0 \leq m \leq M \tag{2.2}$$

Ahol M a frekvenciaszeletek száma, $S_i(n)$ az egyes sávokban kiolvasott energia értékek és $C_i(m)$ az i-edik MFCC együttható. Ezzel a mel-spectogramnál tömörebb reprezentációt kapunk. Ennek hátránya lehet az információvesztés, előnye pedig az apró zajok kiszűrése. [8]



2.5. ábra. MFCC reprezentáció [21]

2.1.3. Music Information Retrieval (MIR)

A bevezetőben már említettem a zenei információk kinyerését (music information retrieval - MIR). Ez egy interdiszciplináris kutatási terület, magában hordozza többek között a zeneelmélet, pszichoakusztika, pszichológia, informatika, jelfeldolgozás és gépi tanulás tudományágakat. Céljára jól utal az elnevezése, zenékből szeretnénk releváns információt kinyerni, és ezeket felhasználni [15]. A felhasználásra szerintem nagyon jó, életszerű példát ad Downie 2003-as cikkének [22] bevezetője, amelyet a következőképp fordíthatunk le:

"Képzeljünk el egy világot, ahol egyszerűen felénekelhetjük egy számítógépnek a dalrészletet, ami már reggeli óta a fejünkben jár. A gép elfogadja a hamis énekünket, kijavítja, és azonnal javaslatot tesz arra, hogy éppen a "Camptown Races" című számra gondoltunk. Miután mi belehallgatunk a gép által talált számos relevánsnak tartott MP3 fájl egyikébe, elfogadhatjuk a gép javaslatát. Ezután elégedetten elutasíthatjuk a felajánlást, hogy az összes további létező verzióját is felkutassa a dalnak, ide értve a nemrég megjelent olasz rap verziót, vagy a skótdudás duettre írt kottát." [22]

Figyeljük meg, hogy ez a hétköznapi eset mennyire össztett probléma. A következő feladatok jelennek meg:

- Az emberi éneklés, vagy dúdolás alapján hangfelismerés.
- Hang alapú lekérdezés egy zenei adatbázisban az előbbi bemenettel.
- Hangelemzés, feldolgozás, hogy a hamis hangokat ki tudjuk javítani, az esetleges háttérzajokat eltávolítsuk, illetve ha kell, a dallamból automatikusan kottát generáljunk.
- Hasonlóságon alapuló keresés zenék között, hogy megtaláljuk a kívánt dalt az adatbázisban.
- Zenei feldolgozások detektálása, hogy további verzióit is megtaláljuk egy adott dalnak.

MIR problémák definiálását több szempontból közelíthetjük meg. Choi cikke [15] két tengelyre osztja fel a problémateret: szubjektivitás és eldöntési időmérték. A

szubjektivitás tengelyen léteznek szubjektívebb feladatok, melyekre nincsenek egyértelmű válaszok. Ilyen lehet például a zene műfajának meghatározása. Objektívebb feladatoknak tekinthetjük azokat, melyek eredménye egyértelműen meghatárhozható, esetleg számszerűsíthető. Ide tartozik a hangszerfelismerés, vagy a tempó észlelés. [15]

A másik tengely, az eldöntési időmérték aszerint sorolja be a feladatokat, hogy mekkora időegységeken értelmezhető egy becslés. Ez egy relatív mérték. Például a dallamfelismerés eldöntési időmértékére azt mondhatjuk, hogy alacsony, mert egy felismert dallam jó eséllyel nem fedi le az egész zenét. Másik kifejezéssel azt mondhatjuk, hogy ez egy időben változó, azaz dinamikus tulajdonság. Ellenben a tempó általában állandó értékű az egész zenében, így a teljes zeneszámot fel tudjuk címkézni egy adott tempóval. Erre azt mondjuk, hogy eldöntési időmértéke relatív magas, azaz ez egy statikus tulajdonsága a zenének.[15]

A hangszerfelismerést tekinthetjük statikus, illetve dinamikus feladatnak is a probléma megközelítésének függvényében. Dinamikus, ha erős címkézést szeretnénk megvalósítani, tehát arra vagyunk kíváncsiak, hogy adott időpillanatban éppen milyen hangszerek szólalnak meg. Gyenge címkézés esetén viszont a feladat statikussá válik. Ebben az esetben az egész zenére vetítve szeretnénk címkéket kapni egyes hangszerek jelenlétével, vagy a többi hangszerrel szembeni dominanciájával kapcsolatban.

