



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Jiří Balhar

Extrakce melodie pomocí hlubokého učení

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Jan Hajič

Studijní program: Informatika

Studijní obor: Programování a softwarové systémy

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Poděkování.

Název práce: Extrakce melodie pomocí hlubokého učení

Autor: Jiří Balhar

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Jan Hajič, ústav

Abstrakt: Abstrakt.

Klíčová slova: klíčová slova

Title: Melody Extraction using Deep Learning

Author: Jiří Balhar

Institute: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Jan Hajič, institute

Abstract: Abstract.

Keywords: key words

Obsah

Úvod	2
1 Datasetsy	3
1.1 MedleyDB	4
1.2 MDB-synth	4
1.3 Orchset	5
1.4 Weimar Jazz Database	6
Závěr	7
Seznam použité literatury	8
Seznam obrázků	9
Seznam tabulek	10
Seznam použitých zkratek	11
Přílohy	12

Úvod

Následuje několik ukázkových kapitol, které doporučují, jak by se měla bakalářská práce sázet. Primárně popisují použití T_EXové šablony, ale obecné rady poslouží dobře i uživatelům jiných systémů.

1. Datasets

Nedostupnost dostatečného množství dat pro automatickou transkripci melodie představuje zejména pro metody strojového učení otevřený problém. Zatímco pro vzdáleně příbuznou úlohu automatického přepisu mluveného slova existuje tisíce hodin nahrávek (například dataset LibriSpeech, který vznikl na základě audioknih), největší dataset s přepsanou melodickou linkou MedleyDB má celkovou délku pod osm hodin. Do roku 2014, kdy MedleyDB vznikl, existovaly datasety, které byly buď rozmanité, ale příliš krátké (ADC04, MIREX05, INDIAN08) nebo naopak celkově větší, ale žánrově a hudebně homogenní (MIREX09, MIR1K, RWC). V roce 2015 byl vydán dataset Orchset, který obsahuje 23 minut výňatků z orchestrálních skladeb různých období. Za dataset pro extrakci melodie by se také dal považovat Weimar Jazz Database, který je sice primárně zaměřený na využití v muzikologii, nicméně obsahuje přes 450 přepsaných jazzových sol. Novinkou z roku 2017 je vydání datasetu MDB-melody-synth, který byl automaticky vygenerován základě vstupní vícestopé hudby (převzaté z MedleyDB), existuje tedy naděje, že současný korpus pro přepis melodie by se mohl v budoucnu rozšířit o velkou část veřejně dostupných vícestopých nahrávek.

Co se týče blízké úlohy transkripce hudby, velikost největších datasetů se pohybuje v řádu desítek hodin, tudíž jde stále o omezené kolekce. Mezi největší se řadí MusicNet (orchestrální, 34 hodin), MAPS (klavír, 18 hodin), MDB-mf0-synth (multižánrový, 4,7 hodin), GuitarSet (kytara, 3 hodiny) a URMP (komorní orchestr, 1,3 hodiny). I když jde o úlohu, která je lépe definovaná (na rozdíl od extrakce melodie v celkovém přepisu nehraje takovou roli subjektivita při volbě hlavního hlasu), s použitím polyfonních nástrojů vyvstává problém náročné ruční anotace.

Vytváření nových datasetů je obecně velmi pracné a nákladné. Obvyklý postup totiž zahrnuje buď kompletní ruční přepis nahrávky nebo alespoň ruční opravu výstupu automatického přepisu, přičemž tuto práci odvedou nejkvalitněji pouze zaškolení hudebníci. Každá vzniklá anotace by se také měla překontrolovat, a to nejlépe jiným hudebníkem. Dalším problémem je vůbec identifikace melodie - jelikož je určení hlavní melodické linie subjektivní, musí se na výsledné anotaci shodnout co nejvíce posluchačů, ve výsledku se proto vybírají takové nahrávky, které nejsou sporné. S tím souvisí také zavedení a pečlivé dodržování anotační politiky u komplexnějších skladeb (zejména orchestrálních), kde může melodii nést více hlasů zároveň (současně či střídajíc se). Také množství výchozích dat pro vznik datasetů není velké, jednak musí být skladby širitelné, pokud má být dataset volně dostupný a jednak by k nim měly být dostupné audiostopy, ze kterých je smíchán finální mix, jelikož ruční anotace pouze s pomocí finálního mixu je mnohem náročnější než anotace stop.

