

Detekcia Akustického Prostredia z Reči

Matúš Dobrotka*



Abstrakt

Téma tejto práce je klasifikácia audio nahrávky do 15 tried. Článok popisuje 2 metódy založené na GMM a i-vectoroch a ich vzájomnú fúziu. Na dátach zo súťaže DCASE dosiahol najlepší GMM systém úspešnosť 59% a i-vector systém 68%. Fúzia týchto dvoch systémov výsledok ešte zlepšila na 69%, čo by v dobe súťaže stačilo na 20. miesto z 97 odovzdaných systémov z celého sveta.

Kľúčové slová: detekcia akustického prostredia — GMM — i-vector — fúzia

Priložené materiály: N/A

*xdobro12@stud.fit.vutbr.cz, Faculty of Information Technology, Brno University of Technology

1. Úvod

Zvuk nesie obrovské množstvo informácie o našom každodennom prostredí a tiež o fyzikálnych udalostiach, ktoré sa v ňom uskutočňujú. Môžeme vnímať prostredie, v ktorom sa nachádzame (napríklad preplnená ulica, kancelária,...) a tiež rozpoznávať jednotlivé zdroje zvukov (auto prechádzajúce okolo, kroky ľudí, apod.).

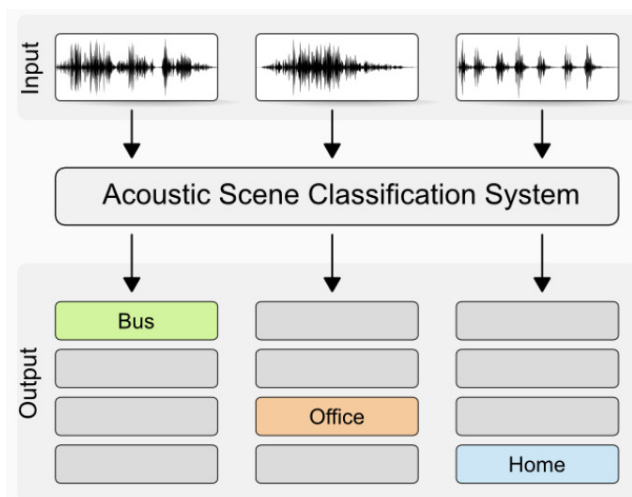
Vývoj metód spracovania signálu, ktoré sú schopné automaticky extrahovať tieto informácie, má preto obrovský potenciál pre rôzne aplikácie. Môže sa jednať napríklad o vyhľadávanie multimédií založené na ich audio obsahu, vytváranie kontextových mobilných zariadení, robotov, automobily,... a tiež inteligentné monitorovacie systémy rozpoznávajúce činnosti v ich okolí využitím akustických informácií. Vzhľadom ale k tomu, že táto oblasť ešte nie je na takej úrovni, že by bolo možné spoľahlivo rozpoznávať zvukové prostredia a jednotlivé zdroje zvuku v realistickej zvukovej palete, kde počuť viaceré zvuky a často súbežne a tiež skreslené prostredím, je ešte potrebné značné množstvo výskumu. Aj preto som sa rozhodol venovať práve tejto téme. Téma vychádza zo zadania medzinárodnej

súťaže *DCASE Challenge*, ktorá je popísaná v nasledujúcej kapitole.

Cieľom tejto práce je vytvoriť systém schopný klasifikovať testovaciu nahrávku do jednej z 15 preddefinovaných tried, ktoré charakterizujú prostredie, kde bola nahrávka nahratá. Výstižne takýto systém popisuje obrázok 1.