A MIR terület főbb részterületei és feladatai Schedl [19] cikkére alapozva a következők:

- Jellemző kinyerés (feature extraction)
  - hangszín leírás pl. [23], [24]
  - kotta- és dallamkinyerés pl. [25], [26], [27]
  - ütem lekövetés, tempó becslés pl. [28], [29], [30]
  - tonalitás becslés pl. [31], [32], [33], [34], [35], [36]
  - struktúrális analízis, szegmentáció pl. [37], [38], [39]
- Hasonlóságon alapuló feladatok
  - hasonlóság mérés pl. [40], [41], [42]

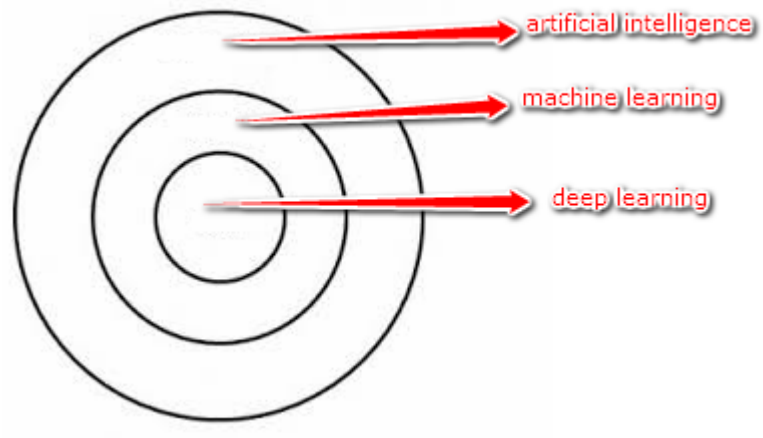
  - zenei feldolgozás felismerés pl. [43], [44]
  - dúdoláson alapuló lekérdezés pl. [45], [46], [47]

- Osztályozási feladatok
  - érzelem- és hangulatfelismerés pl. [48], [49]
  - műfaj szerinti osztályozás pl. [50], [51]
  - hangszerfelismerés pl. [52]
  - szerző / előadó / énekes felismerés pl. [53]
  - automatikus címkézés pl. [54], [55], [56]

- Alkalmazások
  - hanganyaghoz forrásazonosító készítés (fingerprinting) pl. [57], [58]
  - tartalom alapú lekérdezés pl. [59]
  - zene ajánlás pl. [60], [61], [62]
  - lejátszási lista generálás pl. [63], [64], [65], [66]
  - kottázás pl. [67], [68], [69]
  - dal/előadó sikeresség becslés pl. [70], [71], [72]
  - zene vizualizáció pl. [73], [74], [75], [76], [77]
  - felhasználói felületen való zenei böngészés pl. [78], [79], [80], [81], [82]
  - személyre szabott, alkalmazkodó rendszerek pl. [83], [84], [85], [86]

2.2. Mesterséges intelligencia

A mesterséges intelligencia egy általános fogalom az emberi gondolkodás számítógéppel való reprodukálására történő módszerekre. Ahogy arról a bevezetésben is szót ejtettem, a MIR tudományág gyakran használ mesterséges intelligencián alapuló megoldásokat. A továbbiakban két mesterséges intelligenciát megvalósító módszert mutatok be röviden: elsőként a gépi tanulást, majd a mély tanulást, amely a gépi tanulás egy ágazata és napjaink meghatározó trendje. [87]



2.6. ábra. Egyes fogalmak közti tartalmazás szemléltetve, forrás: [87]

2.2.1. Gépi tanulás (machine learning)

A gépi tanulás tehát a mesterséges intelligencia megvalósításának egy módszere. Lényege, hogy explicit utasításokat tartalmazó program helyett a bemeneti adatokat egy tanító algoritmusnak adjuk át. Ha elég mennyiségű és minőségű adatot szolgáltatunk a tanításhoz, akkor a modellünk

Hagyományos gépi tanulásról beszélünk, amikor

//TODO

2.2.2. Mély tanulás (deep learning)

A mély tanulás gyakorlatilag a gépi tanulás egy részhalmaza.

//TODO miben más mint az ML, architektúrák stb

3. fejezet

Adathalmaz

Ebben a fejezetben a munkám kapcsán felkutatott és alkalmazott adathalmazokról lesz szó. Egy deep learning megoldás tervezésének első lépéseként érdemes egy alkalmas kiinduló adathalmazt kiválasztani. Ezt aztán a modell tanítására és tesztelésre használjuk.

3.1. Kiválasztási szempontok

//TODO ismir dataset gyűjtemény pl. többszólam, ingyenesen elérhető, szerteágazó (not biased), stb gyenge címkézés

3.2. Philharmonia Orchestra

Kutatásom első fázisában a Philharmonia Zenekar ingyenesen elérhető hangminta könyvtárát használtam fel. Ebben egyszólamú mintákat találunk. A minták a főkönyvtáron belül a bennük megszólaló hangszer nevével megegyező könytárban találhatóak, ez biztosítja a címkéket.

