VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DETEKCE AKUSTICKÉHO PROSTŘEDÍ Z ŘEČI

ACOUSTIC SCENE CLASSIFICATION FROM SPEECH

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. MATÚŠ DOBROTKA**
AUTHOR

VEDOUCÍ PRÁCE **Ing. PAVEL MATĚJKA, Ph.D.**
SUPERVISOR

BRNO 2018

Zadání diplomové práce/20562/2017/xdobro12

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2017/2018

Zadání diplomové práce

Řešitel: **Dobrotka Matúš, Bc.**
Obor: Počítačová grafika a multimédia
Téma: **Detekce Akustického Prostředí z Řeči**
Acoustic Scene Classification from Speech
Kategorie: Zpracování řeči a přirozeného jazyka

Pokyny:
Cílem práce je klasifikace audio nahrávky do předem definovaných tříd, které charakterizují prostředí, ve kterém byla nahrávka pořízena - například - kancelář, ulice, park, kavárna ....

1. Prostudujte statistické techniky pro modelování řeči - zaměřte se převážně na Gausovské modely.
2. Prostudujte doporučenou literaturu.
3. Seznamte se s metody používanými pro detekci prostředí z audia
4. Seznamte se s daty a základním systémem pro soutěže DCASE 2016 a 2017
5. Zjistěte úspěšnost základního systému na datech z těchto soutěží.
6. Navrhněte a otestujte alespoň dvě změny oproti základnímu systému založené například na změně vstupních příznaků, klasifikátoru či postavit systém pomocí i-vector extraktoru.
7. Proveďte experimenty s fúzí dvou systémů.

Literatura:
- http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification
- http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification
- i-vector http://www.fit.vutbr.cz/~matejkap/pubs.php?id=9657
- x-vector https://sites.google.com/site/dgromeroweb/publications

Při obhajobě semestrální části projektu je požadováno:
- Body 1 až 5.

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci dřívějších projektů (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Matějka Pavel, Ing., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2017
Datum odevzdání: 23. května 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Téma tejto diplomovej práce je klasifikácia audio nahrávky do 15 tried akustických prostredí, v ktorých sa ľudia bežne nachádzajú. Práca popisuje 2 metódy založené na GMM a i-vektoroch a ich vzájomnú fúziu. Na dátach zo súťaže DCASE dosiahol najlepší GMM systém úspešnosť 60.4% a i-vektor systém 68.4%. Fúzia GMM systému a najlepšieho i-vektor systému výsledok ešte zlepšila na 69.3%, čo by v dobe súťaže stačilo na 20. miesto z 98 odovzdaných systémov z celého sveta.

Abstract

The topic of this thesis is an audio recording classification with 15 different acoustic scene classes that represent common scenes and places where people are situated on a regular basis. The thesis describes 2 approaches based on GMM and i-vectors and a fusion of the both approaches. The score of the best GMM system which was evaluated on the evaluation dataset of the DCASE Challenge is 60.4%. The best i-vector system's score is 68.4%. The fusion of the GMM system and the best i-vector system achieves score of 69.3%, which would lead to the 20th place in the all systems ranking of the DCASE 2017 Challenge (among 98 submitted systems from all over the world).

Kľúčové slová

Detekcia akustického prostredia, GMM, banky filtrov, MFCC koeficienty, i-vektor, i-vektor extraktor, fúzia, DCASE Challenge, lineárny Gaussovský klasifikátor.

Keywords

Acoustic scene classification, GMM, filter banks, MFCC, i-vector, i-vector extractor, fusion, DCASE Challenge, linear Gaussian classifier.

Citácia

DOBROTKA, Matúš. *Detekce Akustického Prostředí z Řeči.* Brno, 2018. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Pavel Matějka, Ph.D.

Detekce Akustického Prostředí z Řeči

Prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením pána Ing. Pavla Matějku, Ph.D. Všetky literárne zdroje a publikácie, z ktorých som čerpal, sú uvedené v zozname použitej literatúry.

.....
Matúš Dobrotka
23. mája 2018

Poďakovanie

Chcem poďakovať vedúcemu mojej práce, Ing. Pavlovi Matějkovi, Ph.D., za všetok čas, ktorý mi nielen počas konzultácií ochotne venoval, za priateľský a ľudský prístup pri riešení najrôznejších problémov, za všetky cenné rady, ktoré mi poskytol a za trpezlivosť pri odpovedaní na všetky moje otázky rôzneho druhu. Okrem toho by som rád poďakoval aj mojej Vierke za trpezlivosť, ktorú mi počas celej doby preukazovala, aj keď to pre ňu bolo často náročné hlavne časovo a tiež za podporu, ktorú som v nej mohol cítiť. Poďakovanie patrí aj mojej rodine za podporu a trpezlivosť a tiež všetkým mojim kamarátom a známym, ktorí ma rôznym spôsobom podporovali.

Obsah

Kapitola 1

Úvod

Ľudia klasifikujú zvukové signály v podstate stále a to väčšinou bez toho, aby si to uvedomovali či vynakladali na to akési väčšie úsilie [5]. Úlohy ako napríklad rozpoznávanie hlasu v telefóne či vnímanie rozdielu medzi zvončekom na dverách a zvonením telefónu nepovažujeme za veľmi náročné. Problémy sa ale začnú objavovať v prípade, keď je zvuk slabý alebo je počuť prílišný šum, alebo vtedy, keď sú nejaké zvuky navzájom veľmi podobné. Celkom komplikovaná však môže byť napríklad situácia, keď by sme potrebovali zistiť, ktoré dvere sa práve zatvorili v nejakej veľkej hale alebo budove.

Za ľudskou schopnosťou vnímať a rozpoznávať rozličné zvukové podnety stojí komplexný a komplikovaný systém sluchu napojený na extrémne výkonný počítač - mozog. Bolo by nepochybne veľmi užitočné detailne porozumieť týmto mechanizmom a aplikovať ich vo svete moderných technológií. Keby sme poznali všeobecné systémy, ktoré používame na klasifikovanie zvukových podnetov, boli by sme schopní lepšie diagnostikovať a liečiť sluchové ochorenia. Okrem toho by bolo tiež iste užitočné mať stroj, ktorý by dokázal robiť so zvukom to, čo dokáže aj človek. Príkladom by mohli byť napríklad lekári, ktorí počúvajú, ako pacient dýcha s cieľom diagnostikovať respiračné ochorenia. Ak by odborný medicínsky systém dokázal robiť to isté (keď by mal v sebe takú znalosť naprogramovanú), tak by aj vzdialené oblasti mohli dostať diagnózu rýchlo a bez nákladov na konzultáciu s ľudským odborníkom, ktorý by mohol byť v inej krajine a potreboval by byť transportovaný. Podobne sú aj automechanici schopní diagnostikovať problémy s motorom auta počúvaním zvukov, ktoré vydáva, keď beží. Existuje skrátka veľa oblastí, kde ľudia vo svojej práci používajú svoj sluch. Systémy zamerané na klasifikáciu audio signálu by preto mohli poskytovať možnosť, že by takáto práca bola vykonaná aj vo vzdialených komunitách alebo v iných situáciach, kedy by bol odborník nedostupný alebo príliš drahý.

Takéto systémy majú potenciál dokonca počuť oveľa lepšie ako človek. Ak by nás počítače naučili vnímať zvuk podobným spôsobom ako nás naučili mikroskopy a televízne kamery vnímať vizuálny svet, náš pohľad na svet, v ktorom žijeme, by bol opäť o niečo pestrejší. Ďalšie aplikácie by potenciálne mohli zahŕňať odstraňovanie šumu, separovanie zvukov, automatický prepis hudby, textu, Morseho kódu a podobne.

Zvuk nesie obrovské množstvo informácie o našom každodennom prostredí a tiež o fyzikálnych udalostiach, ktoré sa v ňom uskutočňujú [8]. Môžme 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 ďalšie aplikácie. Môže sa jednať napríklad o vyhľadávanie multimédií založené na ich audio obsahu, vytváranie kontextových mobil-

ných zariadení, robotov, automobilov,... a tiež o inteligentné monitorovacie systémy rozpoznávajúce činnosti v ich okolí využitím akustických informácií. Táto oblasť ešte nie je na takej úrovni, aby 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 diplomovej práce vychádza zo zadania medzinárodnej súťaže *DCASE Challenge*, ktorá je popísaná neskôr v tejto kapitole.

Kapitola 2 popisuje súťaž DCASE Challenge a konkrétne úlohu detekcie akustického prostredia. Okrem toho kapitola popisuje zoznam tried, ktoré sú v rámci tejto práce uvažované pri klasifikácii. Posledná časť tejto kapitoly sa venuje rozboru literatúry v oblasti detekcie akustického prostredia z reči/nahrávky.

V kapitole 3 sa nachádzajú teoretické informácie, ktoré súvisia s mojím riešením a implementáciou. Táto kapitola je relatívne stručná, mojim úmyslom nebolo popisovať teoreticky to, čo už veľakrát popísané takýmto spôsobom bolo, ale skôr som sa v tejto práci chcel viac venovať experimentom a praktickej činnosti, a preto som v tejto časti uviedol viaceré zdroje literatúry, kde sa dá v prípade záujmu nájsť viac podrobnejších teoretických informácií o danej problematike.

Obsahom kapitoly 4 je popis datasetov, ktoré som v súvislosti s touto prácou využíval a s ktorými som prišiel do kontaktu. Prioritne som v mojom riešení využíval dataset súťaže DCASE 2017 – *TUT Acoustic scenes 2017.*

Kapitola 5 sa venuje experimentom, ktoré som vykonal s GMM systémami. Popisuje do detailov niekoľko kategórií experimentov a tiež formát týchto experimentov. Okrem toho kapitola prezentuje všetky dosiahnuté výsledky v súvislosti s implementovanými GMM systémami. Jedná sa hlavne o systémy, ktoré sú založené na bankách filtrov a na MFCC koeficientoch.

Podobne je delená aj kapitola 6, kde sú zase uvedené všetky potrebné informácie o vykonaných experimentoch so systémom založeným na i-vektoroch vrátane popisu formátu týchto experimentov. Nachádzajú sa tu tiež všetky dosiahnuté výsledky súvisiace s i-vektor systémami.

Predposledná kapitola 7 popisuje fúziu viacerých kombinácií implementovaných systémov a výsledky získané fúziou. Taktiež sa v tejto kapitole nachádza zhrnutie výsledkov tejto práce, prehľad najúspešnejších naimplementovaných systémov a ich porovnanie s dôležitými systémami zo súťaže DCASE.

Záverečná kapitola 8 sumarizuje celú prácu, aké metódy boli zvolené pre riešenie tejto problematiky, aké výsledky dosiahli jednotlivé systémy a tiež sa tu píše o možnom budúcom vývoji, ktorým by sa práca na tejto téme mohla uberať.

Kapitola 2

DCASE Challenge

Pojmom DCASE Challenge sa označuje celosvetová súťaž organizovaná fínskou výskumnou skupinou zameranou na audio z Technickej Univerzity v Tampere. Okrem toho sa na organizovaní podieľa aj Centrum pre digitálnu hudbu z londýnskej univerzity Queen Mary a ďalší. 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. Nasledujúce podkapitoly v krátkosti popisujú jednotlivé úlohy súťaže.

Detekcia akustického prostredia

Cieľom tejto úlohy je klasifikovať testovaciu nahrávku do jednej z preddefinovaných tried, ktoré charakterizujú prostredie, kde bola nahrávka nahratá - napríklad *park, ulica* či *kancelária.* Dataset zahŕňa nahrávky z 15 rôznych kontextov. Vzhľadom na to, že táto úloha je kľúčová z pohľadu tejto diplomovej práce, bude podrobnejšie popísaná neskôr v tejto kapitole.

Detekcia zvukových udalostí v syntetickom zvuku

Cieľom tejto úlohy je detekovanie zvukových udalostí (napríklad spev vtákov, prechádzanie auta okolo), ktoré sa nachádzajú v audio nahrávke, odhadnutie počiatočného i koncového času týchto udalostí a tiež určenie triedneho priradenia pre každú z týchto udalostí. Zameriava sa na detekciu zvukových udalostí v kancelárii.

Pre účely trénovania je k dispozícii materiál v podobe oddelených zvukových udalostí pre každú triedu. Okrem toho je tvorený syntetickými zmesami vytvorenými z týchto príkladov v rôznom pomere signálu k šumu a k podmienkam hustoty udalostí. Účastníci majú dovolené použiť ľubovoľnú kombináciu týchto trénovacích dát na natrénovanie ich systému.

Testovacie dáta sú podobne ako trénovacie zložené zo syntetických zmesí získaných na základe príkladových nahrávok za rôznych podmienok (rozličné úrovne pomeru signálu k šumu, hustota udalostí, polyfónia).

Detekcia zvukových udalostí vo zvuku z reálneho života

Úloha používa tak trénovací ako aj testovací materiál nahratý v prostredí reálneho života. Podmienky sú podobné k nášmu každodennému životu, pretože zdroje zvuku sú len zriedkavo izolované od seba. Počet prekrývajúcich sa zvukových udalostí v rovnakom čase nie je striktný ani v trénovacích, ani v testovacích audio dátach, pretože anotácie týchto udalostí boli vykonané manuálne, a preto môžu byť istým spôsobom subjektívne.

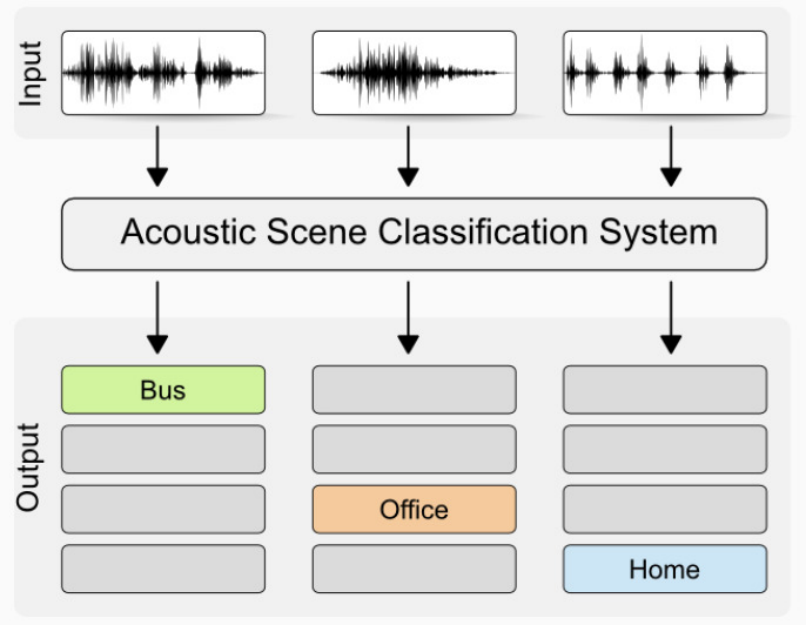
Domáci audio tagging

Táto úloha využíva binaurálne nahrávky vytvorené v domácom prostredí. Hlavné zdroje zvuku v akustickom prostredí sú dvaja dospelí a dve deti, televízia a elektronické prístroje, kuchynské spotrebiče, kroky a klopania vytvorené človekom a zvuky, ktoré vznikli mimo domu.

Audio dáta pre túto úlohu sú poskytnuté v podobe krátkych 4 sekundových nahrávok a cieľom je označiť každú z týchto nahrávok jedným alebo viacerými charakteristickými označeniami, ako napríklad reč dieťaťa alebo reč dospelého muža, a/alebo video hra/televízia.

2.1 Detekcia akustického prostredia

Ako už bolo naznačené v predošlej kapitole, cieľom detekcie akustického prostredia je klasifikovať testovaciu nahrávku do jednej z preddefinovaných tried, ktorá charakterizuje prostredie, v ktorom bola nahrávka nahratá. Takýto systém výstižne popisuje obrázok 2.1.



Obr. 2.1: Prehľad systému pre klasifikáciu akustického prostredia [10]

2.2 Audio dataset a klasifikované triedy

Pre túto úlohu sú určené datasety *TUT Acoustic scenes 2016* a *TUT Acoustic scenes 2017*, ktoré boli vytvorené konkrétne pre účel tejto úlohy vrámci popisovanej súťaže v rokoch 2016 a 2017. Obsahuje nahrávky z rozličných akustických prostredí, pričom každá z nich bolo nahrávaná v inej lokalite.

Jedná sa o 15 rôznych akustických prostredí:

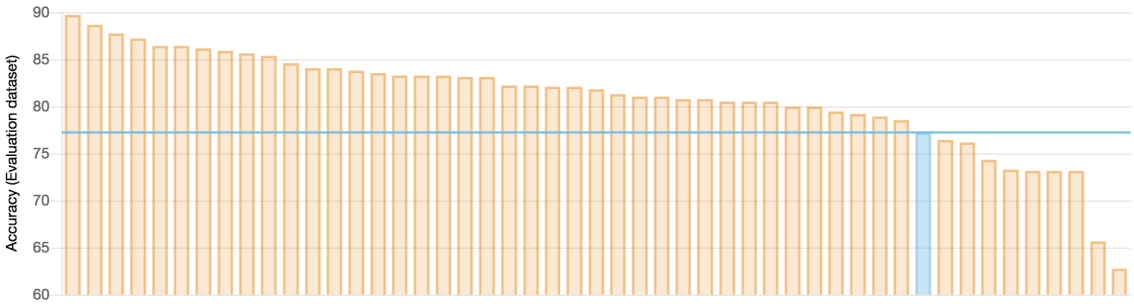
- Autobus - cestovanie autobusom v meste (vozidlo)
- Kaviareň/reštaurácia - malá kaviareň/reštaurácia (vnútorné prostredie)
- Auto - šoférovanie alebo cestovanie ako pasažier, v meste (vozidlo)
- Centrum mesta (vonkajšie prostredie)
- Lesná cesta (vonkajšie prostredie)
- Obchod s potravinami - obchod strednej veľkosti (vnútorné prostredie)
- Obydlie/dom (vnútorné prostredie)
- Pláž pri jazere (vonkajšie prostredie)
- Knižnica (vnútorné prostredie)
- Metro stanica (vnútorné prostredie)
- Kancelária - viacero osôb, typický pracovný deň (vnútorné prostredie)
- Sídlisko/obytná oblasť (vonkajšie prostredie)
- Vlak (vozidlo)
- Električka (vozidlo)
- Mestský park (vonkajšie prostredie)

Všetky audio dáta týchto dvoch datasetov boli nahraté organizátormi súťaže vo Fínsku, prevažne v mestách Helsinki a Tampere a to v období od júna 2015 do januára 2016 (v prípade datasetu pre rok 2017 do januára 2017). Podrobnejší popis týchto datasetov obsahuje kapitola 4.

2.3 Prehľad výskumu pre DCASE 2016

Vzhľadom na to, že téma tejto práce veľmi úzko súvisí so zadaním prvej úlohy súťaže DCASE Challenge, ako prehľad výskumu v tejto oblasti uvádzam najúspešnejšie riešenia autorov, ktorí sa súťaže zúčastnili a zároveň riešili práve úlohu detekcie akustického prostredia. Súťaže sa v roku 2016 zúčastnilo celkovo 37 tímov z celého sveta a odovzdaných systémov bolo spolu 49.