Jedná sa o týchto 15 prostredí:

- Autobus
- Kaviareň/reštaurácia
- Auto
- Centrum mesta
- Lesná cesta
- Obchod s potravinami - obchod strednej veľkosti
- Obydlie/dom
- Pláž pri jazere
- Knížnica
- Metro stanica
- Kancelária - viacero osôb, typický pracovný deň
- Sídliisko/obytná oblasť
- Vlák
- Električka
- Mestský park



Obrázok 1. Prehľad systému pre klasifikáciu akustického prostredia [1]

2. DCASE challenge

Týmto pojmom sa označuje celosvetová súťaž organizovaná fínskou výskumnou skupinou zameranou na audio z Technickej Univerzity v Tampere. Úplný názov súťaže je *Detection and Classification of Acoustic Scenes and Events*. Vznikla už v roku 2013 s cieľom zvýšiť vývoj v oblasti analýzy zvukov z prostredí, v ktorých sa bežne nachádzame. A to porovnávaním rozličných prístupov za použitia spoločného verejne dostupného datasetu a rovnakej vyhodnocovacej metriky.

Po úspešnom úvodnom ročníku došlo v tejto súťaži k dvojročnej prestávke a v roku 2016 organizátori znovu obnovili jej priebeh. Tretí a zároveň zatiaľ posledný ročník sa uskutočnil v roku 2017. Súťaž ako taká obsahuje 4 úlohy a zúčastniť sa jej môže ktokoľvek bez ohľadu na to, či sa rozhodne riešiť iba jednu úlohu, alebo viacero z nich. Každá z týchto úloh je nezávislá, všetky úlohy sú vyhodnocované samostatne.

Táto práca je inšpirovaná prvou z úloh, ktorej názov je *Detekcia akustického prostredia*, pričom som sa sústredil na ročník 2017. Obrázok 2 ukazuje kompletné výsledky zúčastnených tímov a ich systémov. Môžeme vidieť, že baseline systém, ktorý bol uvoľnený autormi súťaže, dosahoval úspešnosť na evaluačných dátach 61% a najlepší systém až 83%.

2.1 Prehľad výskumu

Účastníci súťaže z roku 2017 vo svojich článkoch opisujú metódy a postupy, ktorými sa rozhodli túto úlohu riešiť. Pre aktuálny prehľad uvádzam najlepšie z nich.

Na čele tejto súťaže stojí tím prevažne kórejských autorov, ktorých finálny systém pozostáva z fúzie 2

podsystemov. Podrobne to popisujú vo svojom článku [3]. Hlavnou myšlienkou ich riešenia bolo rozšírenie dátovej sady pomocou metódy *Generative Adversarial Networks (GAN)*, s ktorou celý systém natrénovali. Ďalej použili *Support Vector Machine (SVM)* pre selekciu príznakov. Ako klasifikátor zvolili v prvom prípade *Fully Connected Neural Network (FCNN)* a v druhom prípade SVM. Oba tieto systémy sa podieľali na výslednej fúzii systémov, ktorou dosiahli úspešnosť 83.3%.

Konvolučné neurónové siete sa rozhodli použiť Y.Han, J.Park a K.Lee, pričom vo svojom článku [4] uvádzajú, aké rôzne metódy predspracovania signálu využili. Tento tím zostavil 4 rôzne systémy, s ktorými dosiahol úspešnosť 79.6% – 80.4% a obsadil tak 2., 3., 4. a 5. miesto v celkovom hodnotení súťaže.

Aj ďalší tím v poradí založil svoj systém na konvulčnej neurónovej sieti. Podľa toho, čo autori píšú vo svojom článku [5], špecifikum tohto riešenia je metóda fúzie štandardného spektrogramu a *Constant-Q-Transform (CQT)* spektrogramu. Na to, aby mohli byť navzájom sfúzované príznaky týchto spektrogramov, sú navrhnuté dva fúzovacie mechanizmy - tzv. *voting* a *SVM metódy*. Lepšia z nich - fúzia založená na SVM metóde - obsadila 6.miesto v celkovom poradí s úspešnosťou 77.7%, 7.miesto patrí fúzii na základe *voting* metódy, ktorá získala 74.8%.