//TODO tulajdonságai

3.3. OpenMIC

A többszólamúság bevezetését a kutatásomban az OpenMIC [88] adathalmaz felhasználásával értem el. //TODO openmic cikk alapján

4. fejezet

Módszertan

4.1. Előfeldolgozás

//TODO adattisztítás, reprezentáció változtatás

4.2. Architektúra

//TODO felhasznált deep learning architektúra(?)

4.3. Megvalósítás

//TODO implementáció részletei

4.3.1. Könyvtárak

//TODO fejlesztői könyvtárak, függőségek - nem biztos hogy kell

5. fejezet

Kísérletek, eredmények

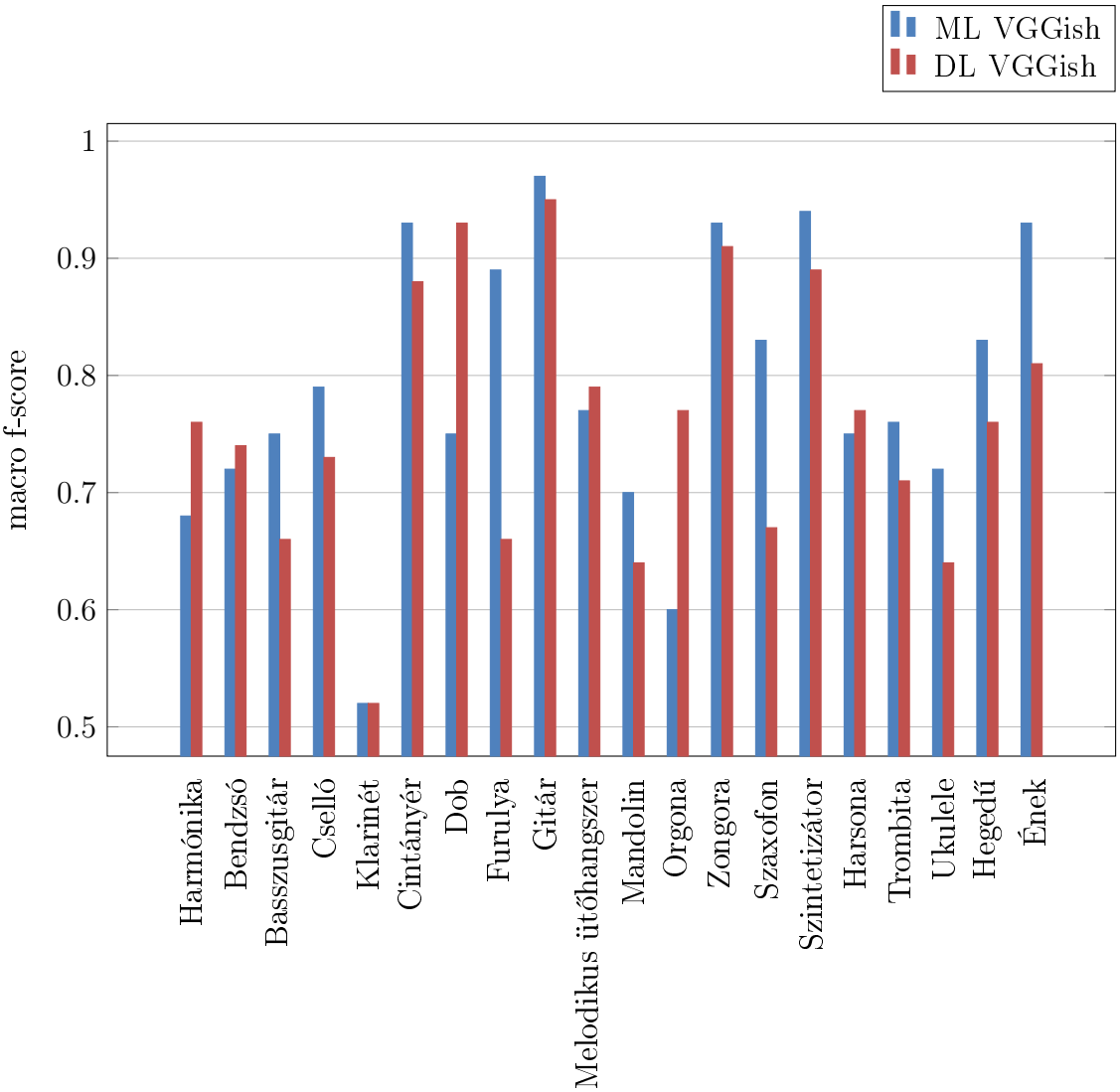
