Ocenjevanje osnovnih frekvenc z uporabo kompozicionalnega hierarhičnega modela

Manca Žerovnik, Matevž Pesek, Matija Marolt

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
e-mail: mz9347@student.uni-lj.si, {matevz.pesek, matija.marolt}@fri.uni-lj.si

Osnutek

V članku predstavljamo pristop ocenjevanja osnovnih frekvenc z uporabo kompozicionalnega hierarhičnega modela. Model predstavlja alternativo drugim predstavnikom učenja globokih arhitektur. Biološko navdahnjen kompozicionalni hierarhični model je po svoji strukturi globoka arhitektura, od drugi pristopov pa se razlikuje po transparentnosti strukture, ki omogoča razumevanje struktur na posameznem nivoju. Interpretacija vseh dodatnih informacij, ki nam jih model zaradi svoje transparentnosti ponuja, je dobro izhodišče za razvoj modela. Model bi lahko deloval na mnogih opravilih s katerimi se spopadamo na področju pridobivanja informacij iz glasbe. V članku evalviramo delovanje modela za opravilo ocenjevanja osnovnih frekvenc.

1 Uvod

Področje pridobivanja informacij iz glasbe (*angl. music information retrieval* - MIR) je mlado, a hitro rastoče področje znanosti, ki se je začelo razvijati v osemdesetih letih prešnjega stoletja. Zajema kombinacijo računalništva, elektrotehnike, informatike, muzikologije, glasbe in psihologije. Kakor že ime pove, gre za pridobivanje informacij, ki se nahajajo v glasbi. Področje zaradi podprtosti z računalnikom ponuja nove razsežnosti procesiranja in uporabnosti le teh. Želi se približati človeškemu zaznavnemu sistemu in ga nadgraditi s tistimi zmogljivosti v katerih sodobna tehnologija presega človeka. MIR se ukvarja s kategoriziranjem, manipulacijo in tudi z ustvarjanjem glasbe.

Eno izmed glavnih razvijajočih se opravil na področju MIR je Samodejna transkripcija glasbe (*angl. Automatic music transcription* - AMT). Predstavljeni avtomatski pristopi so zaenkrat po zmogljivosti še vedno precej pod nivojem človeških strokovnjakov na glasbenem področju, kar ponuja veliko prostora za raziskovanje. Gre za proces samodejnega spreminjanja zvočnega signala glasbe v eno izmed oblik notnega zapisa [1]. Glavni doprinosi, ki jih ponuja AMT, so notni zapisi improvizacij in notni zapisi del pri zvrsteh, kjer se ta večinoma ne zapiše (ljudske pesmi ustnega izročila, pop, jazz ...). Hkrati omogoča hitrejši napredek drugih opravil področja MIR. Problem samodejne transkripcije lahko razdelimo na več nalog: ocenjevanje osnovnih frekvenc (*ang. multiple fundamental frequency estimation ali multi-pitch estimation*), transkripcija melodije (*ang. melody transcription*), zaznavanje pojavitve not (*ang. onset and offset detection*), ocena jakosti in kvantizacija (*ang. loudness estimation and quantisation*), prepoznavanje inštrumentov (*ang. instrument recognition*), pridobivanje ritmičnih informacij (*ang. extraction of rhythmic information*) in časovna kvantizacija (*ang. time quantisation*) [1]. Glavni problem in problem s katerim se bomo ukvarjali je problem ocenjevanja osnovnih frekvenc.

Ocenjevanje osnovnih frekvenc je proces transkripcije polifonične glasbe, ki v določenem časovnem okviru zazna tone, ki se istočasno pojavijo, in so lahko proizvedeni z različnimi inštrumenti. Problem ni enostavno rešljiv. Komponente zvočnih signalov inštrumentov v polifoničnih skladbah, se namreč časovno in frekvenčno prekrivajo. Poleg tega probleme povzročajo odmevi in prehodna stanja. Pristope k problemu lahko v grobem razdelimo v dve skupini: pristop s ponavljajočim ocenjevanjem (*ang. iterative estimation approach*) in pristop s skupnim ocenjevanjem (*ang. joint estimation approach*) [2]. Prvi pri vsakem prehodu v določenem časovnem okviru poišče prevladujočo frekvenco in slednjo ob koncu ponovitve odstrani iz spektra. Pristop s ponavljanjem je manj računsko zahteven, a tudi manj natančen zaradi neaditivnosti posameznih entitet v signalu. Drugi pristop poišče vse prevladujoče frekvence naenkrat, je večinoma natančnejši, a računsko bolj zahteven. Do danes je zaradi večje natančnosti že skoraj povsem prevladal drugi pristop skupnega ocenjevanja. Tudi pri tem pristopu lahko ločimo tehnike reševanja na zaznavanje na podlagi značilnic, zaznavanje s pomočjo statističnih modelov in zaznavanje na podlagi razcepa spektrograma