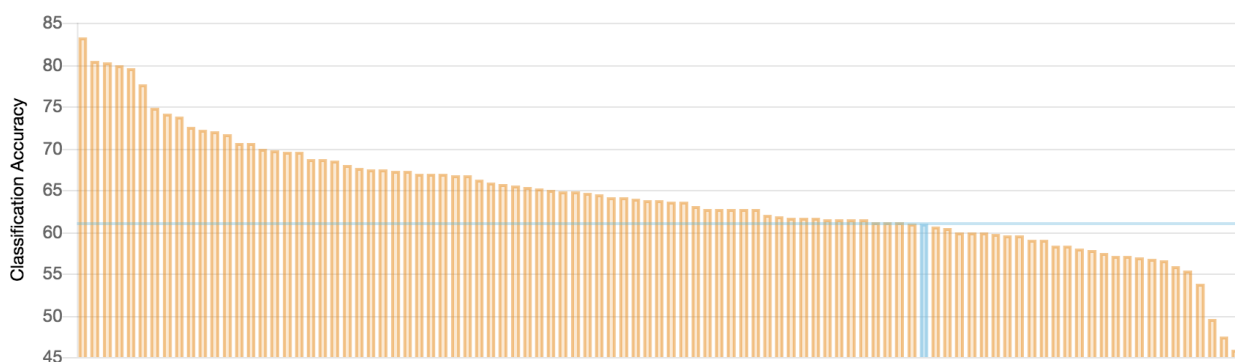
Štvrtý tím v poradí [6] zostavil viaceré systémy, ktorých fúzia priniesla úspešnosť 74.1% a 8.miesto v celkovom poradí. Navrhli systémy založené na konvulčných neurónových sieťach (CNN), systémy založené na CNN super vektoroch s klasifikátorom PLDA, systém založený na GMM super vektoroch a systém doprednej neurónovej siete so súborom akustických príznakov. Vo svojom článku [6] neuviedli úspešnosti jednotlivých subsystémov, iba úspešnosť fúzovaných systémov.

Tím z rakúskej univerzity v Linzi [7], ktorý vyhral súťaž v roku 2016, obsadil 9.miesto s úspešnosťou 73.8% na evaluačných dátach s fúziou i-vector systému a konvulčnej neurónovej siete. Okrem fúzie tento tím odovzdával aj samostatný i-vector systém, ktorý dosiahol 68.7% a obsadil 20.miesto v celkovom poradí.

3. Teoretický základ

Vo svojej práci používam k riešeniu dva prístupy, pričom oba využívajú MFCC príznaky:

1. **Gaussovský klasifikátor** – pre každú triedu sa natrénuje GMM (*popis GMM v nasledujúcej kapitole*), pri vyhodnotení zisťujeme ako GMM danej triedy pasuje na testovaciu nahrávku



Obrázok 2. Graf výsledkov zúčastnených systémov súťaže DCASE 2017 [2]. Modrý stĺpec znázorňuje baseline systém poskytnutý autormi súťaže.

130 2. **i-vector** – je nízorozmerný popis celej nahrávky, pre každú triedu. Tento jednoduchý lineárny klasi- 159
 131 ktorý je klasifikovaný Gaussovským klasifikáto- fikátor sa dá použiť na klasifikáciu i-vectorov i na 160
 132 rom (jedna Gaussovka pre triedu) fúziu niekoľkých systémov. 161

133 3.1 Gaussian Mixture Model

134 V úlohách zaoberajúcich sa automatickým rozpoz-
 135 naním reči sa rozloženie akustických príznakových
 136 vektorov modeluje často pomocou zmesi Gaussovských
 137 (normálnych) rozložení (*angl. Gaussian Mixture Mod-*
 138 *els - GMM*). Táto zmes predstavuje model tvorený
 139 váženou kombináciou Gaussovských rozložení charak-
 140 terizovaných určitou váhou, vektorom stredných hodnôt
 141 a kovariančnou maticou, ako píše aj Ing. Jan Silovský
 142 vo svojej práci [8].