Lorem ipsum

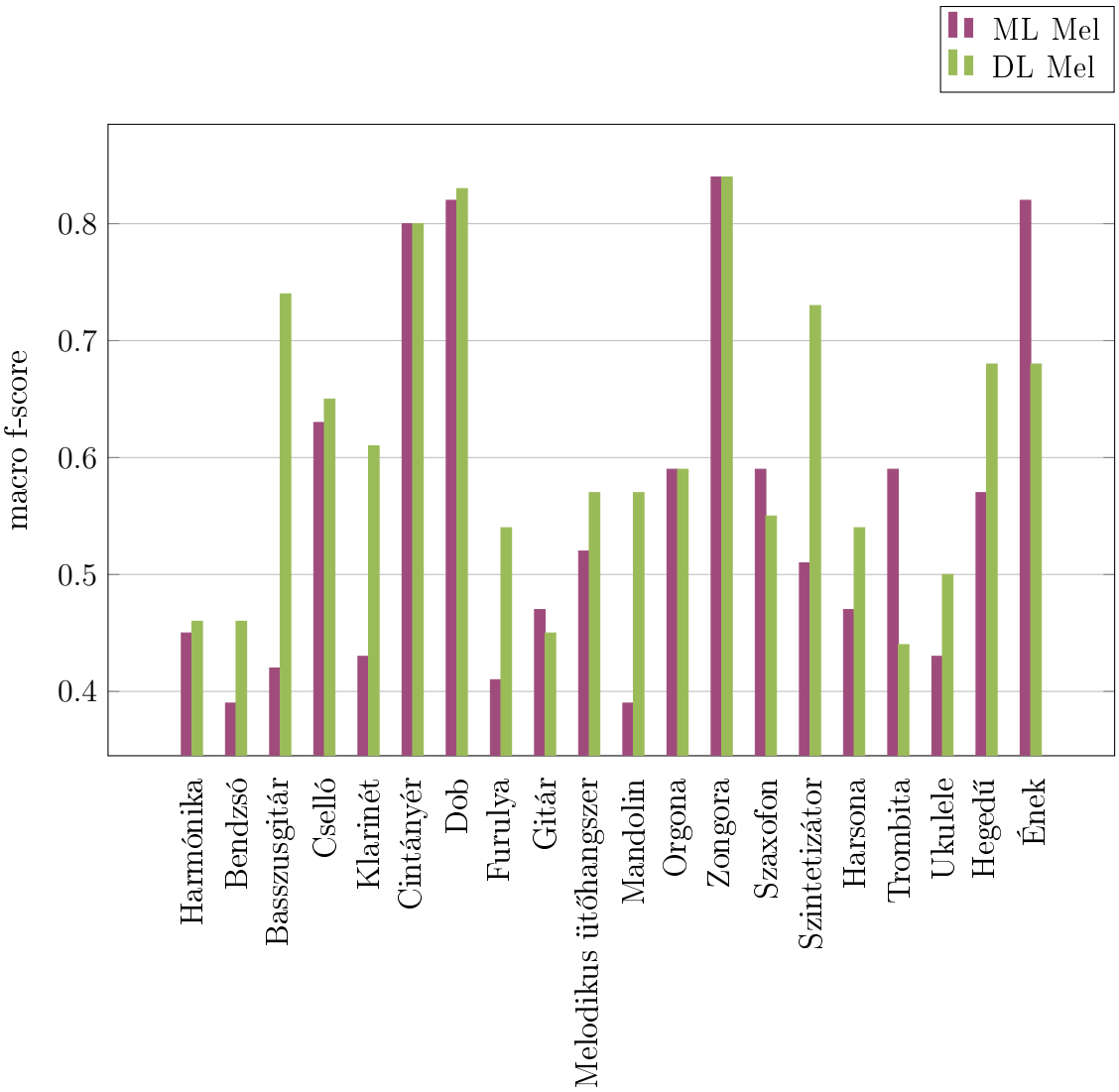
5.1. Mérőszámok

hogyan tudjuk megmérni a modell teljesítményét, pontosság, tanítás ideje, stb... A következő metrikákat vizsgáltuk:

- Pontosság (accuraccy) - a modell az adott lépésben a bemeneti adatok hány százalékára adott helyes kimenetet?
- Veszteség (loss) - a veszteségfüggvény eredménye. A modell predikcióinak a valóságtól való eltérését összeadva kapjuk meg.
- Precizitás (precision) -
- Felidézés (recall) -
- F1 érték (f1 score) - ..... Ennek a súlyozott átlaga a mérvadó.