[1]. V nalogi bomo uporabili pristop s hierarhičnim kompozicionalnim modelom, ki bi ga lahko uvrstili med metode, ki opravljajo zaznavanje na podlagi razcepa spektrograma. Uporabljen pristop predstavlja alternativo pristopom, ki se problema lotevajo z globokimi arhitekturami [3, 4].

Predlagani model predstavlja celovito rešitev, ki se ne osredotoča le na eno nalogo, ampak bi v prihodnosti lahko služil za reševanje mnogih opravil na področju MIR in bi ob tem povsod ponujal rezultate, ki se lahko primerjajo z najboljšimi v svetu. V članku bomo predstavili delovanje modela na opravilu ocenjevanja osnovnih frekvenc (*angl. multiple fundamental frequency estimation* - MFFE) .

V nadaljevanju bomo v poglavju 2 najprej podrobneje predstavili uporabljen kompozicionalni hierarhični model, v poglavju 3 bomo predstavili pridobljene rezultate in jih ovrednotili. V poglavju 4 bomo predstavili smernice za nadaljnje izboljšave ter zaključke do katerih smo prišli.

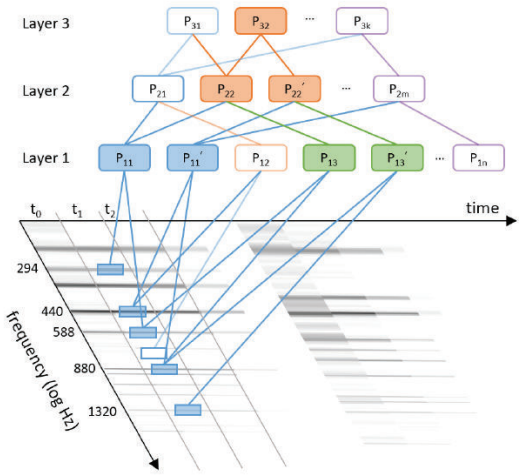
2 Kompozicionalni hierarhični model

Model, ki ga bomo uporabljali je nastal kot prevedba modela za področje računalniškega vida, ki sta ga zasnovala Leonardis in Fidler [5]. Naš model, s podobnim delovanjem, ki se ukvarja s procesiranjem zvočnih informacij, predvsem glasbe, so razvili Pesek, Leonardis in Marolt [6]. Model je podoben drugim globokim arhitekturam, ki temeljijo na pristopu z nevronskimi mrežami, vendar se od slednjih razlikuje po svoji transparentnosti delovanja strukture. Slednje bi se lahko izkazalo kot prednost pri opravilih za pridobivanje informacij iz glasbe.

Model je po principu globokih arhitektur zgrajen iz večih nivojev, ki jih sestavljajo posamezni gradniki - *deli*. Kompleksnost informacije posameznih delov je zaradi hierarhične zgradbe manjša na nižjih nivojih in večja na višjih. Vsak del je, razen na začetnem nivoju, zgrajen iz delov nižjih nivojev. Spodnji ali vhodni nivo predstavlja vhodni zvočni signal transformiran v časovno frekvenčno domeno. Ta nivo označujemo z $\mathcal{L}_0$ in ga imenujemo tudi "ničti" nivo hierarhije. Sestavljen je iz gradnikov, ki predstavljajo vse kanale frekvenčnega spektra signala v izbranem časovnem okvirju.