143 3.2 i-vector

144 i-vector predstavuje elegantný spôsob transformácie
 145 viac dimenzionálnych vstupných dát na menej dimen-
 146 zionálny vektor príznakov, pričom zachováva väčšinu
 147 pôvodnej informácie [9]. Táto technika bola pôvodne
 148 inšpirovaná frameworkom *Joint Factor Analysis*, ktorý
 149 bol uvedený v [10].

Hlavnou myšlienkou je, že GMM supervektor $\mathbf{s}$, ktorý zreľazuje GMM vektory stredných hodnôt, môže byť modelovaný nasledovne:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \quad (1)$$

150 kde $\mathbf{m}$ je supervektor stredných hodnôt modelu
 151 UBM GMM, $\mathbf{T}$ je matica reprezentujúca bázy spájajúce
 152 podpriestor s dôležitou variabilitou v priestore super-
 153 vektoru stredných hodnôt, a $\mathbf{w}$ je viazaná premenná so
 154 štandardnou normálnou distribúciou.

155 3.3 Lineárny Gaussovský klasifikátor

156 Vektory vypovedajúce o úrovni skóre (napríklad i-
 157 vectory) sú modelované tak, že majú zdieľanú ko-
 158 varianciu cez všetky triedy a zvlášť meany/priemery

4. Dataset TUT Acoustic scenes 2017 162

Pre túto úlohu je určený dataset *TUT Acoustic scenes 2017*. Obsahuje nahrávky z rozličných akustických prostredí, pričom každá z nich bolo nahrávaná v inej lokalite. Jedná sa o 15 rôznych akustických prostredí, ktoré boli spomínané v úvodnej kapitole. Audio dáta boli nahrávané ako stereo so vzorkovacou frekvenciou 44.1kHz a s 24bitovou presnosťou. 169

Pôvodné 3-5 minútové nahrávky autori rozdelili na 10 sekundové segmenty. Každý segment predstavuje osobitný súbor. Dataset sa skladá z dvoch častí: 172

- **Development dataset** - pozostáva z 312 segmentov pre každé z 15 prostredí, čo je 52 minút audia. 173-175
- **Evaluation dataset** - obsahuje nahrávky z rovnakých prostredí, no z iných geografických lokalít. Pre každé prostredie je k dispozícii 108 segmentov, čo je 18 minút audia. 176-179

Development dataset dohromady obsahuje 13 hodín audio nahrávok v 4680 segmentoch (súboroch). V tejto práci som ho použil na trénovanie GMM systému a i-vector systému. 180-183

Evaluation dataset pozostáva zo 4 hodín a 30 minút audia rozdeleného do 1620 segmentov (súborov). Tento dataset je použitý na reportovanie výsledkov. 184-186

5. Experimenty 187

V tejto kapitole uvádzam všetky experimenty, ktoré som v tejto práci vykonal. Budú spomínané tak nas-
 tavenia jednotlivých systémov ako aj ich výsledky.
 Týka sa to GMM systému, i-vector systému a taktiež
 fúzie týchto dvoch systémov. 188-192

Hlavné experimenty skúmajú nastavenie MFCC koeficientov, vzorkovacej frekvencie signálu a veľkosti klasifikátorov (počet Gaussoviek a veľkosť i-vectoru) a ich vplyv na výslednú úspešnosť systému.

5.1 GMM

Táto kapitola popisuje experimenty s nastavením príznakov a parametrov GMM systému samotného, ktoré mohli ovplyvniť úspešnosť finálneho systému:

- Počet Gaussoviek
- Počet MFCC koeficientov
- Normalizácia MFCC koeficientov
- Nultý MFCC koeficient (C0)
- Delta a doubledelta koeficienty
- Vzorkovacia frekvencia signálu

Vplyv počtu Gaussoviek na úspešnosť systému zobrazuje tabuľka 1. Najlepší výsledok dosiahol systém pri 128 Gaussovkách a preto v ďalších experimentoch figuruje práve tento počet Gaussoviek.