5.2. Eredmények





5.2.1. próba1

5.2.2. próba2

5.2.3. próba3

5.2.4. próba4

5.3. Összehasonlítás

5.3.1. saját próbálkozásaim

5.3.2. más munkákkal (hagyományos ML)

6. fejezet

Összegzés, kitekintés

Irodalomjegyzék

[1] Florian Eyben és tsai. "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks." *Proceedings of the 11th International Society for Music Information Retrieval Conference* (Utrecht, Netherlands). Utrecht, Netherlands: ISMIR, 2010. aug., 589–594. old. DOI: 10.5281/zenodo.1417131. URL: https://doi.org/10.5281/zenodo.1417131.

[2] Philippe Hamel és Douglas Eck. "Learning features from music audio with deep belief networks". *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*. Utrecht, The Netherlands, 2010. aug., 339–344. old.

[3] Keunwoo Choi és tsai. "A tutorial on deep learning for music information retrieval". *arXiv preprint arXiv:1709.04396* (2017).

[4] Keith Dana Martin és Youngmoo E. Kim. "Musical instrument identification: A pattern-recognition approach". 1998.

[5] Judith Brown. "Computer Identification of Musical Instruments Using Pattern Recognition With Cepstral Coefficients as Features". *The Journal of the Acoustical Society of America* 105 (1999. ápr.), 1933–41. old. DOI: 10.1121/1.426728.

[6] A. Eronen és A. Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features". *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. 2. köt. 2000, II753–II756 vol.2.

[7] Jeremiah Deng, Christian Simmermacher és Stephen Cranefield. "A Study on Feature Analysis for Musical Instrument Classification". *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the*

IEEE Systems, Man, and Cybernetics Society 38 (2008. máj.), 429–38. old. DOI: 10.1109/TSMCB.2007.913394.

[8] Daulappa Bhalke, C. Rao és D. Bormane. "Automatic musical instrument classification using fractional fourier transform based- MFCC features and counter propagation neural network". *Journal of Intelligent Information Systems* 46 (2015. máj.). DOI: 10.1007/s10844-015-0360-9.

[9] Juan José Burred, Axel Roebel és Thomas Sikora. "Dynamic Spectral Envelope Modeling for Timbre Analysis of Musical Instrument Sounds". *Audio, Speech, and Language Processing, IEEE Transactions on* 18 (2010. ápr.), 663 –674. old. DOI: 10.1109/TASL.2009.2036300.

[10] J. Eggink és Guy Brown. "A missing feature approach to instrument identification in polyphonic music". 5. köt. 2003. nov., 49–. old. ISBN: 0-7803-7850-4. DOI: 10.1109/ICASSP.2003.1200029.

[11] Dimitrios Giannoulis és Anssi Klapuri. "Musical Instrument Recognition in Polyphonic Audio Using Missing Feature Approach". *Audio, Speech, and Language Processing, IEEE Transactions on* 21 (2013. szept.), 1805–1817. old. DOI: 10.1109/TASL.2013.2248720.

[12] Jayme Barbedo és George Tzanetakis. "Musical Instrument Classification Using Individual Partials". *Audio, Speech, and Language Processing, IEEE Transactions on* 19 (2011. febr.), 111 –122. old. DOI: 10.1109/TASL.2010.2045186.

[13] Wenxin Jiang és Zbigniew Ras. "Multi-label automatic indexing of music by cascade classifiers". *Web Intelligence and Agent Systems* 11 (2013. ápr.), 149–170. old. DOI: 10.3233/WIA-130268.

[14] Eric J. Humphrey, Juan Pablo Bello és Yann LeCun. "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." *Proceedings of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal). Porto, Portugal: ISMIR, 2012. okt., 403–408. old. DOI: 10.5281/zenodo.1415726. URL: https://doi.org/10.5281/zenodo.1415726.

[15] Keunwoo Choi és tsai. "A Tutorial on Deep Learning for Music Information Retrieval". *CoRR* abs/1709.04396 (2017). arXiv: 1709.04396. URL: http://arxiv.org/abs/1709.04396.

[16] Peter Li, Jiyuan Qian és Tian Wang. "Automatic instrument recognition in polyphonic music using convolutional neural networks". *arXiv preprint arXiv:1511.05520* (2015).

[17] Yoonchang Han, Jaehun Kim és Kyogu Lee. "Deep convolutional neural networks for predominant instrument recognition in polyphonic music". *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (2016), 208–221. old.

[18] Markus Schedl, Arthur Flexer és Julián Urbano. "The Neglected User in Music Information Retrieval Research". *J. Intell. Inf. Syst.* 41.3 (2013. dec.), 523–539. ISSN: 0925-9902. DOI: 10.1007/s10844-013-0247-6. URL: https://doi.org/10.1007/s10844-013-0247-6.