Celkové hodnotenie všetkých odovzdaných riešení zobrazuje obrázok 2.2, na ktorom môžme vidieť, že najlepší systém dosiahol na evaluačných dátach úspešnosť 89.7%. Baseline systém, ktorý poskytli autori súťaže, je zvýraznený modrou farbou a na evaluačných dátach dosahoval úspešnosť 77.2%.



Obr. 2.2: Graf výsledkov zúčastnených systémov súťaže DCASE 2016 na evaluačných dátach[10]. Modrý stĺpec znázorňuje baseline systém poskytnutý autormi súťaže.

2.3.1 Hamid Eghbal-Zadeh

Súťaž DCASE 2016 vyhral tím z rakúskej univerzity v Linzi pod vedením Hamida Eghbal-Zadeho. Vďaka svojmu najlepšiemu systému sa umiestnil na prvom mieste v súťaži spomedzi 49 odovzdaných systémov. Ako tento víťazný tím sám uviedol vo svojom článku [4], v ktorom popisuje svoje riešenie, spoločne vytvorili 4 rôzne systémy, pričom práve posledný z nich bol ten víťazný. Parametre tohto systému budú stručne popísané na nasledujúcich riadkoch.

MFCC konfigurácia a nastavenie i-vektorov

Autori vo svojom článku uvádzajú, že podľa ich experimentov je najlepšie použiť 23 MFFC koeficientov bez nultého koeficientu extrahovaných aplikovaním 20ms pozorovacieho okna bez akéhokoľvek prekrytia. 18 MFCC delta koeficienty (vrátane nultého delta koeficientu) a 20 MFCC double delta koeficientov (vrátane nultého double delta koeficientu) sú extrahované aplikovaním 60ms pozorovacieho okna umiestneného symetricky okolo 20ms rámca. Bez ohľadu na dĺžku pozorovacieho okna používajú 30 mel-škálovateľných filtrov trojuholníkového tvaru v rozmedzí 0 až 11 kHz.

Čo sa týka nastavenia i-vektorov, autori tohto riešenia trénujú ich UBM modely s 256 Gaussovskými komponentami na MFCC príznakoch extrahovaných z jednotlivých častí audia. UBM, matica T, LDA *(Linear Discriminant Analysis)* a WCCN *(Within-Class Covariance Normalization)* projekcie sú trénované na trénovacej sade každého cross-validačného rozdelenia. Dimenzionalitu i-vektorov autori nastavujú na 400. Na vyhodnotenie premietnutých i-vektorov používajú kosínové vyhodnocovanie.

DCNN architektúra

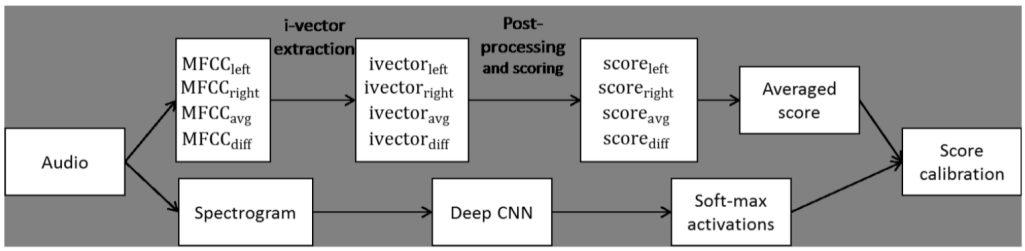
Podrobnú použitú architektúru DCNN možno nájsť priamo v článku autorov [4] v tabuľke 4. Časť modelu súvisiaca s učením príznakov nasleduje *VGG style networks* pre rozpoznávanie objektov. Klasifikačná časť siete je zase navrhnutá ako globálna priemerná pooling vrstva tak, ako je to známe z architektúry *Network in Network*. Veľkosť vstupu siete je jednokanálová časť spektrogramu s veľkosťou 149 x 149. To znamená, že autori trénujú model nie na celých sekvenciách, ale len na malých oknách. Spektrogramy pre tento prístup sú počítané nasledovne: Audio je vzorkované ako 22050 vzorkov na sekundu. Následne sa spočíta *Short Time Fourier Transform (STFT)* na 2048 vzorkových oknách pri snímkovej frekvencii 31.25 snímkov za sekundu. Potom je STFT dodatočne spracovaná s 24 logaritmickými filterbankami, logaritmickým rozsahom a priepustným pásmom 20 Hz až 16 kHz.

Parametre tohto modelu sú optimalizované metódou *minibatch stochastic gradient decent* a *momentum*. *Mini-batch* veľkosť je nastavená na 100 vzorkov. Trénovanie začína s *learning rate* 0.02, pričom sa polí každých 5 epoch. *Momentum* je pevne nastavené na 0.9 počas celého trénovania. Okrem toho autori aplikujú *L2 weight decay penalty* ako 0.0001 na všetky trénovateľné parametre tohto modelu.

Pre klasifikáciu nevidených vzorkov v čase testovania postupujú nasledovne: najskôr prebehnú sliding oknom cez všetky testovacie sekvencie a zistia pravdepodobnosti jednotlivých tried pre každé okno. Ako druhý krok spriemerujú pravdepodobnosti všetkých príspevkov a priradia triedu s najväčšou priemernou pravdepodobnosťou.

Fúzia

Jedná sa o fúziu hlbokých konvolučných neurónových sietí (DCNN) a binaurálnych i-vektorov. Toto riešenie ilustruje obrázok 2.3. Po extrakcii binaurálnych i-vektorov sa na tes-



Obr. 2.3: Blokový diagram víťazného tímu súťaže - fúzia medzi binaurálnymi i-vektormi a DCNN. [4]

tovacej sade spočíta ich konečná výsledková matica. Okrem toho je natrénovaná DCNN (hlboká konvolučná neurónová sieť) a spočítajú sa aj soft-max aktivácie na testovacej sade. Prostredníctvom metódy *linear logistic regression score calibration* skombinovali autori dosiahnuté výsledky binaurálnych i-vektorov a soft-max aktivácie v prípade DCNN skombinovali do jednej výsledkovej matice. Projekčné modely sa učia pomocou výsledkov binaurálnych i-vektorov dosiahnutých na validačnej sade a tiež pomocou soft-max aktivácií z validačnej sady. Výsledky binaurálnych i-vektorov a soft-max aktivácií na testovacej sade sú následne sfúzované použitím modelov naučených na validačnej sade. Výsledok tejto fúzie je použitý pre finálnu predikciu. Fúzia dosiahla na evaluačných dátach úspešnosť 89.7%.

Pre zlepšenie úspešnosti autori experimentovali pri i-vektoroch aj s kalibráciou skóre. Kalibračnú transformáciu získali prostredníctvom *lineárnej logistickej regresie*. Takto transformované výsledky použili pre finálnu predikciu. Tieto modifikované i-vektory predstavovali ďalší odovzdaný systém tohto tímu, ktorý obsadil 2. miesto v celkovom poradí a získal 88.7% na evaluačných dátach.

2.3.2 Victor Bisot

Aj francúzsky tím Victora Bisota vyvinul vrámci súťaže veľmi úspešný systém [2]. Obsadil 3. miesto v celkovom poradí s úspešnosťou 87.7% na evaluačných dátach. V nasledujúcich riadkoch sa bližšie pozrieme na tento systém.

Autori navrhujú prístup učenia príznakov nasledovaný myšlienkou dekompozície časovo-frekvenčných reprezentácií s faktorizáciou nezáporných matíc. Ich cieľom bolo naučenie

spoločného slovníka, ktorý reprezentuje dáta, a tiež použiť projekcie na tento slovník ako príznaky pre klasifikáciu. Tento systém je založený na novom rozšírení faktorizácie nezáporných matíc s učiteľom. Vrámci prístupu, ktorý navrhujú, je slovník a klasifikátor spoločne optimalizovaný s cieľom nájsť vhodnú reprezentáciu na minimalizáciu ceny klasifikácie.

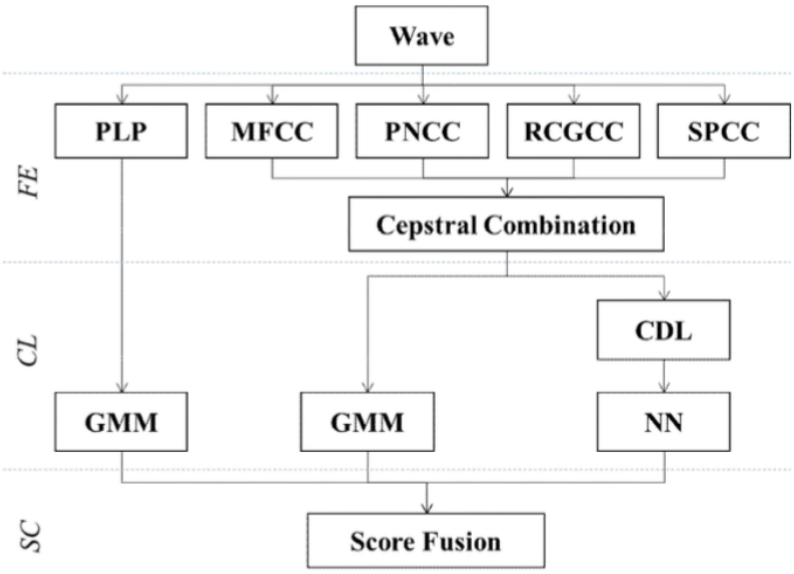
Navrhovaná metóda podstatne prekonáva baseline systém a poskytuje zlepšené výsledky v porovnaní s faktorizáciou nezáporných matíc bez učiteľa. Obrázok 2.4 zobrazuje úspešnosť systému s metódou bez učiteľa *(Sparse NMF)* a s metódou s učiteľom *(Nonnegative TDL)*. Jedná sa o výsledky na development sade nahrávok, je vidno, že metóda s učiteľom výrazne prekonáva metódu bez učiteľa.

	K=128	K=256	K=512
Sparse NMF	81.0	81.2	82.6
Nonnegative TDL	84.2	**85.0**	**84.8**

Obr. 2.4: Porovnanie úspešnosti metódy bez učiteľa a metódy s učiteľom na development sade *(NMF=Nonnegative Matrix Factorization, TDL=Task-driven Dictionary Learning)*. Tabuľka pochádza z článku [2].

2.3.3 Hanseok Ko

Základom ďalšieho úspešného systému je taktiež fúzia, jeho architektúru zobrazuje obrázok 2.5.



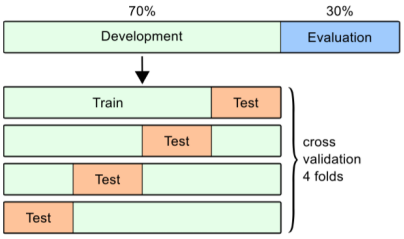
Obr. 2.5: Architektúra úspešného kórejského systému pre klasifikáciu akustického prostredia *(FE=Feature Extraction, CL=Classification, SC=Score Fusion)*. Obrázok prevzatý z [20].

V článku kórejských autorov pod vedením Hanseok-a Ko-a [20] sa píše o fúzii 3 rôznych systémov. V prvom rade je to GMM (Gaussian Mixture Model) systém postavený nad PLP

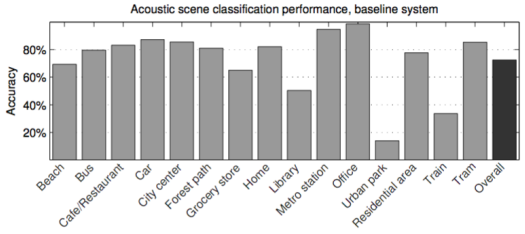
koeficientami. Druhým fúzovaným systémom je GMM systém postavený nad zlúčenými koeficientami MFCC, PNCC, RCGCC a SPCC. Tretím systémom je klasifikátor založený na metóde najbližšieho suseda, ktorý autori vytvorili nad kombináciou koeficientov spomínaných v predošlom systéme. Takto zostavený systém dosiahol úspešnosť na evaluačných dátach 87.2% a obsadil tak 4. miesto v celkovom poradí.

2.3.4 Baseline systém 2016

Baseline systém, ktorý poskytli autori súťaže [9], sa skladá z klasických mel-frekvenčných cepstrálnych koeficientov (MFCC) a z klasifikátoru založeného na zmesi Gaussovských rozložení (GMM). MFCC koeficienty boli spočítané pre celé audio použitím 40ms rámcov s Hammingovým oknom, s prekrytím 50% a so 40 mel-filter bankami. Prvých 20 koeficientov bolo uchovaných včetne 0. koeficientu. Odvodené delta a double delta koeficienty boli tiež spočítané oknom širokým 9 rámcov, čo znamená, že dimenzia vektoru príznakov založených na rámcoch je 60.



Obr. 2.6: Rozdelenie databáze na trénovaciu a evaluačnú sadu [9].



Obr. 2.7: Výkon baseline systému pre jednotlivé triedy na development dátovej sade [9].

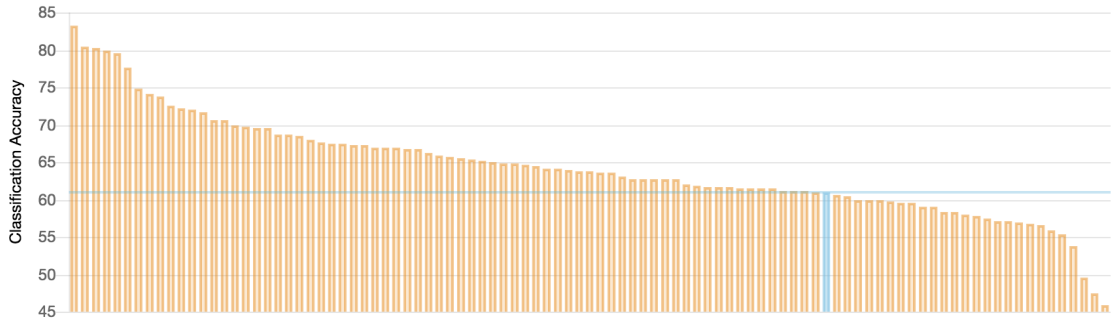
Pre každé akustické prostredie bol natrénovaný GMM model triedy s 32 komponentami na základe popísaných príznakov pomocou algoritmu *Expectation Maximization*. Štádium testovania používa rozhodnutie založené na maximálnej pravdepodobnosti spomedzi všetkých modelov tried akustických prostredí. Výkon systému je meraný pomocou úspešnosti: počet správne klasifikovaných segmentov na celkový počet testovacích segmentov. Výsledky klasifikácie pri použití kros-validačného nastavenia (obrázok 2.6) development dátovej sady zobrazuje obrázok 2.7: celkový výkon je 72.5%. Na evaluačnej dátovej sade dosiahol baseline systém úspešnosť 77.2%. Systém bol implementovaný v programovacom jazyku Python.

2.4 Prehľad výskumu pre DCASE 2017

Nasledujúci ročník súťaže prilákal viacerých autorov a záujemcov o spracovanie zvuku. Účastníkov súťaže v roku 2017 bolo spolu 41, odovzdaných systémov 97. Umiestnenie jednotlivých tímov a ich odovzdaných systémov znázorňuje obrázok 2.8. Môžme skonštatovať, že prvé miesto patrí systému s úspešnosťou 83.3%, čo je podstatný nárast v porovnaní s baseline systémom, ktorého úspešnosť je 61% na evaluačných dátach. Stručný popis najlepších systémov uvádzam v nasledujúcich podkapitolách.

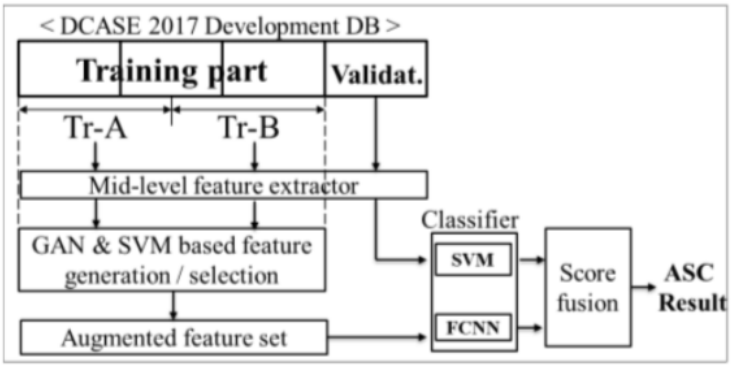
2.4.1 Seongkyu Mun

Tím na čele so Seongkyu-om Mun-om sa na túto úlohu pozeral trochu z iného pohľadu. Keďže pravidlá súťaže zakazujú použitie externých dát na trénovanie a vývoj systému,



Obr. 2.8: Graf výsledkov zúčastnených systémov súťaže DCASE 2017 [11]. Modrý stĺpec znázorňuje baseline systém poskytnutý autormi súťaže.

autori víťazného systému [18] navrhujú nagenerovať si dodatočné trénovacie dáta bez použitia externých dát. K tomu im slúži metóda *Generative Adversarial Networks (GAN).* Vzhľadom na to, že nie je jasné, či by mala každá vygenerovaná vzorka rovnaký vplyv na klasifikáciu, navrhujú použiť *Support Vector Machine (SVM)* pre selekciu vzorkov, ktoré majú diskriminatívnu informáciu o triede.



Obr. 2.9: Blokový diagram navrhovaného systému víťazného tímu DCASE 2017[18].

Navrhovaný systém je popísaný blokovým diagramom na obrázku 2.9. Autori najprv rozdelili development dátovú sadu v pomere 3:1 na 2 časti: trénovaciu a validačnú. Pre validáciu vygenerovaných vzorkov rozdelili trénovaciu časť na polovicu (Tr-A a Tr-B na obrázku 2.9). Generovanie a selekcia príznakov na základe GAN boli vykonané pre každú triedu individuálne. Preto bolo dohromady natrénovaných 15 GAN sietí. Po tom, čo boli príznakové vzorky vygenerované a vyselektované prostredníctvom GAN a SVM, boli rozšírené sady príznakov použité na trénovanie a validáciu s neurónovou sieťou *Fully Connected Neural Network (FCNN)* a SVM pre finálnu klasifikáciu.

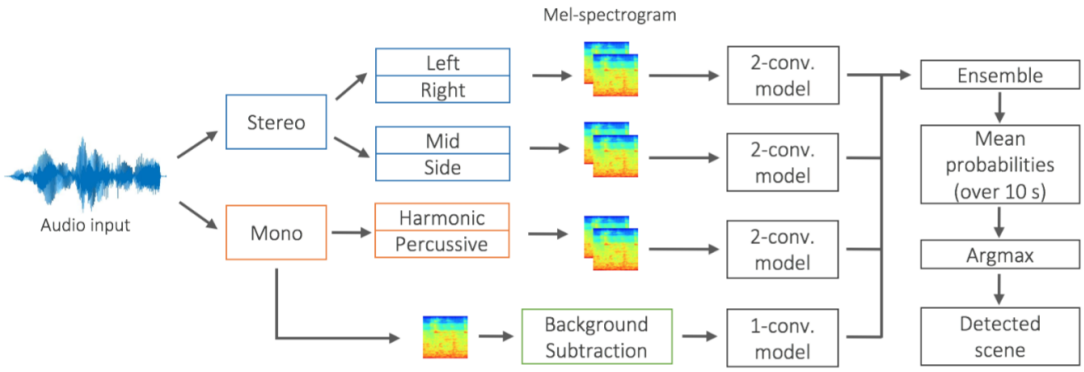
Na zlepšenie výsledkov autori ešte vykonali fúziu výsledkov z SVM a FCNN klasifikátorov. Porovnanie výsledkov úspešnosti systému s rozšírenou dátovou sadou, resp. bez nej vyobrazuje obrázok 2.10. Na evaluačných dátach dosiahol tento systém 83.3%, čím jednoznačne s náskokom necelých 3% od druhého miesta vyhral súťaž DCASE 2017.