Deli so kompozicije entitet, ki opisujejo signal. Del lahko opisuje posamezne frekvence v signalu, njihove kombinacije, poleg tega pa tudi tone, akorde in časovne vzorce [6]. Posameznim delom pripišemo aktivacijo, ki jo definirata lokacija $\mathcal{L}_p$ in magnituda $\mathcal{A}_p$. Lokacija predstavlja frekvenco, magnituda pa moč aktivacije. Glede na moč aktivacije se na podlagi izbrane pragovne vrednosti določijo deli, ki so aktivni.

Shematičen prikaz zgradbe modela je prikazan na sliki 1. Dele na višjih nivojih, imenovane tudi



Slika 1: Slika prikazuje zgradbo kompozicionalnega hierarhičnega modela. Abscisa predstavlja čas. Na sliki je prikazan model zgrajen iz treh nivojev označenih z Layer 1, Layer 2 in Layer 3 ter vhodnega nivoja v časovnem okvirju t1. Obarvani deli predstavljajo aktivne dele na vsakem izmed nivojev. Vir slike: [6].

kompozicije, lahko definiramo kot sestav osrednjega in sekundarnih sestavnih delov z nižjega nivoja. Naslednji nivoji modela $\mathcal{L}_n$ so torej sestavljeni iz kompozicij, ki so zgrajene iz delov nižjega nivoja $\mathcal{L}_{n-1}$. Vsak del je lahko gradnik poljubnega števila kompozicij višjega nivoja in prav tako lahko vsako kompozicijo sestavlja poljubno število delov nižjega nivoja.

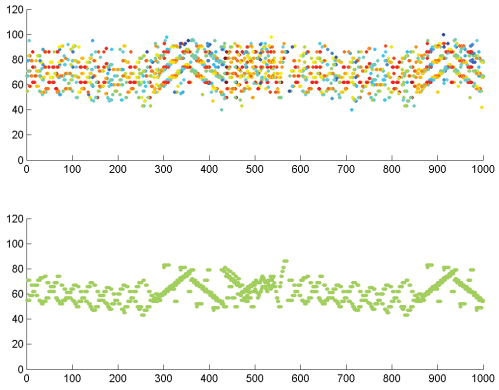
Jakost aktivacije $\mathcal{A}_p$ in lokacijo $\mathcal{L}_p$ določene kompozicije izračunamo glede na vrednosti aktivacij njenih sestavnih delov, kar je bolj natančno opisano v [6]. Kompozicija je aktivna, kadar so aktivni vsi njeni sestavni deli na nižjih nivojih. V vsakem časovnem okvirju nam model vrača množico aktivnih delov iz katerih lahko razberemo razne informacije o procesiranem signalu.

Model za izgradnjo uporablja nenadzorovano učenje. Proces učenja na podlagi statistične metode sestavlja kompozicije iz sopojavitev delov, ki so med seboj čimbolj disjunktni in se pogosto aktivirajo na podobnih razdaljah. Za zmanjšanje števila delov se na koncu uporabi še požrešen algoritem.

Model za povečanje robustnosti uporablja tri biološko navdahnjene mehanizme halucinacijo, inhibicijo ter mehanizem za samodejno uravnavanje jakosti (*ang. automatic gain control* - *AGC*).

2.1 Mehanizmi

Inhibicija je mehanizem, ki posnema delovanje človeškega avditornega sistem in odstranjuje odvečne informacije ter zmanjšuje šum, ki je lahko prisoten v



Slika 2: Na sliki je primer delovanja modela pri ocenjevanju osnovnih frekvenc. Na zgornjem grafu so z vizualizacijo predstavljene hipoteze našega modela, s tem da smo prikazali enako število hipotez na časovni okvir kakor jih je prisotnih v osnovnem posnetku. Vzeli smo tiste hipoteze z najmočnejšo aktivacijo. Spodnji graf pa prikazuje ročno transkribiran posnetek. Pri prikazu rezultata modela smo z vizualizacijo v jet barvnem prostoru prikazali tudi moč aktivacije posamezne hipoteze, za samo skladbo - transkripcijo - informacija o jakosti žal ni bila na voljo. Pri obeh grafih nam abscisa predstavlja časovne okvirje, ordinata pa MIDI vrednosti označene točke.

signalu. Glede na moč najmočnejših hipotez mehanizem izbere tiste, ki imajo v primerjavi z njimi zelo nizko jakost aktivacije in jih odstrani.