Existence dostatečně velkého množství dat je obecně vzato zásadním předpokladem pro využití metod strojového učení pomocí hlubokých neuronových sítí, zejména pak pro netriviální úlohy, jakou je například přepis melodie, jelikož dovoluje zvětšení celkové kapacity modelu, aniž by docházelo k overfittingu. Také pro evaluaci metod, například i v soutěži MIREX, jsou potřeba takové datasety, které dobře reprezentují reálná data, přitom MedleyDB vzniklo mimo jiné z důvodu, že stávající datasety nestačily ani pro tento účel.

Možností řešení nastíněného problému nedostatku dat je více. Příným řešením by byl návrh metody, která by celý proces vzniku datasetů výrazně ulehčila. O to se snaží článek Salamon a kol. (2017) a princip této metody popisuje kapitola XXX.

1.1 MedleyDB

Bittner a kol. (2014)

Multimodální, vícestopý dataset obsahující 122 nahrávek, k 108 z nich je dostupná anotace melodie. Kromě té obsahuje také metadata o všech písních s informacemi o žánru a instrumentalizaci. S celkovou délkou 7.3 hodiny jde o nejdelší dataset, který se zaměřuje na více hudebních žánrů. O rozmanitosti svědčí i to, že se v datasetu vyskytuje řada nástrojů mimoevropského původu, a že jen přibližně polovina písní obsahuje zpěv. Na rozdíl od ostatních datasetů jsou nahrávky ve většině případů celé písně, tedy nejde pouze o krátké výňatky, a ke každé jsou poskytnuty audiostopy, ze kterých je vytvořen výsledný mix. Na základě diskuze, kterou shrnuji v kapitole o definici melodie, autoři poskytují tři verze anotací, na základě různých obecných definic:

1. Základní frekvence nejvýraznějšího melodického hlasu, jehož zdroj zůstává po dobu nahrávky neměnný.¹
2. Základní frekvence nejvýraznějšího melodického hlasu, jehož zdroje se mohou měnit.
3. Základní frekvence všech melodických hlasů, potenciálně pocházejících z více zdrojů.

Ačkoli třetí definice umožňuje, aby v anotaci znělo více melodických linek zároveň, v datasetu se nejedná o kompletní přepis nahrávek (použitelný pro úlohu *multif0 estimation*), ten autoři neposkytují.

Dataset vznikl obvyklou cestou ruční anotace, ze shromážděného vícestopého materiálu byly vybrány stopy s potenciálním výskytem melodie, stopy s přeslechem byly filtrovány pomocí source-separation algoritmu s ručně doladěnými parametry pro každou jednotlivou stopu, následně byl na monofonní stopy spuštěn pitch-tracker pYIN a výsledné automaticky získané anotace opravilo a vzájemně zkontrolovalo pět anotátorů s hudebním vzděláním.

1.2 MDB-synth

Hlavním přínosem práce Salamon a kol. (2017) je navržení způsobu anotace základní frekvence monofonních audiostop takovým způsobem, že výsledná dvojice zvukové stopy a anotace nevyžaduje další manuální kontrolu. Anotace stopy probíhá ve dvou krocích, nejprve získáme pomocí libovolného monopitch trackeru křivku základní frekvence a poté na základě této křivky, která může obsahovat chybné úseky, syntetizujeme novou stopu, která zachovává barvu nahrávky, ale

¹Tato definice je shodná pro evaluační datasety používané v soutěži MIREX, s výjimkou Orchsetu

výšku tónu určuje právě tato anotace. Díky tomu je pak přesnost anotace pro tuto novou, syntetickou nahrávku stoprocentní, přitom (v ideálním případě) neztrácí charakteristiky původní nahrávky.