Acc. [%]	Baseline [14]	Fusion w/o augmented DB case	Fusion on all cases
Beach	75.3	70.9	71.8
Bus	71.8	82.1	87.2
Café	57.7	71.8	87.2
Car	97.1	89.0	88.5
City	90.7	85.6	98.7
Forest	79.5	97.3	94.9
Groce.	58.7	83.3	79.5
Home	68.6	76.0	89.7
Lib.	57.1	82.0	96.2
Metro	91.7	90.7	84.6
Office	99.7	95.1	96.2
Park	70.2	69.9	71.8
Resid.	64.1	71.8	87.2
Train	58.0	71.8	82.1
Tram	81.7	84.6	91.0
Avg.	74.8	81.5	**87.1**

Obr. 2.10: Porovnanie výsledkov baseline systému a systému bez a s rozšírenou dátovou sadou. Výsledky sú vyhodnocované na development dátovej sade [18].

2.4.2 Yoonchang Han

Trojčlenný kórejský tím na čele s Yoonchang-om Han-om aplikoval použitie konvolučnej neurónovej siete vo svojom riešení [7]. Uvádzajú viacero metód predspracovania, ktoré zvýrazňujú rôzne akustické charakteristiky ako binaurálne reprezentácie, separácia harmonicko-perkusívneho zdroja a odčítanie pozadia. Architektúru celého systému schematicky znázorňuje obrázok 2.11.



Obr. 2.11: Architektúra systému navrhnutého tímom, ktorý videol Yoonchang Han. Viaceré modely konvolučnej neurónovej siete *(conv. model)* sú individuálne trénované použitím rozličných metód predspracovania a zlúčené do skupinového modelu *(Ensemble)*. Tento model potom spočíta priemerné pravdepodobnosti pre celý audio klip, aby bol schopný detekovať prostredie [7].

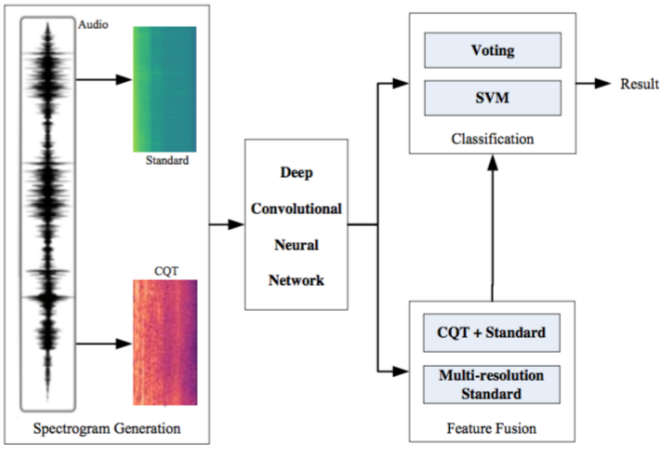
Štruktúru siete nadizajnovali za účelom obsiahnutia čo najväčšej priestorovej informácie v stereu tak, aby mala párový vstup. Vo svojom článku [7] ukazujú, ako efektívne sa navrhované štruktúry siete spolu s metódami predspracovania dokážu naučiť akustické charakteristiky z audio nahrávok a tiež, ako ich skupinový model *(ensemble)* výrazne redukuje

chybovosť klasifikácie. Tím odovzdal celkom 4 jemne líšiace sa systémy, s ktorými dosiahol úspešnosť 79.6% – 80.4% na evaluačných dátach a obsadil tak 2., 3., 4. a 5. miesto v celkovom hodnotení súťaže.

2.4.3 Xing Xiaotao

Model hlbokých konvolučných neurónových sietí *(DCNN)* vo svojom riešení aplikoval čínsky tím, ktorý viedol Xing Xiaotao [24]. Okrem toho použil na klasifikáciu akustického prostredia aj metódu fúzie viacerých spektrogramov. Schému celého systému názorne zachytáva obrázok 2.12. Najskôr sa oddelene uskutoční generovanie štandardného spektrogramu a CQT *(Constant-Q-Transform)* spektrogramu. Následným vstupom týchto spektrogramov do navrhovaného DCNN modelu dôjde k extrakcii príslušných príznakov.

Fúzia príznakov z takýchto spektrogramov je realizovaná dvoma mechanizmami - konkrétne ide o voting a SVM metódy. Autori do súťaže odovzdali dva systémy, ktoré sa líšili práve v mechanizme fúzie. Podľa nich vedie fúzia DCNN príznakov štandardného a CQT spektrogramu k podstatnému zlepšeniu úspešnosti systému v porovnaní so systémom s jedným druhom spektrogramu. Túto skutočnosť potvrdzuje i fakt, že prvý systém (SVM fúzia) obsadil 6.miesto v celkovom poradí súťaže so 77.7% a druhý systém (Voting fúzia) obsadil 7.miesto s úspešnosťou 74.8% na evaluačných dátach, čím sa stal tretím najúspešnejším tímom súťaže DCASE 2017.



Obr. 2.12: Schéma systému tretieho najúspešnejšieho tímu DCASE 2017[24].

2.4.4 Taufiq Hasan

Štvrtý tím v poradí [13] zostavil viaceré subsystémy, ktorých fúzia priniesla úspešnosť 74.1% a 8.miesto v celkovom poradí. Ich subsystémy pozostávajú hlavne z modelov založených na konvolučných neurónových sieťach (CNN), ktoré sú trénované na príznakoch spektrogramového obrazu (SIF[1]) za použitia Mel- a Log-škálovateľných bánk filtrov. Použili taktiež nový viacpásmový prístup, ktorý sa učí CNN modely z rôznych frekvenčných pásiem oddelene použitím jediného spektrogramu.

[1] Anglická skratka pre *Spectrogram Image Features*

V jednom variante týchto CNN subsystémov sa extrahujú príznakové vektory úrovne veľkorozmerných audio segmentov označované ako supervektory. Tieto supervektory sú extrahované z vyrovnávacej vrstvy natrénovaného CNN modelu a neskôr klasifikované pomocou modelu pravdepodobnostnej lineárnej diskriminatívnej analýzy (PLDA[2]). Okrem tohto subsystému implementovali autori ešte GMM supervektor systém založený na MFCC príznakoch s klasifikátorom PLDA a systém doprednej neurónovej siete založený na súbore akustických príznakov. Finálny systém pozostával z fúzie týchto subsystémov.

2.4.5 Bernhard Lehner

Tím z rakúskej univerzity v Linzi [16], ktorý vyhral súťaž v roku 2016, obsadil 9.miesto s úspešnosťou 73.8% na evaluačných dátach s fúziou i-vektor systému modelovaného pomocou MFCC príznakov odvodených z ľavého a pravého audio kanálu a hlbokej konvolučnej neurónovej siete trénovanej na raw spektrogramoch.

Okrem fúzie tento tím odovzdával aj samotné subsystémy, ktorých úspešnosť na evaluačných dátach prehľadne zobrazuje obrázok 2.13. Na základe tohto obrázku môžme skonštatovať, že i-vektor subsystém, ktorý autori zostavili, dosiahol úspešnosť 68.7%. Podľa výsledkov súťaže tento systém zaujal 20. miesto v celkovom hodnotení a bol zároveň najlepším i-vektor systémom súťaže.

(%)	Base	$IVEC_{calib}$	All_{avg}	CNN_{calib}	All_{calib}
Evaluation	61.0	68.7	66.8	64.8	73.8

Obr. 2.13: Úspešnosť jednotlivých systémov rakúskeho tímu na evaluačných dátach v porovnaní s baseline systémom súťaže [16].

Tím Bernharda Lehnera v závere svojho článku [16] konštatuje zaujímavú skutočnosť. Úspešnosť ich najlepšieho systému na development dátovej sade bola približne 90%. Tieto výsledky porovnávajú autori s výsledkami, ktoré získali z experimentu medzi svojimi študentami na univerzite, ktorí mali rovnakú úlohu s rovnakými podmienkami - klasifikovať jednotlivé audio nahrávky do zadaných tried. Úspešnosť študentov na development dátovej sade bola zhruba 50%, z čoho plynie, že úspešnosť implementovaného systému je teda až prekvapivo vysoká v porovnaní s „ľudským" klasifikátorom.

2.4.6 Baseline systém 2017

Implementácia baseline systému [9], ktorý bol poskytnutý autormi súťaže, je založená na viacvrstvovej perceptron architektúre *(Multilayer Perceptron - MLP)* a ako príznaky používa log mel-pásmové energie. Príznaky sú počítané v 40ms rámcoch s prekrytím 50% za použitia 40 mel pásiem pokrývajúc frekvenčný rozsah 0 až 22050Hz. Príznakový vektor bol skonštruovaný použitím 5-rámcového kontextu, z čoho plynie, že dĺžka tohto príznakového vektoru je 200. MLP sa skladá z dvoch dense vrstiev, pričom každá z nich má 50 skrytých jednotiek s 20% stratou.

Sieť je trénovaná algoritmom Adam pre optimalizáciu založenú na gradientoch, trénovanie je vykonané pre maximálne 200 epoch s rýchlosťou učenia 0.001 *(angl. learning rate).* Okrem toho trénovanie využíva skoré ukončovacie kritérium (tzv. *early stopping criteria*)

[2]Anglická skratka pre *Probabilistic Linear Discriminant Analysis*

s monitorovaním, ktoré začne po 100 epochách a s 10 epochovým parametrom *patience*. Výstupná vrstva pozostáva z neurónov typu softmax, ktoré reprezentujú 15 tried. Verdikt klasifikácie je založený na výstupe neurónov, pričom naraz môže byť aktívny len jeden. Na rámcoch založené rozhodnutia boli kombinované s väčšinovým hlasovaním, aby sa získalo jediné návestie *(label)* na klasifikovaný segment. Výkon systému bol meraný prostredníctvom úspešnosti definovanej ako pomer medzi počtom správnych výstupov systému a celkovým počtom výstupov. Systém bol trénovaný a testovaný na poskytnutom štvordielnom kros-validačnom setupe, čím dosiahol na development dátovej sade priemernú úspešnosť klasifikácie 73.8% a na evaluačnej dátovej sade 61.0%.

Baseline systém bol implementovaný v jazyku Python a pre strojové učenie využíval Keras. Obsahuje všetku potrebnú funkcionalitu pre zaobchádzanie s datasetom, ukladanie a prístup k príznakom a modelom a tiež pre vyhodnocovanie výsledkov. Okrem toho umožňuje priamočiaru adaptáciu a modifikáciu rôznych krokov.

2.5 DCASE 2018

V roku 2018 sa súťaž DCASE koná taktiež [12]. Organizátori tentokrát obmenili štruktúru úlohy detekcie akustického prostredia, ktorá po novom zahŕňa 3 podúlohy:

- **Detekcia akustického prostredia** – klasifikácia dát nahratých rovnakým zariadením ako dostupné trénovacie dáta
- **Detekcia akustického prostredia s rôznymi nahrávacími zariadeniami** – klasifikácia dát nahratých zariadeniami, ktoré sa líšia od tých, s ktorými boli nahraté trénovacie dáta
- **Detekcia akustického prostredia s použitím externých dát** – použitie externých dát pri trénovaní systému je povolené

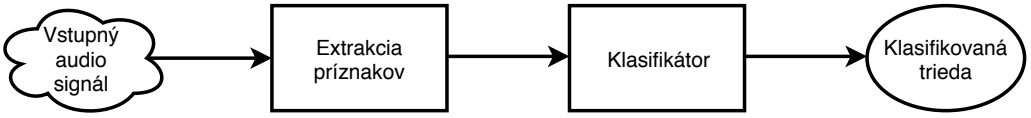
Ďalšou novinkou v tomto ročníku súťaže je fakt, že rozpoznávať sa bude len 10 rôznych akustických prostredí, ktoré sa sčasti líšia od predošlých 15 prostredí. Baseline systém implementuje metódu založenú na konvolučnej neurónovej sieti. Deadline pre odovzdávanie systémov je 31.júla 2018, pričom výsledky budú známe 15.septembra 2018.

Kapitola 3

Teoretický úvod

V tejto kapitole sú uvedené teoretické informácie súvisiace s metódami, ktoré som si na riešenie tejto práce zvolil. Vzhľadom ale na to, že teória už bola popísaná niekoľkokrát v rôznych iných prácach a článkoch, rozhodol som sa, že túto kapitolu nebudem veľmi rozpisovať. Namiesto toho radšej citujem zdroje, kde sú dané veci popísané detailnejšie a kde v prípade záujmu možno nájsť aj dodatočné informácie. O to viac som sa ale snažil venovať experimentom a ich popisom.

Štandardný postup klasifikácie reči znázorňuje obrázok 3.1. Môžme vidieť, že prvá časť sa zaoberá extrakciou príznakov zo vstupného audio signálu. Extrahované príznaky sú následne vstupom do klasifikátora, ktorý určí finálny verdikt, do ktorej z tried patrí vstupný signál.



Obr. 3.1: Bloková schéma štandardného postupu klasifikácie audio nahrávky.

Na základe prehľadu literatúry, ktorý môžme zároveň prehlásiť aj za state-of-the-art v danej oblasti, sa dá povedať, že na tému detekcie akustického prostredia možno pozerať z rôznych pohľadov za použitia viacerých metód a prístupov. Autori spomínaných článkov k riešeniu tejto úlohy použili hlavne tieto metódy:

- Gaussovský klasifikátor - zmes Gaussovských rozložení (GMM)
- i-vektor prístup
- konvolučné neurónové siete
- hlboké neurónové siete
- hlboké konvolučné neurónové siete
- klasifikátor Support Vector Machines
- klasifikátor založený na metóde najbližšieho suseda
- fúzia štandardných a CQT spektrogramov

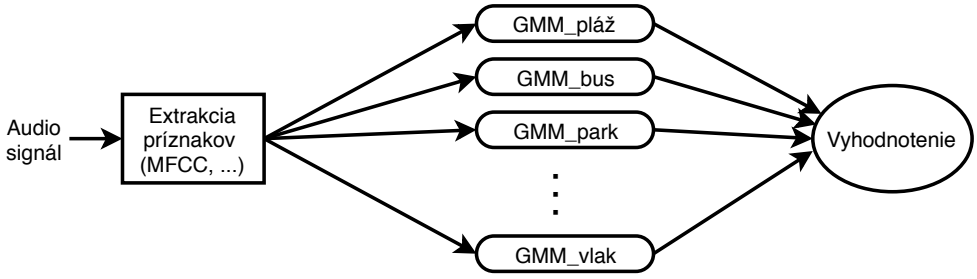
- fúzia rôznych kombinácii týchto metód

V súvislosti s použitými metódami sa líšili tiež príznaky, ktoré sa jednotlivé tímy rozhodli využiť vo svojom riešení. Jedná sa predovšetkým o príznaky pre metódy GMM a i-vektor. Najlepšie tímy hovoria o týchto príznakoch, pričom jednoznačne najčastejšie používané boli práve prvé z nich:

- MFCC koeficienty (Mel-Frequency Cepstral Coefficients)
- PLP koeficienty (Perceptual Linear Prediction)
- PNCC koeficienty (Power Normalized Cepstral Coefficients)
- RCGCC koeficienty (Robust Compressive Gamma-chirp filterbank Cepstral Coefficients)

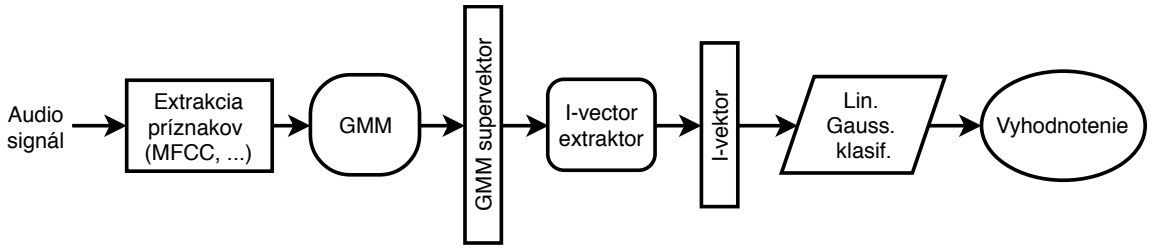
Na základe týchto zistení som sa rozhodol vo svojom riešení napokon použiť nasledujúce dva prístupy, pričom oba využívajú MFCC príznaky:

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 3.2 znázorňuje blokovú schému tohto prístupu.



Obr. 3.2: Bloková schéma GMM systému.

2. **I-vektor** – je nízkorozmerný popis celej nahrávky, ktorý je klasifikovaný Gaussovským klasifikátorom (jedna Gaussovka pre triedu). Schéma, ktorá zobrazuje tento prístup, je na obrázku 3.3.

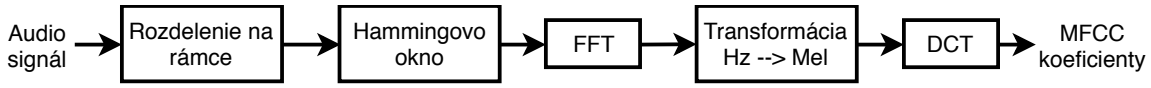


Obr. 3.3: Bloková schéma i-vektor systému.

Oba prístupy sú detailnejšie popísané v ďalšej časti, tak ako aj príznaky.

3.1 Mel-frekvenčné cepstrálne koeficienty - MFCC

MFCC koeficienty sú často používané v oblasti spracovania reči a podľa [19] fungujú dobre aj pri klasifikácii reči. Používajú sa najmä ako vstup pre klasifikátory a boli vytvorené s cieľom priblížiť sa k ľudskému počutiu. Blokový diagram na obrázku 3.4 ukazuje postupnosť krokov potrebných pre výpočet MFCC koeficientov.



Obr. 3.4: Blokový diagram zobrazujúci kroky výpočtu MFCC koeficientov.

V prvej časti celého procesu sa spojitý audio signál rozdelí do rámcov. Účel tohto rámcovania je modelovať malé úseky audio signálu, ktoré sú štatisticky stacionárne. Každý rámec pozostáva z n vzorkov, pričom od susedných rámcov je oddelený m vzorkami. Nasledujúci rámec začína m vzorkov po prvom vzorku a prekrýva prvý rámec $(n - m)$ vzorkami. Podobne tretí rámec začína m vzorkov po druhom rámci a prekrýva ho $(n - m)$ vzorkami.

V ďalšom kroku sa s cieľom minimalizovať nesúvislosti na začiatku a konci každého rámca použije tzv. *funkcia okna*. Najčastejšie sa jedná o Hammingovo okno.