Vloga halucinacije je, da nadomesti manjkajočo informacijo, ki lahko nastane zaradi različnih razlogov, predvsem napak v signalu. Halucinacija ponazarja človeško percepcijo, ki glede na vsebino informacije, logično dopolni pomanjkljivosti. Mehanizem glede na ostalo vsebino lahko delu, ki sicer ni aktiven, določi moč aktivacije.

Tudi mehanizem za samodejno uravnavanje jakosti odpravlja nepravilnosti, ki se pojavljajo v signalu. Z upoštevanjem časovne komponente glede na hipoteze prejšnjih časovnih okvirov s poudarjanjem pojavitev in stabilizacijo aktivacij ob manjših nihanjih uravnava aktivacije.

3 Ovrednotenje rezultatov

Za ocenjevanje osnovnih frekvenc smo zgradili model, kjer je na ničtem nivoju s konstantno Q transformacijo signal pretvorjen v 345 frekvenčnih kanalov med 55 in 8000 Hz. Časovni okvir je dolg 50 ms. Naučili smo dva nivoja $\mathcal{L}_1$ in $\mathcal{L}_2$, ki vsebujeta 23 in 12 delov. Za ocenjevanje osnovnih frekvenc smo uporabili nivo $\mathcal{L}_2$. S prilagajanjem parametrov vpliva inhibicije, halucinacije in mehanizma za samodejno uravnavanje jakosti smo naučili različne hierarhije in poiskali najboljšo kombinacijo za najoptimalnejše delovanje modela. Informacija, ki jo pridobimo iz modela, je neposredna in ne rabi dodatnega procesiranja. Primer delovanja modela je na primeru krajše skladbe prikazan na sliki 2.

Model smo za ovrednotenje in primerjavo rezultatov s trenutnim stanjem istega opravila na področju MIR preizkusili na prostodostopni zbirki MAPS (MIDI Aligned Piano Sounds) [7], ki vsebuje klavirske posnetke, MIDI zapise teh posnetkov ter tekstovne datoteke z informacijami o posnetkih. Zbirka MAPS je razdeljena na več delov, glede na vrsto klavirja s katerim so proizvedene skladbe. Vsak del vsebuje posnetke posameznih tonov, akordov, naključnih kombinacij tonov in skladb. Mi smo model testirali na skladbah vsakega dela, ki so v vsakem delu zbirke zbrane v direktoriju MUS.

Model smo ovrednotili s tremi merami: točnost razvrščanja (*ang. classification accuracy - CA*), natančnost (*ang. precission - PRE*) in priklic (*ang. recall - REC*) in smo jih izračunali za vsako skladbo posebej. Za vsako skladbo smo naredili tri izračune. Prvi je bil normalen izračun točnosti glede na pridobljene hipoteze, drugi, ki ga v tabeli 1 označujemo s pripono 12 ob imenu dela zbirke, spregleda oktavne napake navzgor in navzdol. Ob osnovnem tonu, ki ga proizvede nek vir, se namreč ponavadi vedno pojavijo višji harmoniki, to so toni, ki zvoku določajo višino tona in barvo. Človeško uho vse skupaj zazna kakor en ton, umetno zaznavanje pa višje harmonike težko loči od osnovnega tona. Prvi harmonik ima dvakrat višjo frekvenco od osnovnega tona. To razdaljo v glasbenem jeziku imenujemo oktava, napačne hipoteze, ki so od pravilnih oddaljene za oktavo smo zato poimenovali oktavne napake. Tretji način računananja točnosti, ki smo ga uporabili, spregleda vse oktavne napake in njihove večkratnike. Ta način primerja le tonske razrede. V tabeli 1 so rezultati tretjega načina prikazani v vrstici, kjer je imenu dela zbirke dodana pripona mod. Za MIDI vrednosti hipotez in transkribiranega posnetka, smo namreč povsod vzeli originalne vrednosti po modulu 12. Model smo preizkusili na vseh devetih delih zbirke MAPS.