[19] Markus Schedl, Emilia Gómez és Julián Urbano. "Music Information Retrieval: Recent Developments and Applications". *Foundations and Trends in Information Retrieval* 8 (2014. jan.), 127–261. old. DOI: 10.1561/1500000042.

[20] Dalya Gartzman. *Getting to Know the Mel Spectrogram.* URL: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0. Felkeresve: 2020. 05. 09.

[21] Brian McFee és tsai. *librosa/librosa: 0.6.3.* 0.6.3. verzió. 2019. febr. DOI: 10.5281/zenodo.2564164. URL: https://doi.org/10.5281/zenodo.2564164.

[22] J. Stephen Downie. "Music information retrieval". *Annual Review of Information Science and Technology* 37.1 (2003), 295–340. old. DOI: 10.1002/aris.1440370108. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370108. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370108.

[23] G. Peeters és tsai. "The timbre toolbox: extracting audio descriptors from musical signals". *Journal of the Acoustical Society of America* 130.5 (2011), 2902–2916. old. DOI: 10.1121/1.3642604. URL: http://eprints.gla.ac.uk/68491/.

[24] Perfecto Herrera, Geoffroy Peeters és Shlomo Dubnov. "Automatic Classification of Musical Instrument Sounds". *Journal of New Music Research* 32 (2010. aug.). DOI: 10.1076/jnmr.32.1.3.16798.

[25] Anssi Klapuri és Manuel Davy. *Signal Processing Methods for Music Transcription*. 2006. jan. DOI: 10.1007/0-387-32845-9.

[26] Justin Salamon és Emilia Gómez. "Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics". *IEEE Transactions on Audio, Speech and Language Processing* 20 (2012), 1759–1770. old. URL: http://hdl.handle.net/10230/42183.

[27] Walter B Hewlett és Eleanor Selfridge-Field. *Melodic similarity: Concepts, procedures, and applications, volume 11.* 1998.

[28] Juan Bello és tsai. "A Tutorial on Onset Detection in Music Signals". *Speech and Audio Processing, IEEE Transactions on* 13 (2005. okt.), 1035 –1047. old. DOI: 10.1109/TSA.2005.851998.

[29] F. Gouyon. *Computational Rhythm Description*. VDM Verlag, 2008. ISBN: 978-3836477697.

[30] Martin F. McKinney és Dirk Moelants. "Extracting the perceptual tempo from music". *ISMIR*. 2004.

[31] Gregory H. Wakefield. "Mathematical representation of joint time-chroma distributions". *Advanced Signal Processing Algorithms, Architectures, and Implementations IX*. Szerk. Franklin T. Luk. 3807. köt. International Society for Optics és Photonics. SPIE, 1999, 637 –645. old. DOI: 10.1117/12.367679. URL: https://doi.org/10.1117/12.367679.

[32] Elaine Chew. "Towards a mathematical model of tonality". 2000.

[33] Emilia Gómez. "Tonal Description of Polyphonic Audio for Music Content Processing". *INFORMS J. on Computing* 18.3 (2006. jan.), 294–304. ISSN: 1526-5528. DOI: 10.1287/ijoc.1040.0126. URL: https://doi.org/10.1287/ijoc.1040.0126.

[34] Hélène Papadopoulos és Geoffroy Peeters. "Large-Scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM". *2007 International Workshop on Content-Based Multimedia Indexing* (2007), 53–60. old.

[35] Laurent Oudre, Yves Grenier és Cédric Févotte. "Template-based Chord Recognition : Influence of the Chord Types." *Proceedings of the 10th International Society for Music Information Retrieval Conference* (Kobe, Japan). Kobe, Japan: ISMIR, 2009. okt., 153–158. old. DOI: 10.5281/zenodo.1414884. URL: https://doi.org/10.5281/zenodo.1414884.

[36] David Temperley. "What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered". *Music Perception: An Interdisciplinary Journal* 17.1 (1999), 65–100. old. DOI: 10.2307/40285812. URL: https://doi.org/10.2307\%2F40285812.

[37] Matthew L. Cooper és Jonathan Foote. "Automatic Music Summarization via Similarity Analysis". *ISMIR*. 2002.

[38] Geoffroy Peeters, Amaury La Burthe és Xavier Rodet. "Toward Automatic Music Audio Summary Generation from Signal Analysis". *In Proc. International Conference on Music Information Retrieval*. 2002, 94–100. old.

[39] Wei Chai. "Semantic Segmentation and Summarization of Music". *IEEE Signal Processing Magazine* (2006. jan.).

[40] Dmitry Bogdanov és tsai. "From Low-Level to High-Level: Comparative Study of Music Similarity Measures". 2009. jan., 453–458. old. DOI: 10.1109/ISM.2009.72.