Ďalšia fáza prekonvertuje každý rámec z časovej domény do frekvenčnej. Na to slúži diskrétna Fourierova transformácia (DFT) implementovaná prostredníctvom algoritmu rýchlej Fourierovej transformácie (FFT). Vzhľadom na to, že amplitúda spektra je oveľa dôležitejšia ako fáza, k ďalšiemu výpočtu postačí amplitúda, fázu zanedbáme.

Transformácia reálnej frekvenčnej stupnice do Mel-frekvenčnej stupnice predstavuje ďalší krok procesu. Prevod frekvencie v Hz na Mel-frekvenciu vyjadruje vzorec 3.1, ktorý čerpám zo študijnej opory kurzu Zpracování řečových signálů [23] , kde sa dá nájsť aj viac súvislostí a podrobností k tejto téme.

$$F_{Mel} = 2959 \log_{10}(1 + \frac{F_{Hz}}{700}) \tag{3.1}$$

Mel-frekvencia je založená na nelineárnom ľudskom vnímaní frekvencií audio signálu a jej jednotka, Mel, je jednotkou subjektívnej výšky tónu.

Vo finálnej fáze je logaritmické Mel-spektrum konvertované späť do časovej domény a výsledkom sú Mel-frekvenčné cepstrálne koeficienty. K tomu je často používaná diskrétna kosínová transformácia (DCT). Výpočet MFCC koeficientov sa teda dá súhrnne vyjadriť vzorcom 3.2.

$$c_{mf}(n) = \sum_{i=1}^{K} \log m_k \cos\left[n(k - 0.5)\frac{\pi}{K}\right] \tag{3.2}$$

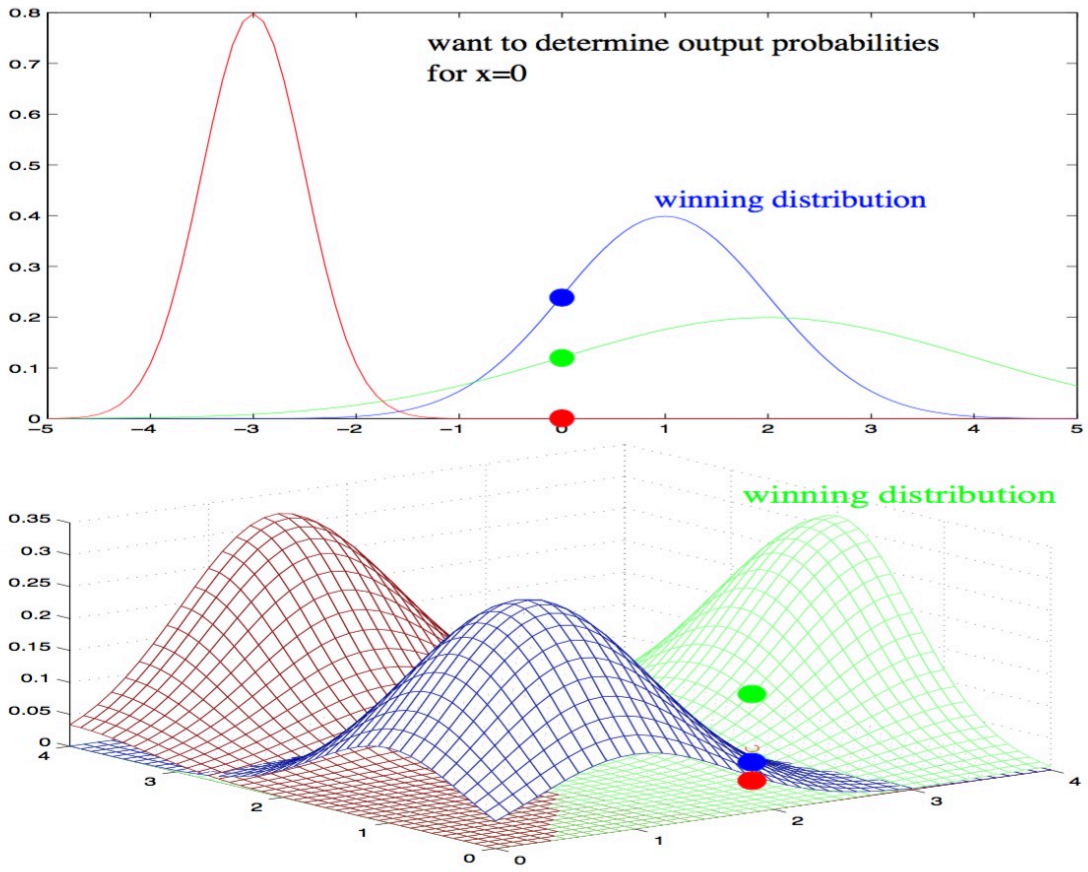
Podrobnejšie informácie o MFCC koeficientoch možno nájsť taktiež v [19].

3.2 Zmes Gaussovských rozložení

V úlohách zaoberajúcich sa automatickým rozpoznaním reči sa rozloženie akustických príznakových vektorov modeluje často pomocou zmesi Gaussovských (normálnych) rozložení – GMM[1]. Táto zmes predstavuje model tvorený váženou kombináciou Gaussovských rozložení

[1] Skratka anglického spojenia *Gaussian Mixture Model*

charakterizovaných určitou váhou, vektorom stredných hodnôt a kovariančnou maticou, ako píše aj Jan Silovský vo svojej práci [21].



Obr. 3.5: Ilustrácia modelovania jednotlivých tried pomocou jednorozmerných (hore) a dvojrozmerných (dole) Gaussovských rozložení [23].

Pravdepodobne najčastejšie využívaným spôsobom odhadu parametrov GMM je iteratívny algoritmus *Expectation Maximization* (EM) [3, 23]. Metóda maximálnej vierohodnosti (ML)[1] a metóda maximálnej aposteriórnej pravdepodobnosti (MAP)[1] patria k najčastejšie používaným metódam odhadu parametrov. Príklad zmesi Gaussovských rozložení vizuálne zobrazuje obrázok 3.5. Výpočty parametrov modelu a postupy trénovania sú detailnejšie popísané napríklad v [1],[3] a [21].

3.3 I-vektor

I-vektor predstavuje elegantný spôsob transformácie viac dimenzionálnych vstupných dát na menej dimenzionálny vektor príznakov, pričom zachováva väčšinu pôvodnej informácie [17]. Táto technika bola pôvodne inšpirovaná frameworkom *Joint Factor Analysis,* ktorý bol uvedený v [15].

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

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{3.3}$$

kde $\mathbf{m}$ je supervektor stredných hodnôt univerzálneho modelu (UBM) GMM, $\mathbf{T}$ je matica reprezentujúca M báz spájajúcich podpriestor s dôležitou variabilitou v priestore supervektoru stredných hodnôt, a $\mathbf{w}$ je vektor o veľkosti M so štandardnou normálnou distribúciou.

Pre každé pozorovanie $\mathcal{X}$ je cieľom vypočítať parametre posteriórnej pravdepodobnosti $\mathbf{w}$:

$$\mathrm{p}(\mathbf{w}|\mathcal{X}) = \mathcal{N}(\mathbf{w}; \mathbf{w}_{\mathcal{X}}, \mathbf{L}_{\mathcal{X}}^{-1}) \tag{3.4}$$

I-vektor $\boldsymbol{\phi}$ je *Maximum a Posteriori* (MAP) odhad premennej $\mathbf{w}$, t.j. stredná hodnota $\mathbf{w}_{\mathcal{X}}$ posteriórnej distribúcie $\mathrm{p}(\mathbf{w}|\mathcal{X})$. Mapuje väčšinu relevantnej informácie z pozorovania premennej dĺžky na vektor pevného (malého) rozmeru. $\mathbf{L}_{\mathcal{X}}$ je presnosť posteriórnej distribúcie. $\mathbf{T}$ sa označuje ako i-vektor extraktor. Viac informácií ohľadom i-vektorov možno nájsť napríklad v [6, 17, 14].

3.4 Lineárny Gaussovský klasifikátor

Generatívne modelovanie i-vektor odhadov pre rozpoznávanie jazyka sa prejavilo ako efektívna alternatíva k diskriminatívnym klasifikátorom založeným na logistickej regresii alebo Support Vector Machines. V [17] autori navrhli jednoduchý lineárny klasifikátor založený na Gaussovských rozloženiach, ktoré poskytujú presnosti podobné presnostiam lineárnych diskriminatívnych prístupov. Model predpokladá, že pre každé prostredie sú generované odpovedajúce i-vektor odhady $\boldsymbol{\mu}_i$ podľa:

$$\boldsymbol{\mu}_i \sim \mathcal{N}\left(\mathbf{m}_\ell, \boldsymbol{\Lambda}^{-1}\right) \ , \tag{3.5}$$

kde $\mathbf{m}_\ell$ je jazykovo závislý vektor stredných hodnôt a $\boldsymbol{\Lambda}^{-1}$ je kovariančná matica, zdieľaná medzi všetkými triednymi distribúciami. Parametre modelu sa potom dajú ľahko získať metódou maximálnej vierohodnosti. Triedou podmienená logaritmická vierohodnosť pre $\boldsymbol{\mu}_i$ dané prostredie ℓ môže byť spočítaná nasledovne:

$$\log P(\boldsymbol{\mu}_i|\ell) = \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}(\boldsymbol{\mu}_i - \mathbf{m}_\ell)^T\boldsymbol{\Lambda}(\boldsymbol{\mu}_i - \mathbf{m}_\ell) + k \ , \tag{3.6}$$

kde k je dátovo nezávislá konštanta. Takýto klasifikátor bude označovaný ako GLC. Viac informácií o GLC možno nájsť napr. v [22].

Kapitola 4

Dáta

Všetky audio dáta, ktoré sú popísané v nasledujúcich 2 sekciách, boli nahrávané ako stereo so vzorkovacou frekvenciou 44.1 kHz, s 24 bitovou presnosťou a sú uložené ako súbory typu WAV.

4.1 TUT Acoustic scenes 2016

Dataset z roku 2016 obsahoval pôvodne nahrávky o dĺžke 3 až 5 minút. Následne boli všetky tieto nahrávky rozdelené na 30 sekundové segmenty.

Dataset sa skladá z dvoch častí:

- **Development dataset** - pre každé z 15 prostredí je k dispozícii 78 segmentov, čiže 39 minút audia.
- **Evaluation dataset** - každé prostredie disponuje 26 segmentami, čo predstavuje 13 minút audia.

Spolu teda development dataset obsahuje 9 hodín a 45 minút audio nahrávok v 1170 segmentoch a evaluation dataset 3 hodiny a 15 minút v 390 segmentoch.

4.2 TUT Acoustic scenes 2017

Pôvodné nahrávky dĺžky 3 až 5 minút boli rozdelené na 10 sekundové segmenty. Každý segment predstavuje osobitný súbor.

Dataset sa skladá taktiež z dvoch častí:

- **Development dataset** - pozostáva z kompletného datasetu *TUT Acoustic scenes 2016* (development + evaluation dáta), pričom pre každé prostredie je k dispozícii 312 segmentov, takže 52 minút audia.
- **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.

Celkovo obsahuje development dataset 13 hodín audio nahrávok v 4680 segmentoch (súboroch) a evaluation dataset 4 hodiny a 30 minút v 1620 segmentoch (súboroch).

Pre oba tieto datasety platí, že som development dataset použil na trénovanie tak GMM systému ako aj i-vektor systému. Evaluation dataset bol zase použitý na reportovanie výsledkov.

4.3 Obmedzenia súťaže DCASE

Pre súťaž platia tieto všeobecné pravidlá:

- Nie je povolené použitie externých dát za účelom vývoja systému. Dáta z inej úlohy súťaže sú tiež považované za externé dáta.
- Manipulácia poskytnutých dát je povolená, development dataset môže byť rozšírený bez použitia externých dát, napríklad technikou *pitch shifting* alebo *time stretching*, a podobne.
- Súťažiaci nesmú subjektívne posudzovať evaluačné dáta a tiež ich nemôžu ani anotovať. Evaluation dataset nemôže byť použitý na trénovanie systému, zakázané je aj použitie štatistík o evaluačných dátach vo fáze rozhodovania systému.

Podrobnejšie informácie týkajúce sa celej súťaže možno nájsť v článkoch organizátorov oboch ročníkov [9] a [8].

4.4 Evaluačná metrika

Vyhodnocovanie klasifikácie akustického prostredia je založené na klasifikačnej metrike, ktorá je definovaná ako počet správne klasifikovaných (testovacích) nahrávok delený celkovým počtom (testovacích) nahrávok. Každá nahrávka sa považuje za nezávislú testovaciu vzorku.

Kapitola 5

Experimenty a výsledky s GMM

V tejto kapitole budú popísané a diskutované konkrétne experimenty týkajúce sa implementovaného GMM systému, ktoré som zrealizoval. Okrem toho tu takisto spomeniem všetko to, čo so samotnými experimentami súvisí a bez čoho by experimenty nebolo možné spúšťať. Najskôr by som teda rád objasnil, akým spôsobom som sa rozhodol experimenty robiť, aký systém práce som si vlastne zvolil.

5.1 Formát GMM experimentov

Hneď zo začiatku riešenia tejto práce bolo jasné, že tu bude treba vykonať nemálo experimentov na to, aby som mohol konštatovať nejaké zmysluplné a dôveryhodné závery. Potreboval som si preto vytvoriť jednotný systém práce, ktorý by mi umožňoval kedykoľvek sa vrátiť k ľubovoľnému experimentu a spustiť si ho znovu za rovnakých podmienok (s rovnakými vstupnými parametrami apod.), aby vrátil rovnaký výsledok.

Takýto jednotný systém práce považujem za kľúčový postup, vďaka ktorému je možné predísť situácii, pri ktorej by som sa síce dopracoval k dobrému výsledku, no nebol by som ho schopný zreplikovať a spätne potvrdiť. Týmto spôsobom som tiež chcel eliminovať prípadný chaos, ktorý by asi ľahko vznikol v prípade veľkého množstva novovytvorených súborov bez nejakého systému a štruktúry.

Každý experiment v tejto práci všeobecne pozostáva z dvoch častí - z trénovania a evaluácie. Pre každú z týchto častí je k dispozícii príslušný shell skript, ktorý danú funkciu vykoná. Nasledujúce riadky popisujú detailnejšie tieto skripty.

5.1.1 Trénovací skript

Shell skript, ktorý sa zaoberá trénovaním systému, pracuje nasledovným spôsobom. Rozlišovacím faktorom jednotlivých skriptov je ich názov, ktorý zároveň hovorí o tom, čo za systém ten-ktorý skript trénuje. Názov trénovacieho skriptu sa skladá z 10 častí, ktoré sú navzájom oddelené podtržníkom a vyzerá takto:

ID_{run}_{train}_N{G}_DATA_FEA_NORM_N{it}_FS_{GMM.sh}

pričom význam jednotlivých častí je nasledujúci:

- **ID** – jednoznačný identifikátor každého experimentu, ktorý sa pri vytvorení ďalšieho trénovacieho skriptu automaticky inkrementuje. Je reprezentovaný trojciferným celým číslom počínajúc nulou, napr. *025*.
- **{run}** – konštantný textový reťazec *run*

- **{train}** – konštantný textový reťazec *train*
- **N{G}** – počet Gaussoviek trénovaného GMM modelu (napr. *512G*)
- **DATA** – špecifikácia datasetu, ktorý má byť pri trénovaní použitý, skript podporuje tieto 4 datasety:
  - *development2016* – development dataset súťaže DCASE 2016
  - *evaluation2016* – evaluačný dataset súťaže DCASE 2016
  - *development2017* – development dataset súťaže DCASE 2017
  - *evaluation2017* – evaluačný dataset súťaže DCASE 2017

  Pre trénovací skript sú typickejšie development datasety a pre evaluačný skript zase evaluačné, ale je ich možné použiť i opačne.
- **FEA** – špecifikácia príznakov, ktoré majú byť pri trénovaní extrahované z audio signálu. Skript podporuje dve varianty príznakov – banky filtrov a MFCC koeficienty, ktoré budú pre prehľadnosť popísané za týmto odstavcom popisujúcim časti trénovacieho skriptu.
- **NORM** – špecifikácia normalizácie príznakov, ktorá sa má pri ich spracovaní použiť. Skript berie do úvahy 3 možnosti:
  - *NN* – žiadna normalizácia
  - *MN* – mean normalizácia
  - *MVN* – mean and variance normalizácia
- **N{it}** – počet iterácií potrebných na natrénovanie GMM modelu (napr. *10it*)
- **FS** – vzorkovacia frekvencia trénovacích dát v Herzoch, ktorá sa má pri trénovaní použiť. Skript počíta s troma variantmi: *8000, 16000* a *44100*.
- **{GMM.sh}** – konštantný textový reťazec *GMM.sh*

Ako bolo v popise spomenuté, skript podporuje 2 druhy príznakov – banky filtrov a MFCC koeficienty. **FEA** časť z názvu skriptu má pre každý druh z týchto príznakov odlišný tvar:

- **{FBANK}FBC**:
  - **{FBANK}** – konštantný reťazec *FBANK,* ktorý určuje druh použitých príznakov. V tomto prípade sa teda jedná o banky filtrov.
  - **FBC** – počet bánk filtrov, ktoré sa majú použiť pri extrakcii príznakov, napr. *20, 32, 40, 48, ...*
- **{MFCC}CC-ZERO-DERIV1-DERIV2-FBC-CONTEXT**:
  - **{MFCC}** – konštantný reťazec *MFCC* určujúci druh príznakov. V tomto prípade ide o MFCC koeficienty.
  - **CC** – počet MFCC koeficientov, ktoré majú byť extrahované z audio signálu, napr. *12, 20, ...*

- **ZERO** – špecifikácia nultého koeficientu, možné sú tri alternatívy:
  - *C0* – pridanie nultého MFCC koeficientu k priamym koeficientom
  - *E* – nultým koeficientom bude energia jednotlivých koeficientov, ktorá bude pridaná ako nultý koeficient k priamym koeficientom
  - *NC0* – nultý koeficient nebude vôbec pripojený k priamym MFCC koeficientom
- **DERIV1** – (ne)pripojenie odvodených koeficientov (tzv. delta koeficienty) k priamym koeficientom:
  - *D* – pripojenie delta koeficientov k priamym koeficientov
  - *ND* – delta koeficienty nebudú zahrnuté vôbec
- **DERIV2** – (ne)pridanie odvodených (akceleračných) koeficientov 2.rádu (tzv. doubledelta koeficienty) k priamym a delta koeficientom:
  - *DD* – pripojenie doubledelta koeficientov k priamym a delta koeficientom
  - *NDD* – doubledelta koeficienty nebudú zahrnuté vôbec
- **FBC** – počet bánk filtrov, pomocou ktorých budú extrahované MFCC koeficienty, napr. *20, 32, 40, 48, ...*
- **CONTEXT** – parameter určujúci kontext, ktorý sa berie do úvahy v prípade výpočtu delta a doubledelta koeficientov. Možné hodnoty sú *1, 2* alebo *3.*

Pre lepšiu predstavu uvádzam konkrétny príklad názvu trénovacieho skriptu:
`186_run_train_128G_development2017_MFCC20-C0-D-DD-24-1_NN_4it_16000_GMM.sh`
Takýto trénovací skript natrénuje GMM model pomocou 128 Gaussoviek (so 4 iteráciami) na development dátach DCASE 2017 so vzorkovacou frekvenciou 16000 Hz. Extrahované príznaky predstavujú 20 priamych MFCC koeficientov rozšírených o nultý MFCC koeficient, ku ktorým sú pripojené delta a doubledelta koeficienty získané kontextom 1. Tieto MFCC koeficienty budú vytvorené prostredníctvom 24 bánk filtrov a nebudú žiadnym spôsobom normalizované.