Pro vytváření datasetu je toto významné zjednodušení, protože tím algoritmus odstraňuje časově nejnáročnější část práce - ruční kontrolu anotací audiostop. Pokud by se ukázalo, že syntéza významně neubírá na kvalitě dat, použitím navrhované metody by mohlo vzniknout velké množství nových dat (například repozitář Open Multitrack Testbed obsahuje stovky vícestopých nahrávek, které by šlo využít). Autoři v článku provádí kvantitativní analýzu pomocí srovnání state-of-the-art algoritmů pro extrakci melodie a prokazují, že výsledky těchto metod na syntetických datech se významně neliší od výsledků na původních, tím je podle autorů potvrzená možnost použití dat jak pro trénování tak pro evaluaci nových metod.

Metoda má ale bohužel svá omezení, mezi ty zásadní patří, že se dá aplikovat pouze na stopy, které obsahují monofonní signál, vstupní data tedy nesmí obsahovat přeslech a nahrávaný nástroj může hrát pouze jednohlas, v důsledku nelze zpracovat klavír či kytara, které hrají zpravidla vícehlas. To nevadí tolik u generování datasetu pro přepis melodie, jelikož melodii často hraje jeden hlas a doprovod hrají ostatní, velkým nedostatkem je toto spíše pro generování multif0 datasetů.

Dále k článku není zveřejněná kompletní referenční implementace algoritmu, tudíž algoritmus nelze snadno spustit na nových datech. Ve výsledku je tudíž největším praktickým přínosem nová sada syntetických datasetů pro úlohy přepisu melodie, basy, monofonních stop a kompletní partitury, každý dataset obsahuje destičky nahrávek. Vícestopá data použitá pro syntézu byla převzata z MedleyDB, tudíž ve výsledku nové datasety nerozšiřují celkový hudební záběr, pouze zpřesňují ten již existující.

TODO obrázek? Porovnání spektrogramů syntetické a původní nahrávky

Z kvalitativního pohledu je na výstupních syntetických nahrávkách poznat, že jsou syntetické. Autoři sice prokazují, že současné metody na těchto datech dosahují stejných výsledků, nicméně v článku chybí diskuse o tom, zda-li v datech algoritmus nevytváří nové umělé artefakty, které by mohly zneužít metody strojového učení pro spolehlivější výsledky (které by však negeneralizovaly na reálná data). Při pohledu na spektrogram je například zřejmé, že syntetická nahrávka obsahuje mnohem více výrazných alikvótních frekvencí

1.3 Orchset

Dataset vytvořený týmem Bosch a kol. (2016) orientovaný na orchestrální repertoár pocházející z různých historických období včetně 20. století. Obsahuje 64 výňatků délky od 10 do 32 sekund. Výňatky byly vybírány tak, aby obsahovaly zřejmou melodii, dataset tedy obsahuje v porovnání málo pasáží bez melodie (6% z celkové délky). Vzhledem k komplexitě uvažovaných žánrů autoři vycházejí z kombinace rozšířené definice melodie podle Bittner a kol. (2014) a definice Poliner a kol. (2007). Melodii ve výňatcích proto zpravidla nese více hudebních nástrojů (nebo celých sekcí), které se v průběhu střídají, případně mohou části hrát společně v rozdílných oktávách (nebo jiných intervalech, tvoříce tak harmonický doprovod).

Pro zjištění melodie se v takto vrstveném materiálu autoři uchylují k úplnému základu definice melodie (Poliner) a nechávají si skupinou čtyř posluchačů výňatky přezpívat. Tato hrubá data pak autoři sumarizují a odebírají z data-setu ty výňatky, na jejichž melodii se posluchači neshodli. Přezpívané tóny bylo nutné ručně opravit, aby načasováním přesně seděly na výňatek. Lidský hlas také samozřejmě nemá rozsah plného orchestru, proto bylo dalším krokem transponovat anotace tak, aby zněly ve správných oktávách. Zde se opět může vyskytnout problém subjektivity, pokud melodii hrají dva různé nástroje, pouze v jiných oktávách, pak je sporné, který nástroj označit jako hlavní (v některých případech taková otázka ani nedává příliš smysl.). Částečným řešením je zvolit libovolnou anotační politiku a tu konzistentně dodržovat (žádná společná v komunitě MIR neexistuje), v případě Orchsetu byla snaha minimalizovat skoky v melodické kontuře, což zároveň respektuje obecné pozorování, že v melodii se vyskytují mnohem častěji malé skoky (nejčastěji prima a malá/velká sekunda) než větší. Tedy například pokud pasáží hrané ve dvou různých oktávách předcházela pasáž hraná v jedné, anotace obou pasáží lze transponovat do společné oktávy tak, abychom na rozhraní minimalizovali skok v anotaci.