[41] Michael A. Casey és tsai. *Content-Based Music Information Retrieval: Current Directions and Future Challenges*. 2008.

[42] Markus Schedl és tsai. "Exploring the Music Similarity Space on the Web". *ACM Trans. Inf. Syst.* 29.3 (2011. júl.). ISSN: 1046-8188. DOI: 10.1145/1993036.1993038. URL: https://doi.org/10.1145/1993036.1993038.

[43] Joan Serrà, Emilia Gómez és Perfecto Herrera. "Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond". *Advances in Music Information Retrieval*. Szerk. Zbigniew W. Raś és Alicja A. Wieczorkowska. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, 307–332. old. ISBN: 978-3-642-11674-2. DOI: 10.1007/978-3-642-11674-2_14. URL: https://doi.org/10.1007/978-3-642-11674-2_14.

[44] Thierry Bertin-Mahieux és Daniel Ellis. "Large-scale cover song recognition using the 2D Fourier transform magnitude". (2012. jan.).

[45] Naoko Kosugi és tsai. "A practical query-by-humming system for a large music database". 2000. jan., 333–342. old. DOI: 10.1145/354384.354520.

[46] Justin Salamon, Joan Serrà és Emilia Gómez. "Tonal representations for music retrieval: From version identification to query-by-humming". *International Journal of Multimedia Information Retrieval, special issue on Hybrid Music Information Retrieval* 2 (2013. márc.), 45–58. old. DOI: 10.1007/s13735-012-0026-0.

[47] Roger Dannenberg és tsai. "A comparative evaluation of search techniques for query-by-humming using the MUSART testbed". *JASIST* 58 (2007. márc.), 687–701. old. DOI: 10.1002/asi.20532.

[48] yi-hsuan Yang és Homer Chen. "Machine Recognition of Music Emotion: A Review". *ACM Transactions on Intelligent Systems and Technology (TIST)* 3 (2012. máj.). DOI: 10.1145/2168752.2168754.

[49] Cyril Laurier és tsai. "Indexing music by mood: Design and integration of an automatic content-based annotator". *Multimedia Tools Appl.* 48 (2010. máj.), 161–184. old. DOI: 10.1007/s11042-009-0360-2.

[50] George Tzanetakis és Perry Cook. "Musical Genre Classification of Audio Signals". *IEEE Transactions on Speech and Audio Processing* 10 (2002. jan.), 293–302. old.

[51] Peter Knees, Elias Pampalk és Gerhard Widmer. "Artist Classification with Web-Based Data." 2004. jan.

[52] Perfecto Herrera, Anssi Klapuri és Manuel Davy. "Automatic Classification of Pitched Musical Instrument Sounds". 2006. jan., 163–200. old. DOI: 10.1007/0-387-32845-9_6.

[53] Youngmoo Kim és Brian Whitman. "Singer Identification in Popular Music Recordings Using Voice Coding Features". (2002. szept.).

[54] Mohamed Sordo. "Semantic annotation of music collections: A computational approach". Dissz. 2012. jan.

[55] Emanuele Coviello, Antoni Chan és Gert Lanckriet. "Time Series Models for Semantic Music Annotation". *Audio, Speech, and Language Processing, IEEE Transactions on* 19 (2011. aug.), 1343 –1359. old. DOI: 10.1109/TASL.2010.2090148.

[56] Riccardo Miotto és Nicola Orio. "A Probabilistic Model to Combine Tags and Acoustic Similarity for Music Retrieval". *ACM Transactions on Information Systems - TOIS* 30 (2012. máj.), 1–29. old. DOI: 10.1145/2180868.2180870.

[57] Xinxi Wang, David Rosenblum és Ye Wang. "Context-Aware Mobile Music Recommendation for Daily Activities". 2012. okt. DOI: 10.1145/2393347.2393368.

[58] Pedro Cano és tsai. "Audio Fingerprinting: Concepts And Applications". 2. köt. 2005. szept., 233–245. old. DOI: 10.1007/10966518_17.

[59] Michael Casey és tsai. "Content-Based Music Information Retrieval: Current Directions and Future Challenges". *Proceedings of the IEEE* 96 (2008. máj.), 668 –696. old. DOI: 10.1109/JPROC.2008.916370.

[60] Oscar Celma. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space.* 2010. jan. ISBN: 978-3-642-13286-5. DOI: 10.1007/978-3-642-13287-2.

[61] Yuan Zhang és tsai. "Auralist: Introducing serendipity into music recommendation". 2012. febr., 13–22. old. DOI: 10.1145/2124295.2124300.