Funkcionalita trénovacieho skriptu

Popisovaný trénovací skript na začiatku svojho vykonávania vytvorí na tej istej úrovni zložku s rovnakým názvom, ktorá sa líši akurát v koncovke `.dir`. Táto zložka, nazvime ju *výstupná zložka experimentu,* slúži na uchovávanie všetkých výstupov z (nielen) trénovacieho skriptu, čím je zabezpečené spätné dohľadanie potrebných informácií o tom-ktorom experimente. V tejto zložke ďalej trénovací skript vytvorí 3 podzložky:

- `GMM/ -` slúži na uloženie natrénovaných modelov (modelov je viac, pretože každý trénovací skript vytvorí jeden model na jeden počet Gaussoviek, čiže napr. pri počte Gaussoviek 128 sa uložia modely aj pre všetky menšie počty Gaussoviek, t.j. 1, 2, 4, 8, 16, 32 a 64)
- `results/ -` slúži na uloženie textového výstupu tak trénovacieho ako aj evaluačného skriptu
- `scripts/ -` slúži na uloženie všetkých zdrojových súborov, ktoré boli k vykonaniu skriptu potrebné

Poslednou a zároveň najhlavnejšou fázou trénovacieho skriptu je spustenie pythonového skriptu s parametrami, ktoré boli špecifikované v názve trénovacieho skriptu. Úlohou tohto pythonového skriptu je natrénovať žiadané GMM modely so špecifikovanými parametrami, ktoré následne uloží do podzložky `GMM/`. Textový výstup zaznamenávajúci priebeh trénovania sa uloží do podzložky `results/`. Tento výstup je použitý na kontrolu, či trénovanie prebehlo v poriadku a nenastal žiaden problém.

5.1.2 Evaluačný skript

Evaluačný skript je shell skript, ktorý sa zaoberá evaluáciou natrénovaného systému. Názov tohto skriptu je taktiež kľúčový ako aj v prípade trénovacieho skriptu, no vďaka pomocnému skriptu sa automaticky odvodí od názvu trénovacieho skriptu, jeho formát je rovnaký. Vzhľadom na to, že nemožno evaluovať systém, ktorý nie je natrénovaný, skript predpokladá, že trénovací skript už existuje, a preto môže byť jeho názov odvodený od toho trénovacieho. Jediná odlišnosť v názve evaluačného skriptu je tá, že namiesto reťazca **{train}** figuruje reťazec **{evaluate}**.

Pre úplnosť uvádzam tiež príklad názvu evaluačného skriptu k trénovaciemu skriptu, ktorý je uvedený v predošlej sekcii:
`186_run_evaluate_128G_evaluation2017_MFCC20-C0-D-DD-24-1_NN_4it_16000_GMM.sh`
Tento evaluačný skript vyhodnotí na evaluačných dátach súťaže DCASE 2017 všetky systémy, ktoré boli natrénované príslušným trénovacím skriptom. V tomto prípade sa jedná o systémy s 1, 2, 4, 8, 16, 32, 64 a 128 Gaussovkami, t.j. evaluačný skript vyhodnotí celkovo 8 systémov.

Funkcionalita evaluačného skriptu

Evaluačný skript na začiatku svojho vykonávania skontroluje, či sú natrénované všetky modely, ktoré sa chystá evaluovať a či je tiež vytvorená výstupná zložka experimentu vrátane všetkých svojich podzložiek, ktorých vytvorenie má na starosti trénovací skript. V prípade, že je všetko v poriadku, nasleduje samotná evaluácia jednotlivých systémov pomocou špeciálneho pythonového skriptu, ktorý evaluuje postupne tie systémy, ktoré plynú zo vstupných parametrov špecifikovaných v názve evaluačného shell skriptu.

Výsledky vyhodnotenia uloží pre každý evaluovaný systém do osobitného textového súboru v rámci zložky `results/`.

5.2 GMM systém založený na bankách filtrov

V tejto sekcii bude analyzovaný GMM systém, ktorý bol vybudovaný na príznakoch označovaných ako *banky filtrov.* Úspešnosť takéhoto systému môže byť ovplyvňovaná niekoľkými parametrami, s ktorými je možné experimentovať. Ja som prostredníctvom takýchto experimentov hľadal vhodnú kombináciu parametrov, ktorá by vo výsledku viedla k čo možno najúspešnejšiemu systému detekcie akustického prostredia.

V tejto sekcii budú postupne predstavené jednotlivé experimenty a hlavne ich výsledky tak, ako aj osobitné kombinácie parametrov a ich vplyv na úspešnosť systému. Vo svojich experimentoch som sa venoval konfigurácii nasledovných parametrov systému, ktoré možno zároveň nazvať kategóriami experimentov:

1. počet príznakov – bánk filtrov

2. normalizácia príznakov
3. vzorkovacia frekvencia vstupných audio nahrávok
4. audio kanál vstupných audio nahrávok
5. počet Gaussoviek

Nasledujúce podsekcie popisujú jednotlivé parametre systému, ktoré boli predmetom experimentov a tiež prezentujú úspešnosť osobitných GMM systémov v podobe grafov. Spočiatku bolo treba stanoviť východiskové hodnoty týchto parametrov, keďže napríklad v prípade prvej kategórie experimentov založených na počte príznakov ešte nevieme, aká normalizácia je pre daný GMM systém najvhodnejšia. Prehľad východiskových hodnôt, ktoré sú v prípade potreby použité v experimentoch, je zahrnutý v tabuľke 5.1.

Tabuľka 5.1: Prehľad východiskových hodnôt jednotlivých parametrov pre GMM systém založený na bankách filtrov.

Parameter	**Východisková hodnota**
Normalizácia	žiadna
Vzorkovacia frekvencia	16 kHz
Audio kanál	pravý
Počet Gaussoviek [1]	1 – 1024

Pokiaľ nie je explicitne uvedené inak, sú všetky sytémy spomenuté v ďalšej časti práce trénované na development dátovej sade zo súťaže DCASE 2017 a evaluované na evaluačnej dátovej sade súťaže DCASE 2017. Radšej som experimentoval s datasetom súťaže z roku 2017. Táto dátová sada je ťažšia, obsahuje kratšie nahrávky a má ich viac ako dáta z roku 2016. Preto sú aj výsledky vierohodnejšie.

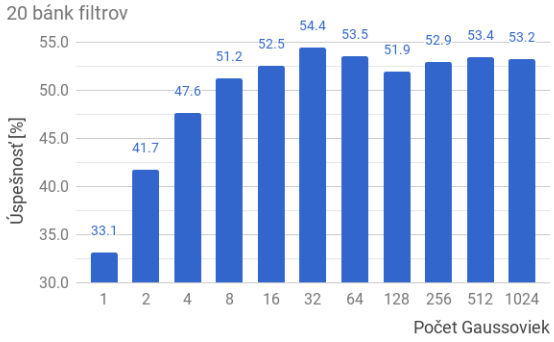
5.2.1 Počet bánk filtrov

Cieľom experimentov týkajúcich sa počtu príznakov – bánk filtrov, bolo zistiť, ktorý systém s koľkými bankami filtrov dosahuje najvyššiu úspešnosť detekcie akustického prostredia. To znamená, že tieto experimenty boli zakaždým spustené s východzími hodnotami parametrov (pozri tabuľku 5.1), pričom sa v jednotlivých experimentoch menil iba počet bánk filtrov, aby bolo zrejmé, do akej miery sa tento činiteľ podieľa na finálnej úspešnosti systému.

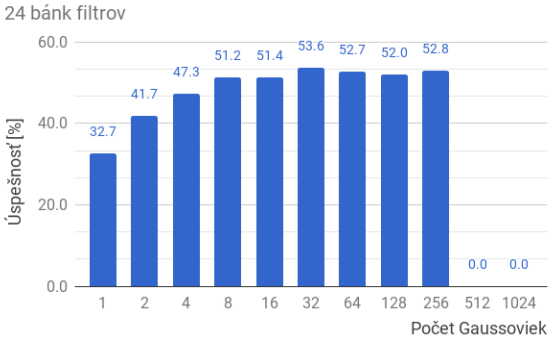
Obrázok 5.1 zobrazuje úspešnosti GMM systémov líšiacich sa v počte bánk filtrov, ktoré predstavujú príznaky. Ako reprezentatívnu vzorku mojich experimentov som sa rozhodol použiť 8 rozličných počtov bánk filtrov. Konkrétne počty som stanovil na základe literatúry a odborných článkov, ktoré som čítal na túto tému, podľa toho, koľko autori bežne používali v rámci svojich systémov. Jedná sa teda o systémy s:

- 20 bankami filtrov (graf s úspešnosťou systému ilustruje obrázok 5.1a)
- 24 bankami filtrov (obrázok 5.1b)
- 28 bankami filtrov (obrázok 5.1c)

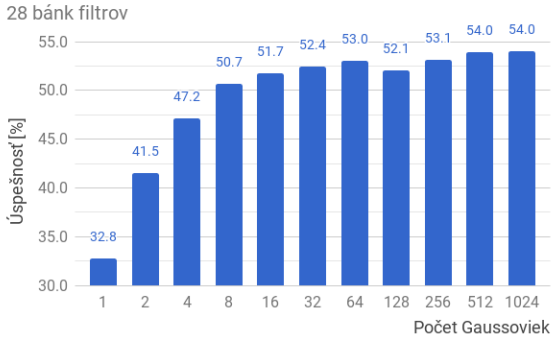
[1] Každý experiment bol spustený pre 1 až 1024 Gaussoviek, aby bolo nejakým spôsobom možné sledovať vývoj úspešnosti systému



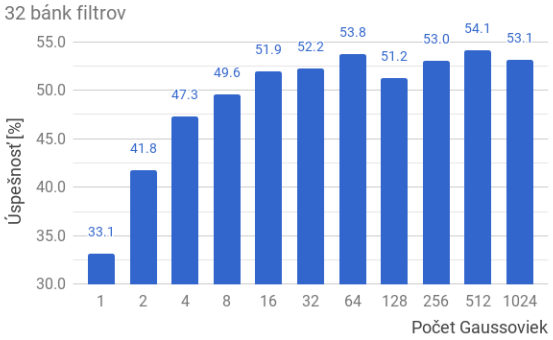
(a) 20 bánk filtrov



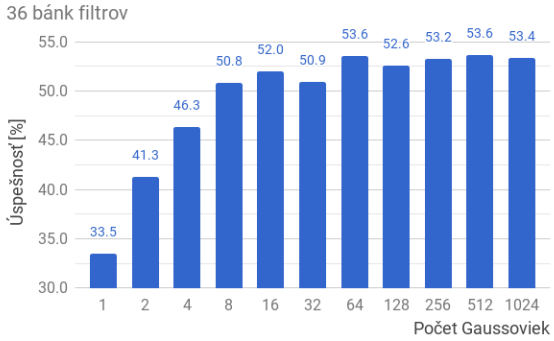
(b) 24 bánk filtrov



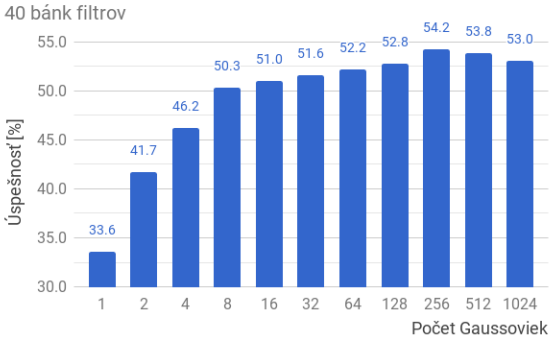
(c) 28 bánk filtrov



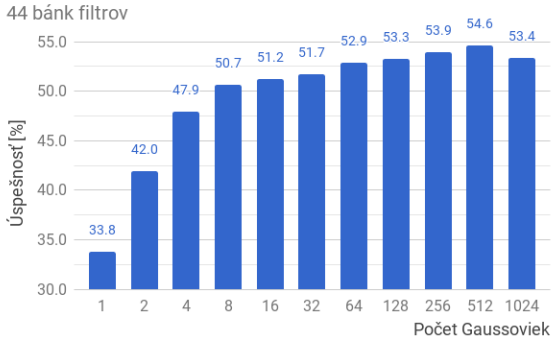
(d) 32 bánk filtrov



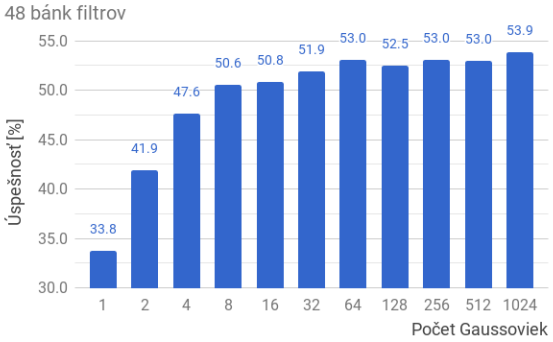
(e) 36 bánk filtrov



(f) 40 bánk filtrov



(g) 44 bánk filtrov



(h) 48 bánk filtrov

Obr. 5.1: Grafy znázorňujúce vplyv počtu príznakov na úspešnosť GMM systému. V tomto prípade sú príznaky reprezentované bankami filtrov.

- 32 bankami filtrov (obrázok 5.1d)
- 36 bankami filtrov (obrázok 5.1e)
- 40 bankami filtrov (obrázok 5.1f)
- 44 bankami filtrov (obrázok 5.1g)
- 48 bankami filtrov (obrázok 5.1h)

Na základe grafov z obrázku 5.1 môžme usúdiť, že najvhodnejším počtom bánk filtrov z hľadiska úspešnosti GMM systému je 44. Takýto systém dosiahol úspešnosť 54.6%, čo je najviac spomedzi 8 uvažovaných systémov. S úspešnosťou 54.4% je druhým najúspešnejším systémom ten s 20 bankami filtrov. Tretie miesto v poradí obsadil systém so 40 bankami filtrov a úspešnosťou 54.2%, štvrté miesto patrí systému s 32 bankami filtrov a úspešnosťou 54.1%.

Na základe výsledkov môžme teda konštatovať, že úspešnosť GMM systému nerastie a ani neklesá priamo úmerne s počtom bánk filtrov, nie je medzi nimi priama závislosť.

5.2.2 Normalizácia bánk filtrov

Ďalšou kategóriou experimentov, ktorým som sa venoval, bolo zahrnutie, resp. nezahrnutie normalizácie príznakov. Jedná sa o normalizáciu pre každú nahrávku, tzv. *file based normalizáciu.* Ako bolo spomenuté v predošlej časti, experimenty s počtom príznakov využívali východiskovú hodnotu normalizácie, t.j. príznaky neboli žiadnym spôsobom normalizované.

V tejto časti nadviažem na výsledky z predošlej časti a na 4 najúspešnejších systémoch vykonám pokusy s normalizáciou príznakov za účelom zistenia jej vplyvu na úspešnosť systému. V súvislosti s normalizáciou som sa v tejto práci zaoberal nasledovnými troma možnosťami:

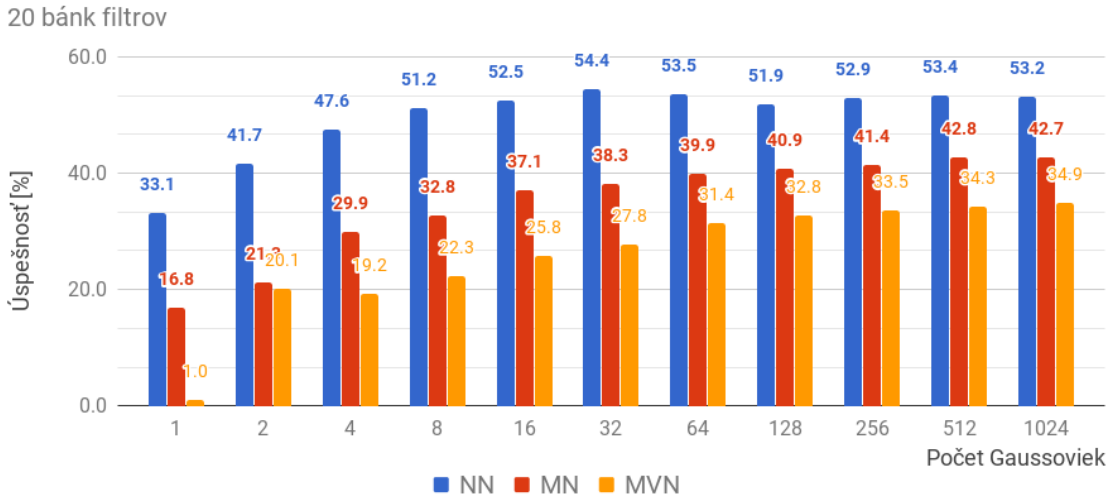
- *žiadna normalizácia* – v grafoch označovaná ako NN (východisková hodnota)
- *mean normalizácia* – v grafoch označovaná ako MN
- *mean and variance normalizácia* – v grafoch označovaná ako MVN

Obrázok 5.2 znázorňuje grafy dvoch konfigurácií systémov, ktoré po aplikovaní uvažovaných možností normalizácie dosahovali najlepšie výsledky. Pri pohľade na tieto grafy je celkom zjavné a jednoznačné, aký účinok mala aplikácia mean normalizácie na úspešnosť systémov, a tiež ako vplýva aplikovanie mean and variance normalizácie na finálnu úspešnosť GMM systému.