Dataset obsahuje pouze hrubé anotace tónů melodie, nikoli přesnou základní frekvenci nástroje, který v danou chvíli melodii hraje. Článek o tomto rozhodnutí příliš nediskutuje, vychází ale opět logicky z volby dat. U orchestrálních dat je tento abstraktnější pojem melodie mnohem méně sporný. Pokud hraje melodii sekce nástrojů v unisonu, přesná základní frekvence není dobře definovaná, jelikož se základní frekvence znějících hlasů vzájemně překrývají.

1.4 Weimar Jazz Database

Weimar Jazz Database Pfeiderer a kol. obsahuje přes 450 transkripcí jazzových sol ze všech období vývoje jazzu. Data původně zamýšlená pro muzikologické studie využívající statistické metody ale lze využít i pro potřeby extrakce melodie, jelikož uvažované nahrávky spadají zřejmě pod nejrestriktivnější definici melodie (definici používanou v soutěži MIREX) - melodii nese jistě právě jeden, sólový nástroj, a po celou dobu výňatku je jistě nejvýraznější. Výběr sólových nástrojů se omezuje pouze na jednohlasé, jelikož ruční anotace vícehlasých je příliš obtížná. Hlavním problémem při využívání je restriktivní licence, která platí na nahrávky, tudíž zdrojové audio, na základě kterého anotace vznikaly, není veřejně přístupné. Jelikož pro data neexistují jednotlivé stopy, ruční anotace probíhala přímo z finální nahrávky, což je obtížný úkol -

Závěr

Seznam použité literatury

- BITTNER, R., SALAMON, J., TIERNEY, M., MAUCH, M., CANNAM, C. a BELLO, J. (2014). MedleyDB: A multitrack dataset for annotation - intensive mir research. *International Society for Music Information Retrieval Conference*.
- BOSCH, J. J., MARXER, R. a GÓMEZ, E. (2016). Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, **45** (2), 101–117. ISSN 17445027. doi: 10.1080/09298215.2016.1182191. URL https://repositori.upf.edu/bitstream/handle/10230/26985/Bosch{}_NewMusic{}_Eval.pdf?sequence=1&isAllowed=y.
- PFLEIDERER, M., FRIELER, K., ABESSER, J., ZADDACH, W.-G. a BURKHART, B. *Inside the Jazzomat*. ISBN 9783959831246. URL http://schott-campus.com/wp-content/uploads/2017/11/inside{}_the{}_jazzomat{}_final{}_rev{}_oa4.pdf.
- POLINER, G. E., MEMBER, S., ELLIS, D. P. W., MEMBER, S., EHMANN, A. F., GÓMEZ, E., STREICH, S. a ONG, B. (2007). Melody Transcription From Music Audio : Approaches and Evaluation. *Ieee Transactions on Audio, Speech, and Language Processing*, **15**(4), 1247–1256. doi: 10.1109/TASL.2006.889797. URL <https://academiccommons.columbia.edu/doi/10.7916/D8NC69RK/download>.
- SALAMON, J., BITTNER, R. M., BONADA, J., BOSCH, J. J., GOMEZ, E. a JUAN PABLO BELLO (2017). An Analysis/Synthesis Framework for Automatic F0 Annotation of Multitrack Datasets. *Proceedings of the International Society for Music Information Retrieval {(ISMIR)} Conference*, pages 71–78.

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

Přílohy