Tabuľka 1. Vplyv počtu Gaussoviek na úspešnosť systému

Počet Gaussoviek	1	2	4	8	16	32	64	128	256
Úspešnosť [%]	40	43	43	43	44	44	44	45	44

Ďalším skúmaným parametrom bol počet MFCC koeficientov. Tabuľka 2 zobrazuje výsledky tohto experimentu. Najlepšie sa ukázalo použitie 19 a 20 MFCC koeficientov. Do ďalších experimentov som teda bral do úvahy 20 MFCC koeficientov.

Tabuľka 2. Vplyv počtu MFCC koeficientov na úspešnosť systému.

Počet koeficientov	4	7	8	11	12	15	16	19	20	23
Úspešnosť [%]	37	39	40	43	44	44	44	45	45	44

Ďalší experiment odzrkadľuje vplyv normalizácie MFCC koeficientov na úspešnosť systému. Detailnejšie výsledky možno nájsť v tabuľke 3. Ako z tabuľky vyplýva, najlepšie sa v tomto prípade javia koeficienty, ktoré nie sú normalizované. Preto v ďalších pozorovaniach uvažujem práve nenormalizované koeficienty.

Tabuľka 3. Vplyv normalizácie MFCC koeficientov

Normalizácia	Úspešnosť [%]
Mean	33
Mean a variance	34
Žiadna	45

Zahrnutie, resp. nezahrnutie nultého MFCC koeficientu taktiež patrí k parametrom, ktoré môžu ovplyvniť úspešnosť systému. Informácie o tomto experimente obsahuje tabuľka 4. Najúspešnejšie sa v tomto prípade ukázalo zahrnúť nultý koeficient C0, a preto nasledujúce experimenty nultý koeficient obsahujú.

Tabuľka 4. Vplyv nultého koeficientu na úspešnosť systému

Nultý koeficient	Úspešnosť [%]
Bez C0	45
C0	54
E (energia)	45

Nasleduje experiment, ktorý sleduje vplyv odvodených koeficientov - tzv. *delta a doubledelta koeficientov* na úspešnosť systému. Z tabuľky 5 možno vyvodiť, že v prípade zahrnutia aj delta aj doubledelta koeficientov systém dosahuje najlepšie výsledky.

Tabuľka 5. Vplyv priamych (P), delta (D) a doubledelta (A) koeficientov

Koeficienty	Úspešnosť [%]
P	54
P + D	56
P + D + A	56

Záverečným experimentom v prípade GMM systému bolo sledovanie vplyvu vzorkovacej frekvencie signálu na úspešnosť systému. Z tabuľky 6 sa javí najlepšie vzorkovacia frekvencia 16000 Hz.

Tabuľka 6. Vplyv vzorkovacej frekvencie signálu

Vzorkovacia frekvencia [Hz]	Úspešnosť [%]
8000	56
16000	59
44100	57

Tento systém dosiahol úspešnosť 59% a je to GMM systém, s ktorým som dosiahol najvyššiu úspešnosť. Baseline systém súťaže z roku 2016 bol založený tiež na GMM, pri použití datasetu z roku 2017 bola jeho úspešnosť 52% na evaluačných dátach.

5.2 i-vector

Na základe skúseností z GMM klasifikátorov, ktoré sú popísané v predchádzajúcej kapitole, som sa po niekoľkých experimentoch ustálil na nasledujúcom nastavení extrakcie príznakov a predspracovaní:

- 24 MFCC koeficientov + nultý koeficient C0

- Všetne delta koeficientov
- 32 Mel-filter bánk
- Vzorkovacia frekvencia nahrávok 16 kHz

S týmto nastavením som zisťoval závislosť veľkosti GMM a i-vectoru a sledoval úspešnosť systému. Vzhľadom na to, že je k dispozícii málo trénovacích dát (v rámci pravidiel súťaže), snažím sa nájsť bod, v ktorom sa systém začne pretrénovávať.