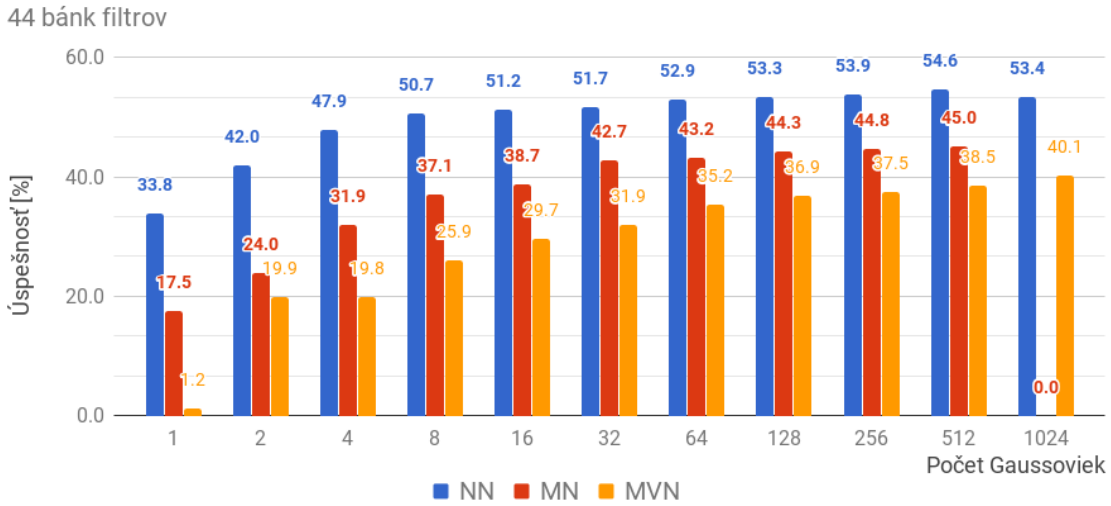
V prípade, keď príznaky systému nie sú normalizované, dosahuje jeho úspešnosť suverénne najlepšie výsledky. Naproti tomu sa použitie mean normalizácie tak ako aj použitie mean and variance normalizácie podpisuje negatívnym spôsobom na celkovej úspešnosti. V prípade systému so 44 bankami filtrov (pozri obrázok 5.2b) sa jedná o pokles úspešnosti v rozsahu zhruba od 9 až po 18% absolútne, keď berieme do úvahy mean normalizáciu. V prípade mean and variance normalizácie je pokles ešte väčší a to priemerne zhruba až o 20% absolútne. Pri pohľade na graf pre systémy s 20 bankami filtrov (pozri obrázok 5.2a), môžme vidieť veľmi obdobný vývoj úspešnosti jednotlivých systémov v závislosti na aplikovanej normalizácii.

Túto kategóriu experimentov možno uzavrieť konštatovaním, že pre systémy detekcie akustického prostredia je najprospešnejšie, keď nie sú extrahované príznaky (banky filtrov)

normalizované žiadnym z uvedených spôsobov, pretože ani mean normalizácia, ani mean and variance normalizácia sa neprejavili pozitívnym spôsobom na úspešnosti testovacích GMM systémov. Systémy vyhodnotené ako najúspešnejšie z predošlej kategórie experimentov teda zostávajú najúspešnejšie aj po vykonaní experimentov s normalizáciou.



(a) Systémy s 20 bankami filtrov.

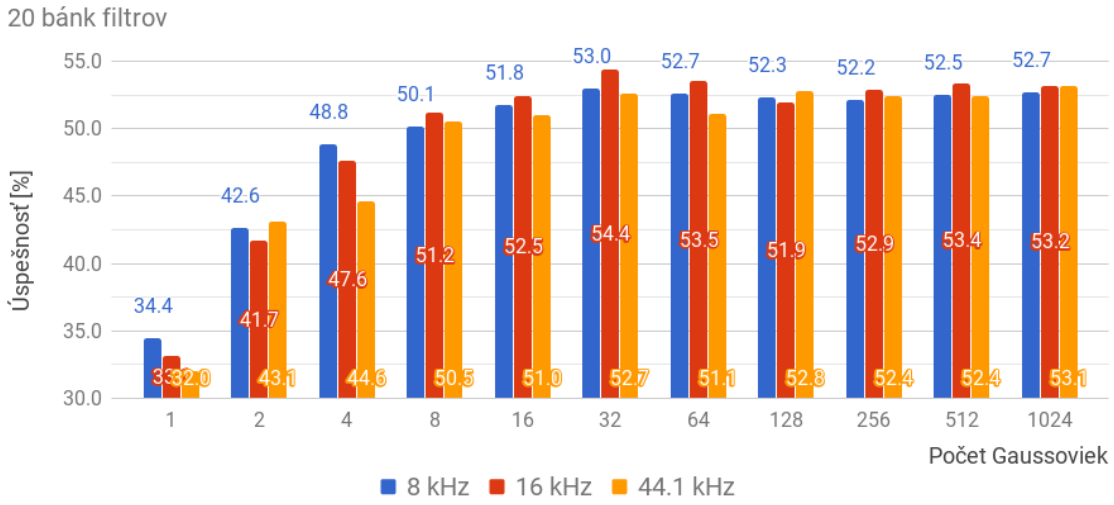


(b) Systémy so 44 bankami filtrov.

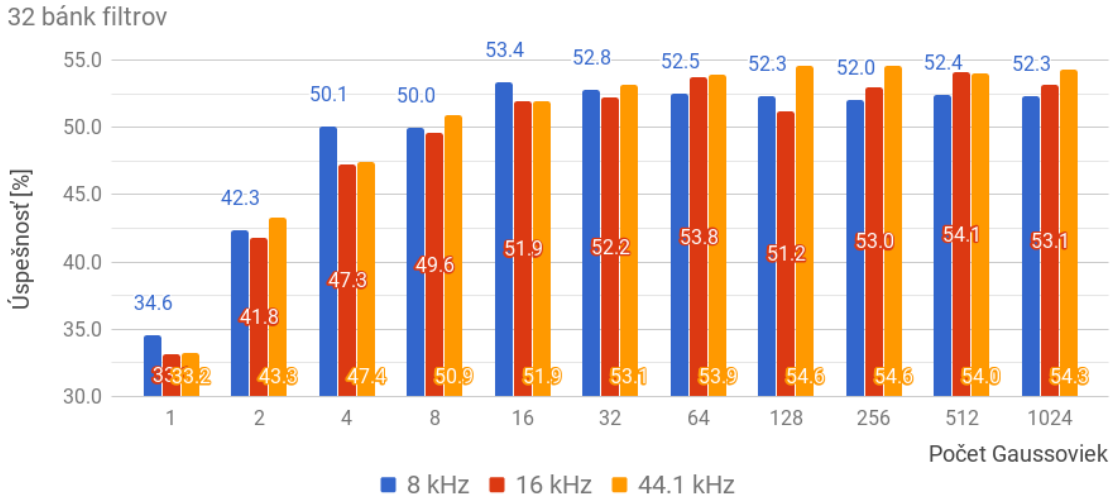
Obr. 5.2: Grafy znázorňujúce vplyv normalizácie príznakov na úspešnosť GMM systémov líšiacich sa v počte bánk filtrov *(NN = žiadna normalizácia, MN = mean normalizácia, MVN = mean and variance normalizácia).*

5.2.3 Vzorkovacia frekvencia vstupných audio nahrávok

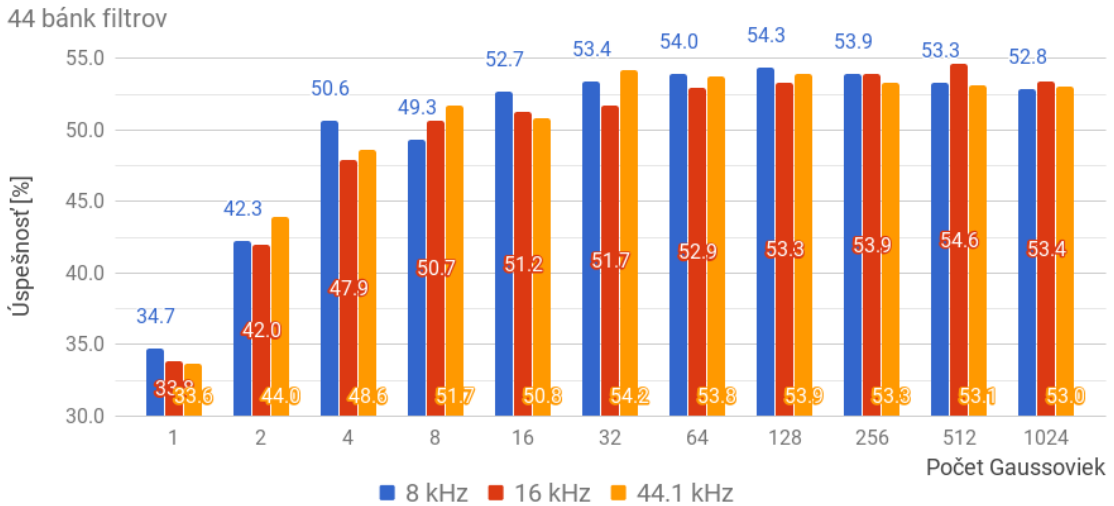
V tejto kategórii experimentov sa zameriame na vzorkovaciu frekvenciu audio nahrávok, ktoré sú vstupom do systému. Cieľom je opäť nájsť závislosť medzi vzorkovacou frekvenciou a úspešnosťou systému. Tento experiment bol taktiež vykonaný na najlepších 4 systémoch z predošlej kategórie, ktoré sú zároveň najlepšími systémami prvej kategórie experimentov



(a) Systémy s 20 bankami filtrov.



(b) Systémy s 32 bankami filtrov.



(c) Systémy so 44 bankami filtrov.

Obr. 5.3: Grafy znázorňujúce vplyv vzorkovacej frekvencie na úspešnosť GMM systému. Vplyv frekvencie je zachytený na troch rôznych konfiguráciach systémov líšiacich sa v počte bánk filtrov.

(vzhľadom na negatívny vplyv použitia normalizácie). Na účely tohto experimentu boli testované nasledovné vzorkovacie frekvencie vstupného audio signálu:

- *8 kHz*
- *16 kHz* – východisková hodnota pre predošlé experimenty
- *44.1 kHz*

Obrázok 5.3 zobrazuje 3 konfigurácie systémov, ktoré v tomto experimente dosiahli najvyššiu úspešnosť. Grafy naznačujú, že nemožno jasne povedať, ktorá zo vzorkovacích frekvencií je pre GMM systém najprospešnejšia, záleží to na kombinácii ostatných parametrov systému.

Najlepší výsledok dosiahol systém so 44 bankami filtrov (obrázok 5.3c) v kombinácii so vzorkovacou frekvenciou 16 kHz a s úspešnosťou 54.63%. Druhým najúspešnejším je systém s 32 bankami filtrov (obrázok 5.3b), vzorkovacou frekvenciou 44.1 kHz a úspešnosťou 54.57%. Tretím najúspešnejším je systém s 20 bankami filtrov (obrázok 5.3a), so vzorkovacou frekvenciou 16 kHz a úspešnosťou 54.4%. Štvrtým najlepším je opäť systém so 44 bankami filtrov, so vzork. frekvenciou 8 kHz a s úspešnosťou 54.3%.

Záverom môžme povedať, že na základe grafov na obrázku 5.3 má vzork. frekvencia 8 kHz v priemere najmenšiu úspešnosť v porovnaní s ostatnými 2 alternatívami, no zároveň nie je možné všeobecne a jednoznačne určiť, či je lepšia frekvencia 16 kHz alebo 44.1 kHz. Subjektívne možno postrehnúť, že o trochu lepšie výsledky dosiahla vzork. frekvencia 16 kHz, pretože medzi prvými štyrmi najlepšími systémami sa nachádzajú práve dva s touto frekvenciou.

5.2.4 Audio kanál vstupných nahrávok

Poslednou kategóriou experimentov v súvislosti s GMM systémami založenými na bankách filtrov je výber a otestovanie rôznych audio kanálov vstupných nahrávok. Originálne nahrávky sú totiž stereo nahrávky toho istého zvuku, čiže obidva audio kanály sú veľmi podobné, avšak nie sú rovnaké. Na základe článkov autorov súťaže, ktorí rôznym spôsobom kombinovali jednotlivé audio kanály vstupných nahrávok, som sa rozhodol zistiť, ako sa tieto audio kanály prejavia na úspešnosti GMM systému. Do úvahy som bral takéto varianty:

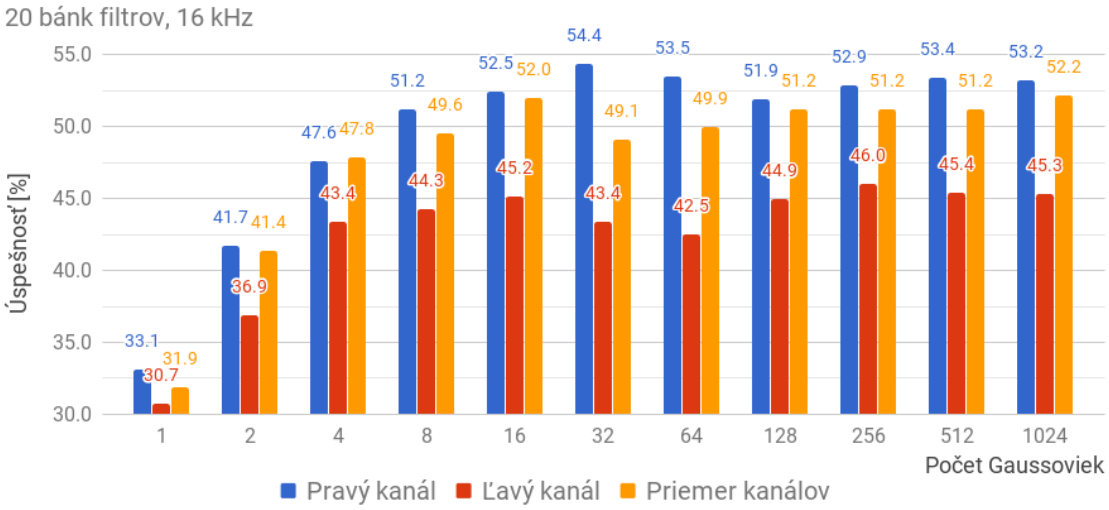
- *pravý audio kanál* – východisková hodnota predošlých experimentov
- *ľavý audio kanál*
- *priemer pravého a ľavého kanálu*

Obrázky 5.4 a 5.5 zobrazujú grafy 4 doteraz najúspešnejších konfigurácií GMM systémov v súvislosti s vplyvom jednotlivých variant audio kanálov na úspešnosť takýchto systémov. Trend grafov je vzhľadom k variantom audio kanálov rovnaký, resp. veľmi podobný pre všetky 4 konfigurácie, čím možno s určitosťou prehlásiť, aká závislosť v tomto prípade jestvuje medzi jednotlivými variantmi.

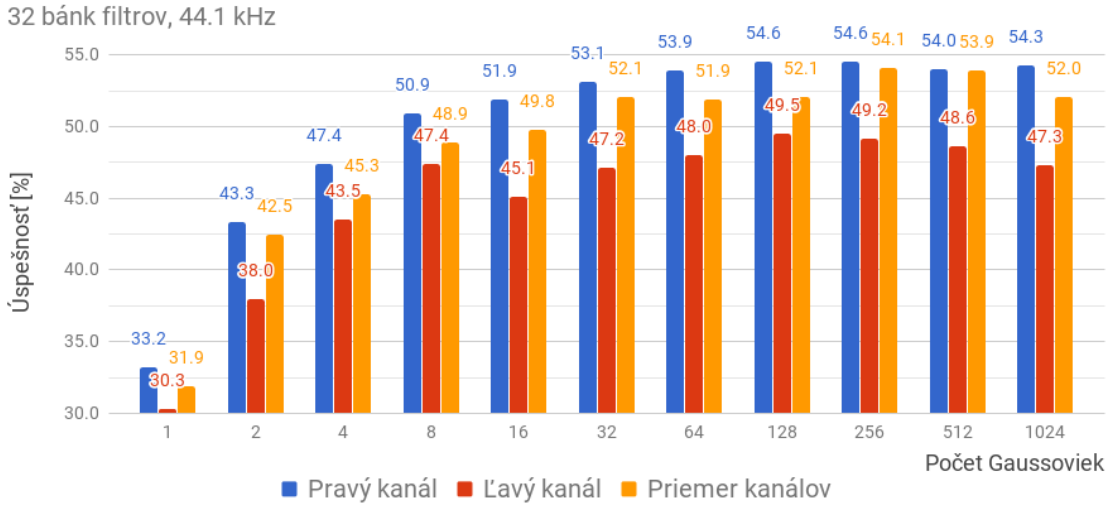
Výrazne najvyššiu úspešnosť dosahujú systémy, ktoré majú na vstupe nahrávky z pravého audio kanálu. Naopak suverénne najmenej úspešné sú systémy, ktorých vstup je tvorený nahrávkami z ľavého audio kanálu. Z grafov na obrázkoch 5.4a, 5.4b, 5.5a a 5.5b sa dá tiež vyčítať, že nahrávky získané spriemerovaním oboch audio kanálov – pravého aj ľavého,

sa úspešnosť systému zvýši v porovnaní so systémami s iba ľavým kanálom, no priemerná úspešnosť aj tak nie je vyššia, ako pri systémoch tvorených len pravým audio kanálom.

Dôvodom, prečo je taký rozdiel v úspešnostiach ľavého a pravého kanálu, by mohol byť fakt, že všetky audio dáta boli nahraté jedným zariadením, ktorého ľavý mikrofón nepracoval úplne korektne. V tom prípade by stálo za to, aby bola vykonaná hlbšia analýza organizátormi súťaže, pretože z našej strany sú to len domnienky. Na posluch a aj pohľad sú nahrávky z jednotlivých kanálov veľmi podobné.

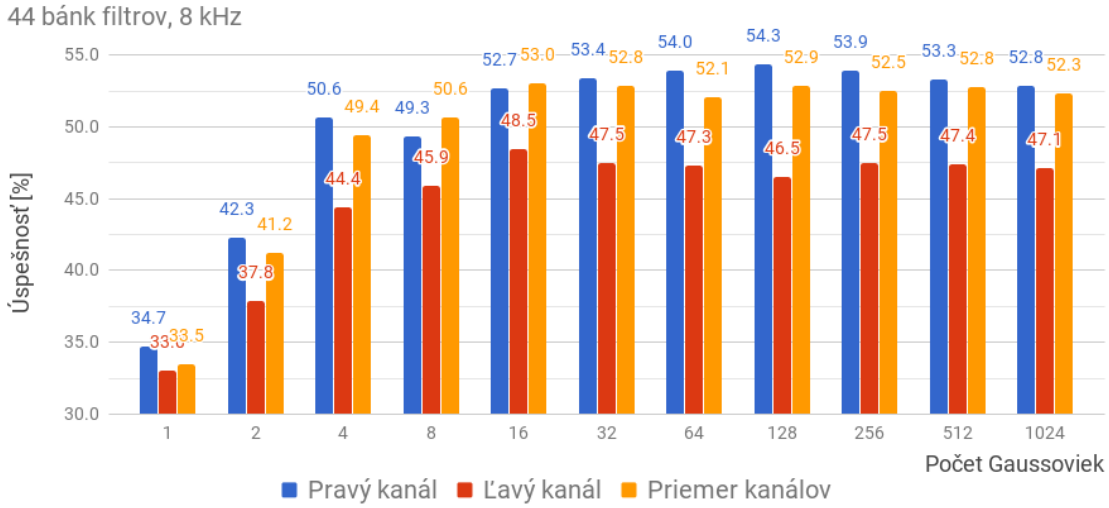


(a) Systémy s 20 bankami filtrov pri vzork. frekvencii 16 kHz.