[62] Marius Kaminskas, Francesco Ricci és Markus Schedl. "Location-aware music recommendation using auto-tagging and hybrid matching". 2013. okt., 17–24. old. DOI: 10.1145/2507157.2507180.

[63] Tim Pohle és tsai. ""Reinventing the Wheel": A Novel Approach to Music Player Interfaces". *Multimedia, IEEE Transactions on* 9 (2007. máj.), 567 – 575. old. DOI: 10.1109/TMM.2006.887991.

[64] Gordon Reynolds és tsai. "Towards a Personal Automatic Music Playlist Generation Algorithm: The Need for Contextual Information". (2007. jan.).

[65] Elias Pampalk, Tim Pohle és Gerhard Widmer. "Dynamic Playlist Generation Based on Skipping Behavior." 2005. jan., 634–637. old.

[66] J.-J Aucouturier és Francois Pachet. "Scaling up music playlist generation". 26. köt. 2002. febr., 105 –108 vol.1. ISBN: 0-7803-7304-9. DOI: 10.1109/ICME.2002.1035729.

[67] Dixon, Simon Lui és Gerhard Widmer. "MATCH: A Music Alignment Tool Chest". 2005. szept.

[68] Meinard Müller, Henning Mattes és Frank Kurth. "An Efficient Multiscale Approach to Audio Synchronization". 2006. okt.

[69] Bernhard Niedermayer és Gerhard Widmer. "A Multi-Pass Algorithm for Accurate Audio-to-Score Alignment". 2010. dec., 417–422. old.

[70] Markus Schedl és tsai. "What's Hot? Estimating Country-specific Artist Popularity." 2010. jan., 117–122. old.

[71] Francois Pachet és Pierre Roy. "Hit Song Science Is Not Yet a Science." 2008. jan., 355–360. old.

[72] Noam Koenigstein és Yuval Shavitt. "Song Ranking based on Piracy in Peer-to-Peer Networks." 2009. jan., 633–638. old.

[73] Meinard Müller és Nanzhu Jiang. "A scape plot representation for visualizing repetitive structures of music recordings". *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012* (2012. jan.).

[74] Arpi Mardirossian és Elaine Chew. "Visualizing Music: Tonal Progressions and Distributions." 2007. jan., 189–194. old.

[75] Matthew Cooper és tsai. "Visualization in Audio-Based Music Information Retrieval". *Computer Music Journal* 30 (2006. jún.), 42–62. old. DOI: 10.1162/comj.2006.30.2.42.

[76] Jonathan Foote. "Visualizing Music and Audio using Self-Similarity". 1999. jan., 77–80. old. DOI: 10.1145/319463.319472.

[77] Emilia Gómez és Jordi Bonada. "Tonality visualization of polyphonic audio". (2005. jan.).

[78] Sebastian Stober és Andreas Nürnberger. "MusicGalaxy: A Multi-focus Zoomable Interface for Multi-facet Exploration of Music Collections". 6684. köt. 2010. jún., 273–302. old. DOI: 10.1007/978-3-642-23126-1_18.

[79] Stefan Leitich és Martin Topf. "Globe of Music - Music Library Visualization Using Geosom." 2007. jan., 167–170. old.

[80] Paul Lamere és Douglas Eck. "Using 3D Visualizations to Explore and Discover Music." 2007. jan., 173–174. old.

[81] Elias Pampalk és Masataka Goto. "MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling." 2006. jan., 367–370. old.

[82] Rebecca Stewart és Mark Sandler. "The amblr: A mobile spatial audio music browser". 2011. júl., 1–6. old. DOI: 10.1109/ICME.2011.6012203.

[83] Markus Schedl és Dominik Schnitzer. "Hybrid retrieval approaches to geospatial music recommendation". 2013. júl., 793–796. old. DOI: 10.1145/2484028.2484146.

[84] Sebastian Stober. "Adaptive Methods for User-Centered Organization of Music Collections". Dissz. 2011. nov.

[85] Marius Kaminskas, Francesco Ricci és Markus Schedl. "Location-aware music recommendation using auto-tagging and hybrid matching". 2013. okt., 17–24. old. DOI: 10.1145/2507157.2507180.

[86] Linas Baltrunas és tsai. "InCarMusic: Context-Aware Music Recommendations in a Car". 85. köt. 2011. aug., 89–100. old. DOI: 10.1007/978-3-642-23014-1_8.

[87] Dr. Michael J. Garbade. *Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences.* URL: `https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb`. Felkeresve: 2020. 05. 04.

[88] Eric Humphrey, Simon Durand és Brian McFee. "OpenMIC-2018: An Open Data-set for Multiple Instrument Recognition." *ISMIR*. 2018, 438–444. old.