Detailné výsledky sú v tabuľke 7. Najlepšie sa javí počet Gaussoviek okolo 512 a väčší rozmer i-vectorov – pre 400 dosahuje najlepších výsledkov 68%. S vyšším počtom Gaussoviek sa úspešnosť znižuje, čo znamená, že systém je už pretrénovaný z dôvodu nedostatku trénovacích dát. Najlepšie umiestnený tím súťaže, ktorý sa venoval tiež i-vector systému[7] dosiahol úspešnosť 68.7%, z čoho môžeme usúdiť, že úspešnosť tu popisovaného i-vector systému sa líši len minimálne.

Tabuľka 7. Úspešnosť i-vector systému [%] vzhľadom na počet Gaussovských rozložení a dimenzionalitu i-vectoru.

Počet Gaussoviek	Rozmer i-vectoru		
	100	200	400
32	64.75	62.72	61.73
64	65.43	65.43	63.33
128	65.68	66.42	66.42
256	65.93	66.54	66.98
512	64.63	67.90	68.27
1024	64.44	65.62	66.11
2048	64.82	64.63	64.75

5.3 Fúzia

Na záver som uskutočnil fúziu najlepšieho GMM systému a najlepšieho i-vector systému použitím Gaussovského lineárneho klasifikátoru. Porovnanie úspešností týchto systémov sa nachádza v tabuľke 8.

Tabuľka 8. Úspešnosť najlepších systémov z predošlých 2 kapitol a ich vzájomnej fúzie.

Systém	Úspešnosť [%]
GMM	59
i-vector	68.27
GMM + i-vector	69.01

Fúzia sa prejavila zvýšením úspešnosti systému na 69.01%, prekonala vyvinutý i-vector systém, čo potvrdzuje jej pozitívny vplyv. Toto zvýšenie úspešnosti však nebolo až tak významné, pretože je medzi jednotlivými systémami príliš veľký rozdiel v úspešnosti. GMM systém má značne nižšiu úspešnosť oproti

i-vector systému na to, aby v tejto fúzii výraznejšie pomohol.

Tabuľka 9 zobrazuje confusion maticu, ktorá detailnejšie popisuje chovanie tohto systému. Môžeme vidieť, že systém si dokázal najlepšie poradiť s triedami *pláž*, *obydlie/dom* a *stanica metra*, ktoré dokázal klasifikovať len s minimálnou chybou. Naopak najväčší problém mal systém s triedami *autobus*, *knižnica*, *park* a *električka*. *Sídliisko* si často pletol s *parkom*, *obchod* s *potravínami* si systém neraz splietol s *autobusom*. *Stanica metra* bola viackrát zamotaná s *obchodom* s *potravínami* alebo s *kaviarňou*. Ako *obydlie/dom* boli často klasifikované nahrávky z *knižnice* a *kancelárie*, čo nasvedčuje, že systém sa zameriaval viac na pozadie nahrávok, ktoré bolo v týchto prípadoch veľmi podobné.

6. Záver

V tejto práci som implementoval systém pre klasifikáciu akustického prostredia z audio nahrávky. Použil som dva prístupy: GMM systém a i-vector systém. Okrem toho som na záver vytvoril fúziu najlepšieho GMM systému a i-vector systému.

Baseline systém súťaže DCASE 2017 dosiahol úspešnosť 61%. Môj najlepší systém, ktorým bola spomínaná fúzia, dosiahol 69%, čo by znamenalo umiestnenie na 20.mieste spomedzi 97 odovzdaných systémov, ktoré sa súťaže zúčastnili.

Na základe prehľadu najlepších systémov som zisťtil, že autori týchto systémov často fúzovali svoje systémy s neurónovými sieťami a preto by ďalším krokom tejto práce mohla byť práve implementácia systému založeného na neurónovej sieti a následná fúzia s už existujúcim napr. i-vector systémom.