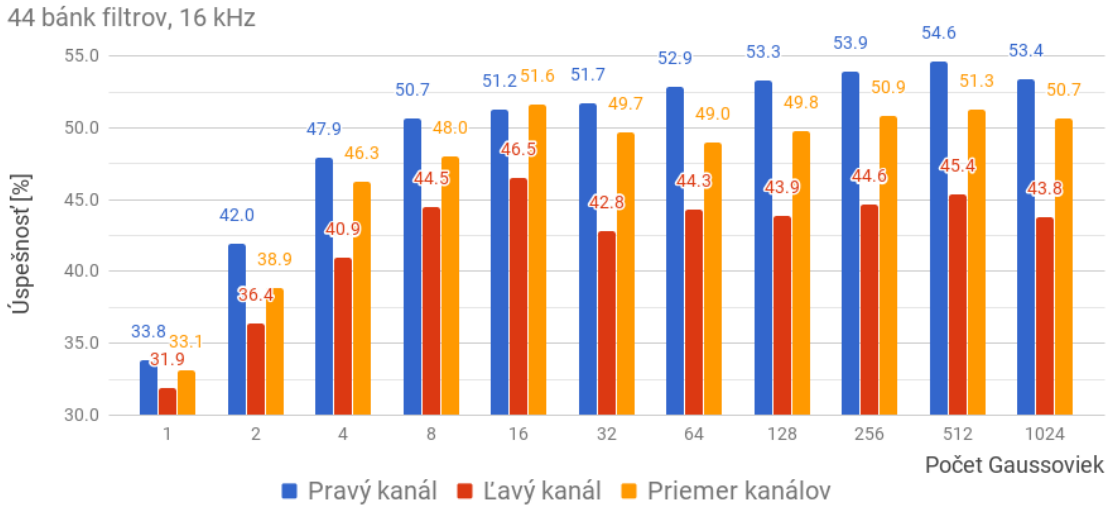


(b) Systémy s 32 bankami filtrov pri vzork. frekvencii 44.1 kHz.

Obr. 5.4: Grafy znázorňujúce vplyv daného audio kanálu na úspešnosť GMM systému. Vplyv audio kanálu je zachytený na dvoch rôznych konfiguráciach systémov líšiacich sa v počte bánk filtrov a vzork. frekvencii.



(a) Systémy so 44 bankami filtrov pri frekvencii 8 kHz.



(b) Systémy so 44 bankami filtrov pri frekvencii 16 kHz.

Obr. 5.5: Grafy znázorňujúce vplyv daného audio kanálu na úspešnosť GMM systému. Podobne ako na obrázku 5.4, aj tu sú zobrazené ďalšie dve konfigurácie systému, ktoré spolu tvoria 4 najúspešnejšie konfigurácie.

5.2.5 Zhrnutie experimentov na GMM systémoch s bankami filtrov

Celá predošlá sekcia sa venovala experimentom na GMM systéme, ktorý je založený na príznakoch nazývaných banky filtrov. Spočiatku sa jednalo o akúsi východiskovú konfiguráciu systému, prostredníctvom ktorej mohli začať experimenty nad daným systémom. Postupne boli obmieňané jednotlivé parametre systému, až sa napokon podarilo nájsť tú najvhodnejšiu kombináciu vedúcu k najvyššej možnej úspešnosti.

Na základe vykonaných experimentov teda môžme skonštatovať, že GMM systém založený na bankách filtrov s najvyššou úspešnosťou detekcie akustického prostredia z audio nahrávok má parametre, ktoré sú popísané v tabuľke 5.2. Táto tabuľka obsahuje aj dosiahnutú úspešnosť najlepšieho systému a tiež počet správne klasifikovaných nahrávok/súborov.

Tabuľka 5.2: Prehľad hodnôt jednotlivých parametrov najlepšieho GMM systému založeného na bankách filtrov a tiež dosiahnutá úspešnosť. Hodnoty sú získané na základe vykonaných experimentov.

Parameter	**Hodnota**
Počet bánk filtrov	44
Počet Gaussoviek	512
Normalizácia	žiadna
Vzorkovacia frekvencia	16 kHz
Audio kanál	pravý
Úspešnosť [%]	54.63
Úspešnosť [nahrávky]	885 z 1620

Experimenty takisto ukázali, že úspešnosť systému nie je priamo úmerná počtu bánk filtrov, najlepšie sa javí 44, 20, 40 a 32 bánk filtrov. Okrem toho sa ukázalo, že použitie normalizovaných príznakov vplýva jednoznačne negatívne na úspešnosť systému. V prípade vzorkovacej frekvencie experimenty nedokázali nájsť zreteľne úspešnejší variant, z grafov je vidno, že 8 kHz variant dosiahol v priemere najmenšiu úspešnosť. Zvyšné dva varianty (16 kHz a 44.1 kHz) dosahujú ale veľmi podobnú úspešnosť. V prípade audio kanálov sa najprospešnejšie vzhľadom na úspešnosť ukázal celkom jednoznačne pravý kanál, pričom naopak ľavý kanál dosahoval zjavne najnižšiu úspešnosť.

5.3 GMM systém založený na MFCC koeficientoch

V tejto sekcii bude analyzovaný GMM systém, ktorý je na rozdiel od predchádzajúceho GMM systému založený nie na bankách filtrov, ale na *Mel-frekvenčných cepstrálnych koeficientoch,* ktoré budem skrátene označovať ako *MFCC koeficienty.* Tieto príznaky som si vybral na základe viacerých článkov, ktoré som na túto tému čítal. Z týchto článkov som zistil, že autori často využívajú práve tieto príznaky, pretože majú relatívne dobrú úspešnosť v porovnaní s ostatnými príznakmi, ktoré sa všeobecne pri rozpoznávaní reči a v súvislosti s GMM systémom používajú. V mojom prípade by som rád zistil, do akej miery sa bude líšiť úspešnosť GMM systému s bankami filtrov v porovnaní so systémom s MFCC koeficientami a či aplikácia týchto MFCC koeficientov bude viesť k zlepšeniu celkovej úspešnosti GMM systému, alebo nie. K tomu budú slúžiť práve experimenty, ktoré postupne predstavím v tejto sekcii.

Tak ako aj v predošlej sekcii, aj v tomto prípade je GMM systém tvorený viacerými parametrami, ktoré môžu ovplyvniť úspešnosť celkového systému. Pri systéme s MFCC koeficientami je dokonca týchto parametrov ešte viac, aj keď niektoré sú rovnaké. Cieľom je teda opäť nájsť správnu kombináciu všetkých parametrov, ktoré dokážu systém vytvoriť tak, aby dosahoval čo najlepšie výsledky pri detekcii akustického prostredia. Parametre, resp. kategórie experimentov, ktorým sa budem venovať, sú nasledovné:

1. počet príznakov – MFCC koeficientov
2. normalizácia príznakov
3. nultý MFCC koeficient

4. odvodené koeficienty 1. rádu – delta koeficienty
5. odvodené koeficienty 2. rádu – doubledelta koeficienty
6. vzorkovacia frekvencia vstupných audio nahrávok
7. audio kanál vstupných audio nahrávok
8. počet Gaussoviek

Pre účely hlavne prvých experimentov je potrebné znovu tak, ako aj v predošlej časti stanoviť východiskové hodnoty pre jednotlivé parametre, aby bolo jasné, s akými parametrami bol ten-ktorý systém spúšťaný. Prehľad východiskových hodnôt pre jednotlivé parametre zobrazuje tabuľka 5.3.

Pri MFCC koeficientoch je treba tiež určiť počet bánk filtrov, ktoré budú použité pri tvorbe MFCC koeficientov. S týmto parametrom som neexperimentoval, počet bánk filtrov som určil fixne podľa toho, akú hodnotu používali autori pri MFCC koeficientoch najčastejšie vo svojich článkoch, ktoré som čítal. V experimentoch teda budú prezentované systémy s 24 bankami filtrov.

Tabuľka 5.3: Prehľad východiskových hodnôt jednotlivých parametrov pre GMM systém založený na MFCC koeficientoch.

Parameter	**Východisková hodnota**
Normalizácia	žiadna (NN)
Nultý MFCC koeficient	nie (NC0)
Delta koeficienty	nie (ND)
Doubledelta koeficienty	nie (ND)
Vzorkovacia frekvencia	8 kHz
Audio kanál	priemer pravého a ľavého kanálu
Počet Gaussoviek [2]	1 – 128

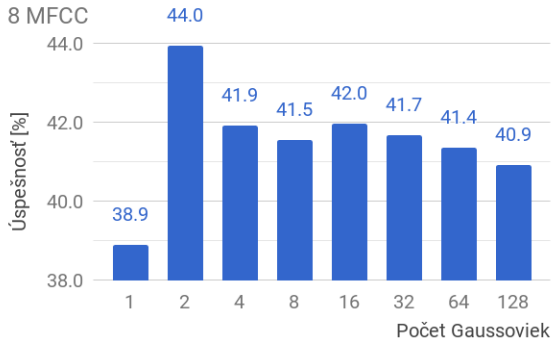
Pokiaľ nie je explicitne uvedené inak, sú všetky systémy spomenuté v ďalšej časti trénované na development dátovej sade zo súťaže DCASE 2017 a evaluované na evaluačnej dátovej sade súťaže DCASE 2017. Ako som sa už v predošlej časti vyjadril, radšej som experimentoval s dátovou sadou z roku 2017, pretože je ťažšia a výsledky sú vierohodnejšie.

5.3.1 Počet MFCC koeficientov

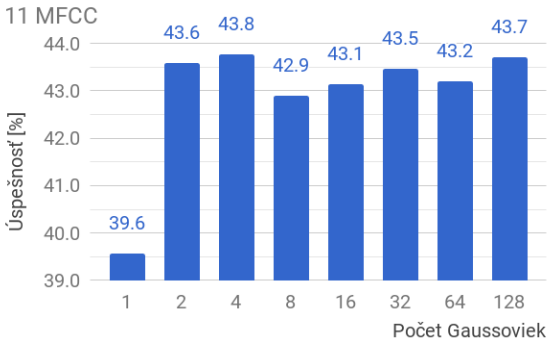
Prvá kategória experimentov – experimenty s počtami MFCC koeficientov – majú za cieľ jedinú vec – nájsť optimálny počet koeficientov, ktorý zabezpečí GMM systému čo možno najlepšiu možnú úspešnosť. Vzhľadom na to, že toto je počiatočná kategória experimentov s MFCC koeficientami, všetky parametre (mimo počtu MFCC koeficientov) uvádzaných systémov v tejto časti majú východiskovú hodnotu, t.j. hodnotu, ktorá je uvedená v tabuľke 5.3.

Obrázok 5.6 znázorňuje grafy úspešnosti jednotlivých systémov, ktoré sa líšia počtom koeficientov. Rozhodol som sa použiť nasledovné počty koeficientov, pričom bolo treba

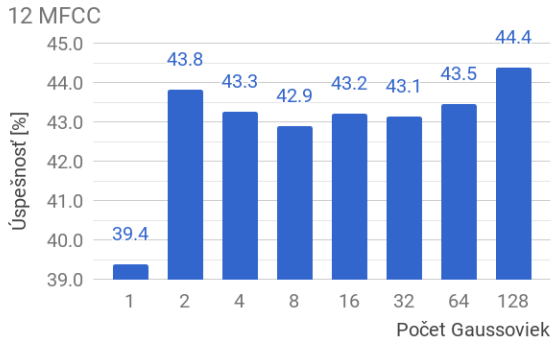
[2] Každý experiment bol spustený pre 1 až 128 Gaussoviek, aby bolo nejakým spôsobom možné sledovať vývoj úspešnosti systému. Toto neplatí pre záverečnú kategóriu experimentov, ktorá hľadá optimálny počet Gaussoviek.



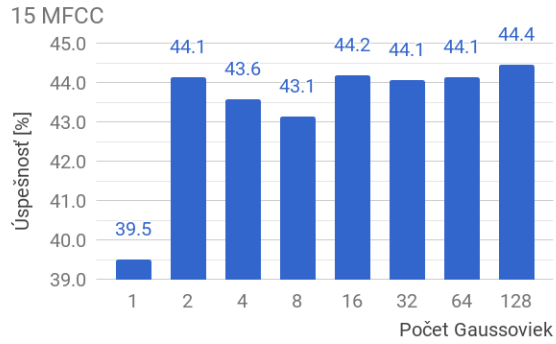
(a) 8 MFCC koeficientov



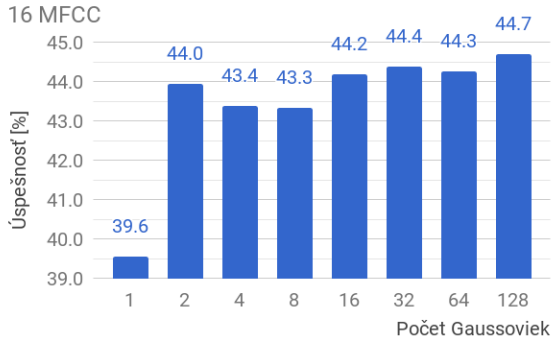
(b) 11 MFCC koeficientov



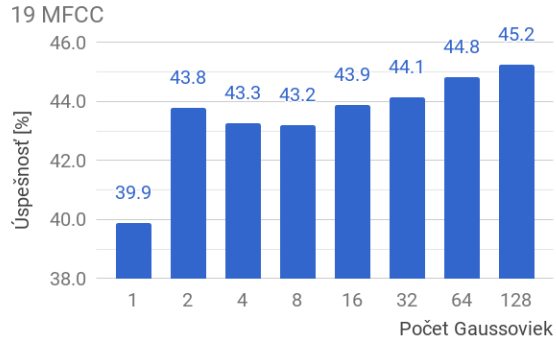
(c) 12 MFCC koeficientov



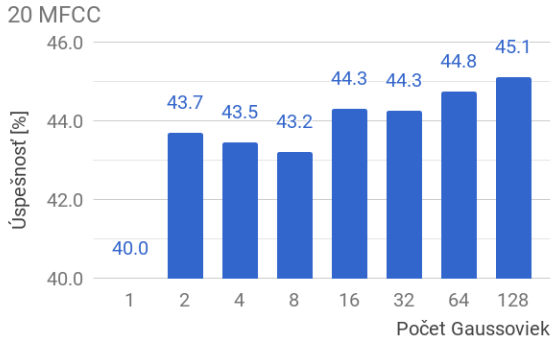
(d) 15 MFCC koeficientov



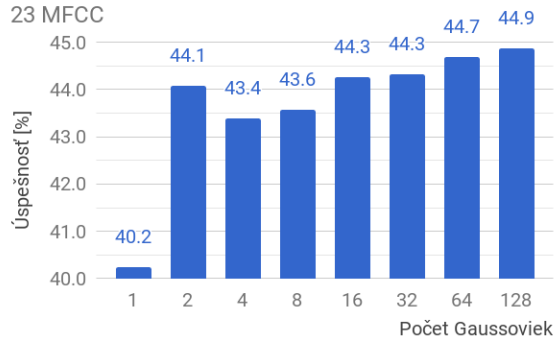
(e) 16 MFCC koeficientov



(f) 19 MFCC koeficientov



(g) 20 MFCC koeficientov



(h) 23 MFCC koeficientov

Obr. 5.6: Grafy znázorňujúce vplyv počtu príznakov na úspešnosť GMM systému. V tomto prípade sú príznaky reprezentované MFCC koeficientami.

zohľadniť skutočnosť, že v prípade 24 bánk filtrov je možné použiť maximálne 23 koeficientov (totiž, počet koeficientov nesmie byť väčší alebo rovný ako počet bánk filtrov, z ktorých sa koeficienty vytvárajú):

- 8 MFCC koeficientov (graf s úspešnosťou tohto systému ilustruje obrázok 5.6a)
- 11 MFCC koeficientov (obrázok 5.6b)
- 12 MFCC koeficientov (obrázok 5.6c)
- 15 MFCC koeficientov (obrázok 5.6d)
- 16 MFCC koeficientov (obrázok 5.6e)
- 19 MFCC koeficientov (obrázok 5.6f)
- 20 MFCC koeficientov (obrázok 5.6g)
- 23 MFCC koeficientov (obrázok 5.6h)

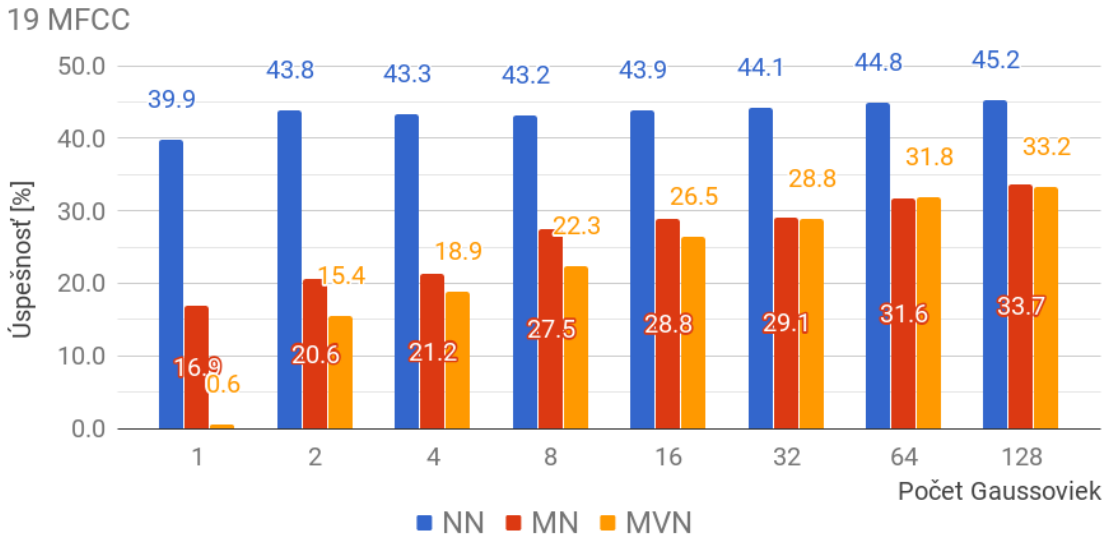
Najlepšiu úspešnosť dosiahol podľa grafov na obrázku 5.6 systém s 19 MFCC koeficientami, ktorý pri 128 Gaussovkách dosiahol 45.2% úspešnosti. Systém s 20 koeficientami sa svojou úspešnosťou 45.1% zaradil veľmi tesne za najlepší systém. Preto môžme aj pri ďalších experimentoch uvažovať, že práve systémy s 19 a 20 koeficientami sú tie, ktoré z hľadiska počtu koeficientov dosahujú najlepší výsledok.

Na základe priebehov jednotlivých grafov môžme tiež konštatovať, že úspešnosť GMM systému stúpa priamo úmerne s počtom MFCC koeficientov až do bodu, kedy systém dosiahne maximum. Potom úspešnosť zase klesá. V prípade 24 bánk filtrov je toto maximum v bode 19 koeficientov. V porovnaní s GMM systémom založeným na bankách filtrov experimenty ukázali, že úspešnosť nie je priamo úmerná počtu bánk filtrov, to znamená, že už prvé experimenty s MFCC koeficientami poukazujú na odlišné vzájomné chovanie týchto systémov. Dôvodom takéhoto chovania je pravdepodobne to, že banky filtrov sú na sebe vzájomne závislé, sú korelované, zatiaľ čo MFCC sú dekorelované banky filtrov, čo sa viac páči GMM systému.