Neposredna primerjava nakazuje, da so predstavljeni rezultati slabši od trenutno najboljših na tem področju. Wenninger [8] predlaga metodo, ki na isti bazi deluje s povprečno klasifikacijsko točnostjo 77.1 %, Böck [9] pa predlaga metodo, ki ima povprečno klasifikacijsko natančnost 68.7 %. Nam [4] predstavi metodo z DBN, ki ima oceno F na isti bazi 74.4 %, vendar je rešitev namenjena zgolj transkripciji klavirske glasbe.

Naša rešitev zaenkrat ne ponuja primerljivih rezultatov, vendar se pristop zelo razlikuje od ostalih

Tabela 1: Rezultati prvih dveh delov zbirke, ki so izračunani kot povprečje vseh skladb posameznega dela. Rezultate smo pridobivali na celotnih skladbah in odražajo trenutno delovanje modela. Za vsak del smo izračunali natančnost na tri načine, ki so opisani v članku.

Ime dela zbirke	CA	PRE	REC
AkPnBcht	0.5797	0.2243	0.5682
AkPnBcht_12	0.7026	0.4263	0.6824
AkPnBcht_mod	0.8753	0.3185	0.6799
AkPnBsdf	0.6321	0.2193	0.6337
AkPnBsdf_12	0.7267	0.5010	0.7078
AkPnBsdf_mod	0.8929	0.3830	0.7219

in veliko obeta. Kompozicionalni hierarhični model deluje obenem tudi kakor klasifikator, kar želimo v prihodnosti nadomestiti z metodo podpornih vektorjev (*ang. support vector machine - SVM*). Predvidevamo, da bi rezultate z dodatnim postprocesiranjem lahko izboljšali. Iz rezultatov lahko sklepamo, da bi potrebovali tudi mehanizem za ustrezno odstranitev oktavnih napak, za katere smo ob evalvaciji ugotovili, da se pojavljajo redno in pogosto z enako močjo aktivacije kakor osnovna frekvenca, zaradi česar jih je pri razvrščanju težko odstraniti. Ob koncu moramo poudariti, da je bil uporabljen model naučen na zelo majhni bazi posnetkov klavirskih tipk in ni za učenje uporabljal nobenih delov podatkovne zbirke MAPS, medtem ko je [8], ki je dosegel najboljši rezultat izmed zgoraj omenjenih, za učenje uporabil približno 80 odstotkov osnovne baze, ki jo predstavljajo skladbe zbirke MAPS na kateri je bil model testiran.

4 Nadaljnje delo in zaključek

Kompozicionalni hierarhični model je bil preoblikovan tako, da smo pridobili hipoteze, ki ocenjujejo frekvence, ki se pojavljajo v določenem časovnem okvirju vhodnega zvoka. Delovanje modela smo ovrednotili in primerjali z rezultati, ki jih na isti podatkovni zbirki, dosegajo drugačne metode. Naši rezultati niso primerljivi z najboljšimi, so pa obetavni in predvidevamo, da jih bomo z dodatnimi nadgradnjami delovanja izboljšali. Trenutno se ukvarjamo z interpretacijo napak v klasifikaciji in iskanjem algoritmov, ki bi te lahko odpravili. V prihodnosti bomo najprej implementirali ne-negativno faktorizacijo matrik in metodo podpornih vektorjev, kar bi z boljšo klasifikacijo utegnilo izboljšati natančnost hipotez. Potem bomo z istimi metodami poskusili izvesti transkripcijo melodije, tako da bomo na obstoječih hipotezah uporabili strojno učenje. Model si želimo približati delovanju človeškega avditornega sistema.

Literatura

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[2] YEH Chunghsin. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, Université Lille 1, 2008.

[3] Eric J Humphrey, Juan P Bello, and Yann LeCun. Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

[4] Juhan Nam, Jiquan Ngiam, Honglak Lee, and Malcolm Slaney. A classification-based polyphonic piano transcription approach using learned feature representations. In *ISMIR*, pages 175–180, 2011.

[5] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[6] Matevž Pesek, Aleš Leonardis, and Matija Marolt. Compositional Hierarchical Model for Music Information Retrieval. *ISMIR*, 2014.

[7] Emiya V., Badeau R., and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. 2010.

[8] F. Weninger, C. Kirst, B. Schuller, and H.-J. Bungartz. A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6–10, May 2013.

[9] S Bock and Markus Schedl. Polyphonic piano note transcription with recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 121–124. IEEE, 2012.