Podakovanie

Rád by som poďakoval vedúcemu mojej práce, Pavlovi Matějkovi, za všetok jeho čas, za jeho priateľský a ľudský prístup a za všetky cenné rady a postrehy, ktoré mi poskytol.

Literatúra

- [1] T. Heittola et al. *Detection and Classification of Acoustic Scenes and Events 2016*. [Online; navštívené 15.03.2018].
- [2] T. Heittola et al. *Detection and Classification of Acoustic Scenes and Events 2016*. [Online; navštívené 15.03.2018].
- [3] Seongkyu Mun, Sangwook Park, David Han, and Hanseok Ko. Generative adversarial network based acoustic scene training set augmentation

Tabuľka 9. Confusion matica najlepšieho systému, t.j. fúzie najlepšieho i-vector a GMM systému, ktorý dosiahol úspešnosť 69.01% na evaluačných dátach. Riadky predstavujú skutočné návestia a stĺpce reprezentujú návestia predpovedané systémom.

	Pláž	Bus	Kavi	Auto	Cent	Les	Potr	Dom	Kniž	Metr	Kanc	Park	Sídl	Vlak	Elek
Pláž	92	0	1	0	0	0	0	0	0	0	0	6	9	0	0
Autobus	0	53	1	6	0	0	28	0	0	13	0	3	1	2	1
Kaviareň	0	0	81	0	0	0	6	0	0	20	0	1	0	0	0
Auto	0	2	0	80	0	0	1	0	0	0	0	0	0	0	25
Centrum mesta	2	0	0	0	90	0	0	0	0	2	0	10	3	1	0
Lesná cesta	1	0	0	0	0	90	0	0	0	0	0	10	7	0	0
Potraviny	0	0	3	0	1	0	74	2	0	28	0	0	0	0	0
Obydlie/Dom	0	0	1	0	0	1	0	99	0	2	2	2	1	0	0
Knižnica	0	0	0	0	0	13	1	22	51	19	2	0	0	0	0
Stanica metra	0	0	4	0	0	2	0	0	0	97	3	2	0	0	0
Kancelária	0	0	0	0	0	18	0	24	0	0	66	0	0	0	0
Park	0	0	1	0	2	3	20	0	0	7	0	42	33	0	0
Sídlisko	0	0	1	0	14	10	0	0	0	1	0	2	80	0	0
Vlak	0	5	0	0	3	0	0	0	3	5	0	0	0	69	23
Električka	0	3	10	0	3	0	22	0	0	16	0	0	0	0	54

and selection using SVM hyper-plane. Technical report, DCASE2017 Challenge, September 2017.

[4] Yoonchang Han and Jeongsoo Park. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. Technical report, DCASE2017 Challenge, September 2017.

[5] Zheng Weiping, Yi Jiantao, Xing Xiaotao, Liu Xiangtao, and Peng Shaohu. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. Technical report, DCASE2017 Challenge, September 2017.

[6] Rakib Hyder, Shabnam Ghaffarzadegan, Zhe Feng, and Taufiq Hasan. BUET bosch consortium (B2C) acoustic scene classification systems for DCASE 2017. Technical report, DCASE2017 Challenge, September 2017.

[7] Bernhard Lehner, Hamid Eghbal-Zadeh, Matthias Dorfer, Filip Korzeniowski, Khaled Koutini, and Gerhard Widmer. Classifying short acoustic scenes with I-vectors and CNNs: Challenges and optimisations for the 2017 DCASE ASC task. Technical report, DCASE2017 Challenge, September 2017.

[8] Jan Silovský. Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvcích. Technical report, Technická univerzita v Liberci, November 2011.

[9] David González Martínez, Oldřich Plchot, L. Burget, O. Glembek, and P. Matějka. Lan-

guage recognition in ivectors space. In *Proceedings of Interspeech 2011*, pages 861–864, 2011.

[10] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2072–2084, 2007.