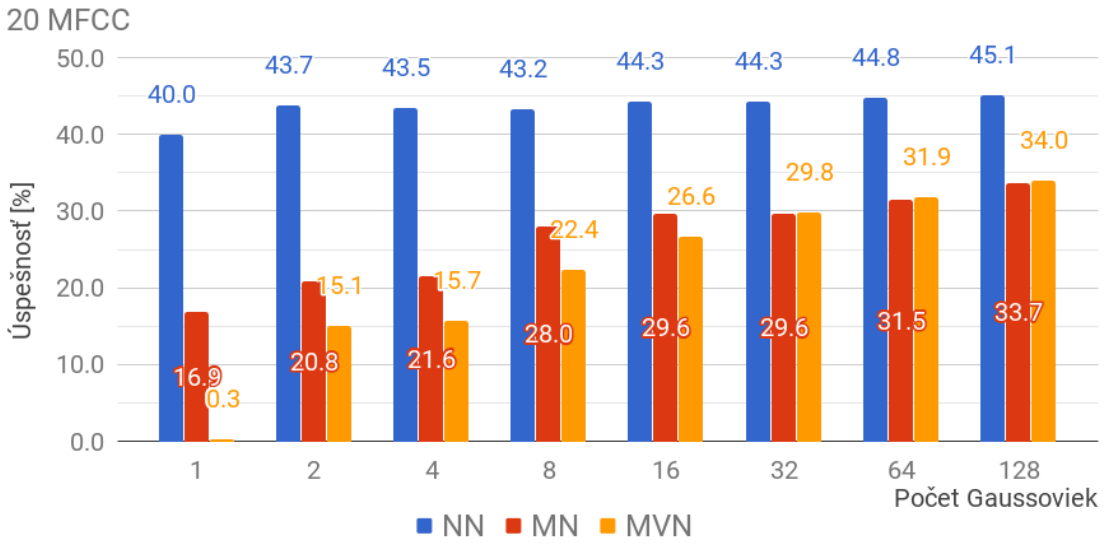
5.3.2 Normalizácia MFCC koeficientov

Tak ako v prípade GMM systému s bankami filtrov, tak aj v prípade systému s MFCC koeficientami sa ďalšia kategória experimentov zameriava na normalizáciu príznakov, čiže normalizáciu MFCC koeficientov. Jedná sa o normalizáciu pre každú nahrávku, tzv. *file based normalizáciu.* Experimenty tejto kategórie nadväzujú na výsledky z predošlej kategórie experimentov, a teda figurovať budú systémy s 19 a 20 MFCC koeficientami, keďže sa ukázalo, že tento počet je optimálny vzhľadom na úspešnosť systému. Varianty normalizácie sú rovnaké ako v prípade bánk filtrov:

- *žiadna normalizácia* – v grafoch označovaná ako NN (východisková hodnota)
- *mean normalizácia* – v grafoch označovaná ako MN
- *mean and variance normalizácia* – v grafoch označovaná ako MVN



(a) Systémy s 19 MFCC koeficientami.



(b) Systémy s 20 MFCC koeficientami.

Obr. 5.7: Grafy znázorňujúce vplyv normalizácie príznakov na úspešnosť GMM systémov líšiacich sa v počte MFCC koeficientov *(NN = žiadna normalizácia, MN = mean normalizácia, MVN = mean and variance normalizácia).*

Obrázok 5.7 ukazuje veľmi zreteľne vplyv normalizácie MFCC koeficientov na úspešnosť GMM systému, či už sa jedná o MN normalizáciu alebo o MVN normalizáciu. Najlepší variant z pohľadu úspešnosti systému je bez pochyby taký variant, kedy MFCC koeficienty nie sú vôbec normalizované (NN). Platí to tak pre systém s 19 MFCC koeficientami (obrázok 5.7a), ako aj pre systém s 20 koeficientami (obrázok 5.7b). MN normalizácia je v priemere úspešnejšia ako MVN normalizácia, no s pribúdajúcim počtom Gaussoviek sa zvyšuje úspešnosť práve MVN normalizácie a v prípade 64 či 128 Gaussoviek má už lepšiu úspešnosť ako MN normalizácia.

Záverom možno pozorovať, že vplyv normalizácie na úspešnosť v prípade systému s bankami filtrov (pozri obrázok 5.2) a systému s MFCC koeficientami je veľmi podobný. Z toho vyplýva, že pri budovaní systémov rozpoznávania reči za účelom detekcie akustického prostredia nie je dobrý nápad normalizovať príznaky GMM systému. V prípade bánk filtrov je aj rozdiel medzi MN a MVN zjavný, zatiaľ čo pri MFCC koeficientoch sa tento rozdiel skôr postupne stráca.

5.3.3 Nultý MFCC koeficient

Experiment s (ne)pridaním nultého MFCC koeficientu k ostatným koeficientom, ako aj názov napovedá, je špecifický pre GMM systém s MFCC koeficientami, preto nefiguroval v experimentoch s bankami filtrov. Nultý koeficient sa v odbornej literatúre označuje tiež ako *C0* a podľa viacerých autorov je aj veľmi užitočný. C0 je totiž podľa [25] súbor priemerných energií každého frekvenčného pásma v analyzovanom signále. Okrem štandardného nultého MFCC koeficientu existuje tiež variant, ktorý môže tento C0 nahradiť. Jedná sa o logaritmickú energiu *E,* ktorej hodnota sa použije namiesto C0 koeficientu. V tejto kategórii experimentov boli teda skúmané 3 varianty:

- *len priame MFCC koeficienty* – v grafoch označené ako NC0 (východisková hodnota)
- *pridanie nultého MFCC koeficientu* – v grafoch označené ako C0
- *logaritmická energia namiesto C0* – v grafoch označené ako E

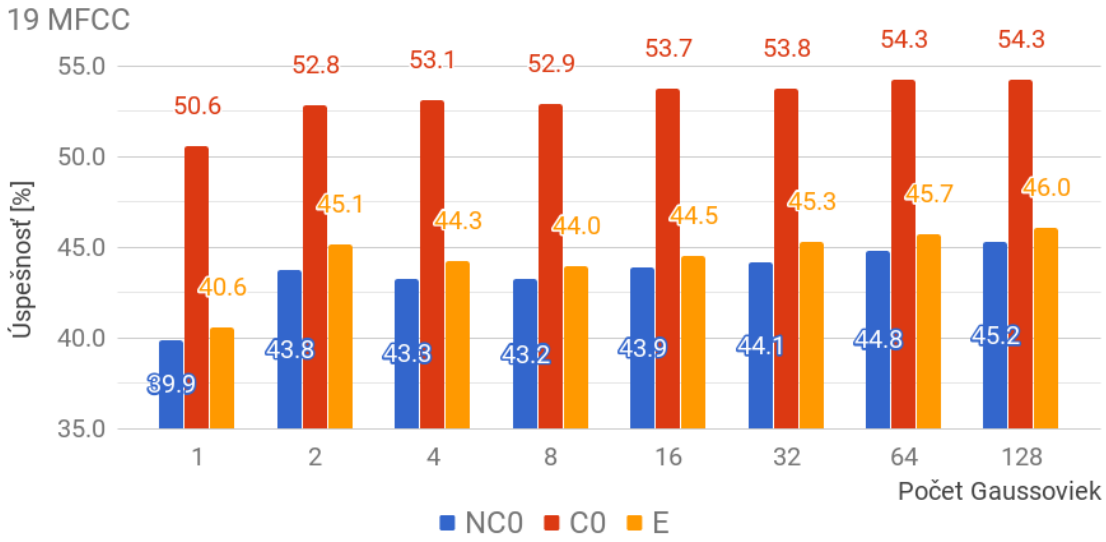
Grafy na obrázku 5.8 znázorňujú vplyv nultého koeficientu na úspešnosť systému. Z týchto grafov je zreteľne vidno, že pridanie nultého MFCC koeficientu k ostatným koeficientom vplýva až prekvapivo pozitívne na celkovú úspešnosť systému. Potvrdil to tak systém s 19 MFCC koeficientami (na obrázku 5.8a), ako aj systém s 20 MFCC koeficientami (na obrázku 5.8b). V prípade logaritmickej energie E ako nultého koeficientu sa úspešnosť systému zvýšila síce minimálne, no stále sa jedná o pozitívny vplyv na úspešnosť.

Nakoniec možno teda s určitosťou povedať, že obe varianty nultého koeficientu (tak C0 ako aj E) reagujú pozitívne na úspešnosť systému, pretože v oboch prípadoch je celková úspešnosť systému vyššia ako v prípade variantu bez zahrnutia nultého koeficientu k ostatným koeficientom. Preto poznatok vyplývajúci z tohto experimentu zahrniem v nasledujúcich experimentoch, v ktorých budú figurovať systémy s MFCC koeficientami rozšírenými o nultý koeficient C0.

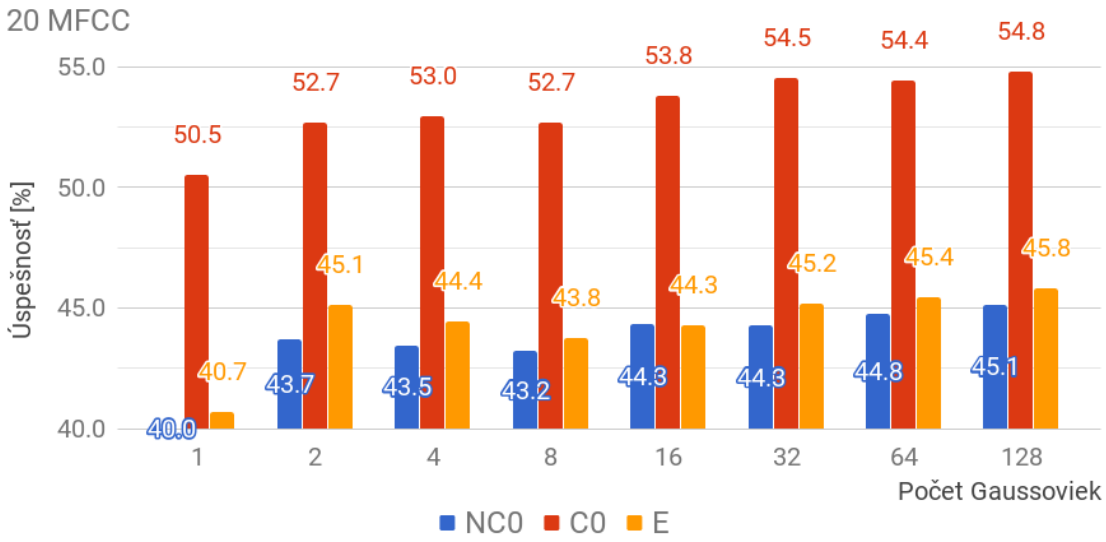
5.3.4 Delta a doubledelta koeficienty

Ďalšia kategória experimentov sa venuje *odvodeným koeficientom 1.rádu,* ktoré sú známe tiež pod pojmom *delta koeficienty,* alebo *diferenčné koeficienty* a *odvodeným koeficientom 2.rádu,* ktoré sa označujú aj ako *doubledelta koeficienty,* alebo *akceleračné koeficienty.* Experimenty hľadajú odpoveď na otázku, či je z hľadiska úspešnosti systému výhodnejšie pripojiť k priamym (statickým) MFCC koeficientom aj odvodené koeficienty a ak áno, či je výhodnejšie pripojiť iba delta koeficienty, alebo je lepšie pripojiť rovno aj doubledelta koeficienty.

Delta koeficienty predstavujú odchýlky medzi jednotlivými priamymi MFCC koeficientami, zatiaľ čo doubledelta koeficienty predstavujú odchýlky medzi jednotlivými delta koeficientami. Z predošlého popisu je teda zrejmé, že v tejto kategórii experimentov budú vystupovať opäť 3 alternatívy:



(a) Systémy s 19 MFCC koeficientami.

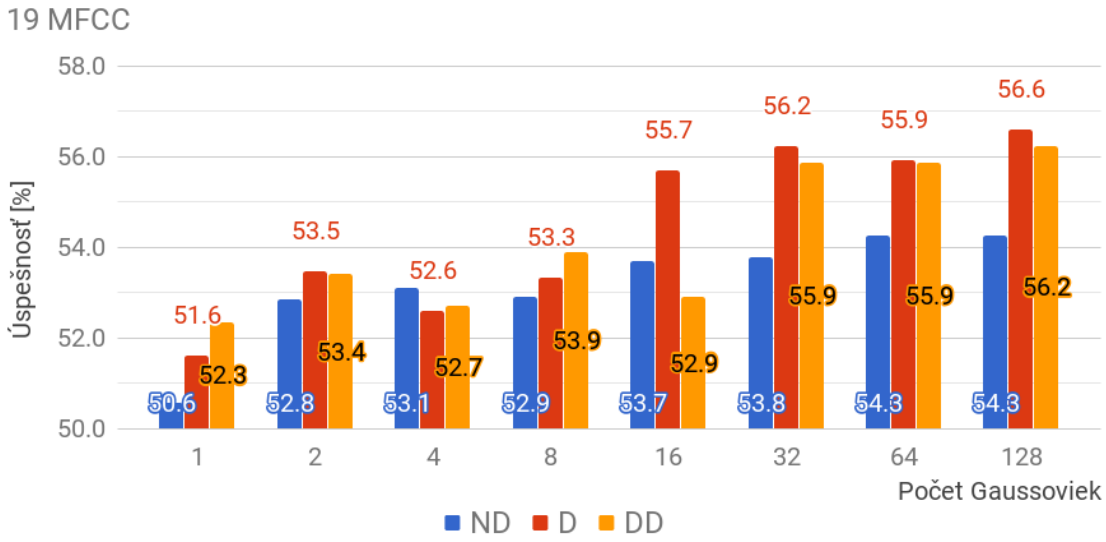


(b) Systémy s 20 MFCC koeficientami.

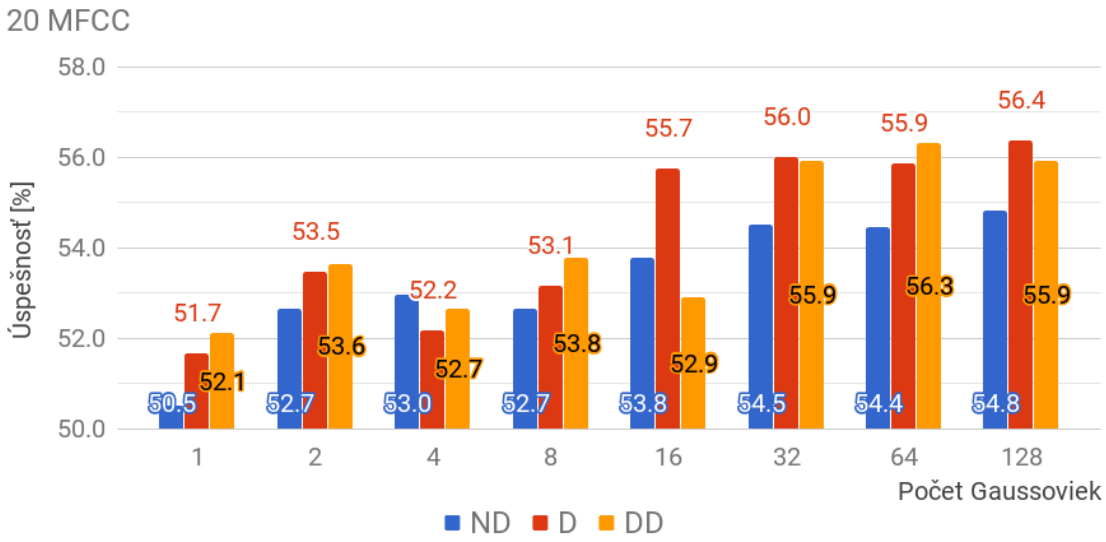
Obr. 5.8: Grafy znázorňujúce vplyv nultého MFCC koeficientu na úspešnosť GMM systémov líšiacich sa v počte MFCC koeficientov *(NC0 = len priame MFCC koeficienty, C0 = pridanie nultého MFCC koeficientu, E = logaritmická energia použitá ako hodnota nultého koeficientu).*

- *len priame koeficienty* – v grafoch označené ako ND (východisková hodnota)
- *pridanie delta koeficientov* – v grafoch označené ako D
- *pridanie delta aj doubledelta koeficientov* – v grafoch označené ako DD

Priebeh experimentov tradične vyobrazujú grafy na obrázku 5.9. Na prvý pohľad môžme vidieť, že pripojenie odvodených koeficientov k priamym MFCC koeficientom má rozhodne pozitívny vplyv na úspešnosť celého GMM systému. Oproti variantu bez odvodených koefi-



(a) Systémy s 19 MFCC koeficientami.



(b) Systémy s 20 MFCC koeficientami.

Obr. 5.9: Grafy znázorňujúce vplyv odvodených koeficientov na úspešnosť GMM systémov líšiacich sa v počte MFCC koeficientov *(ND = len priame koeficienty, D = pridanie delta koeficientov, DD = pridanie delta a doubledelta koeficientov).*

cientov sa úspešnosť zdvihla o niekoľko jednotiek percentuálnych bodov v závislosti na konkrétnom počte Gaussoviek.

Na druhú stranu, priebehy grafov tiež ukazujú, že vplyv doubledelta koeficientov na úspešnosť systému nie je tak výrazný ako v prípade delta koeficientov. Častokrát je dokonca úspešnosť systému s doubledelta koeficientami nižšia ako bez týchto koeficientov. Najvyššiu úspešnosť 56.6% v prípade systému s 19 MFCC koeficientami (na obrázku 5.9a) dosiahol systém s delta koeficientami, ale bez doubledelta koeficientov. Systém s 20 MFCC koeficientami (na obrázku 5.9b) dosiahol najvyššiu úspešnosť 56.4% taktiež s delta koeficientami a bez doubledelta koeficientov.

Výsledky experimentov teda naznačujú, že pripojenie delta koeficientov k priamym MFCC koeficientom je dobrý krok k zlepšeniu úspešnosti systému, zatiaľ čo pripojenie doubledelta koeficientov nevedie k významnejšiemu zvýšeniu úspešnosti GMM systému. Preto systémy, ktoré budú figurovať v ďalších experimentoch, budú obsahovať delta koeficienty, no bez účasti doubledelta koeficientov.

5.3.5 Vzorkovacia frekvencia vstupných audio nahrávok

Toto je ďalšia z kategórií experimentov, ktorá bola použitá i v prípade systému s bankami filtrov a preto ju asi ani netreba veľmi popisovať. Experimenty majú za cieľ zistiť, ktorá vzorkovacia frekvencia vstupných audio nahrávok je z pohľadu úspešnosti najúčinnejšia. Do úvahy sa berú rovnaké 3 možnosti ako v prípade bánk filtrov, čiže:

- *8 kHz* – východisková hodnota pre predošlé experimenty
- *16 kHz*
- *44.1 kHz*

Grafy na obrázku 5.10 zobrazujú, do akej miery sa ktorá vzorkovacia frekvencia podieľa na úspešnosti systému. Je celkom zjavné, že variant, ktorý sa na zvýšení úspešnosti podieľa najviac, je v tomto prípade frekvencia 16 kHz. Platí to tak pre systém s 19 MFCC koeficientami, ktorý je na obrázku 5.10a, ako aj pre systém s 20 MFCC koeficientami z obrázka 5.10b. Úspešnosť sa vyšplhala až na úroveň 58.8%, resp. 59.1% v prípade systému s 20 koeficientami.

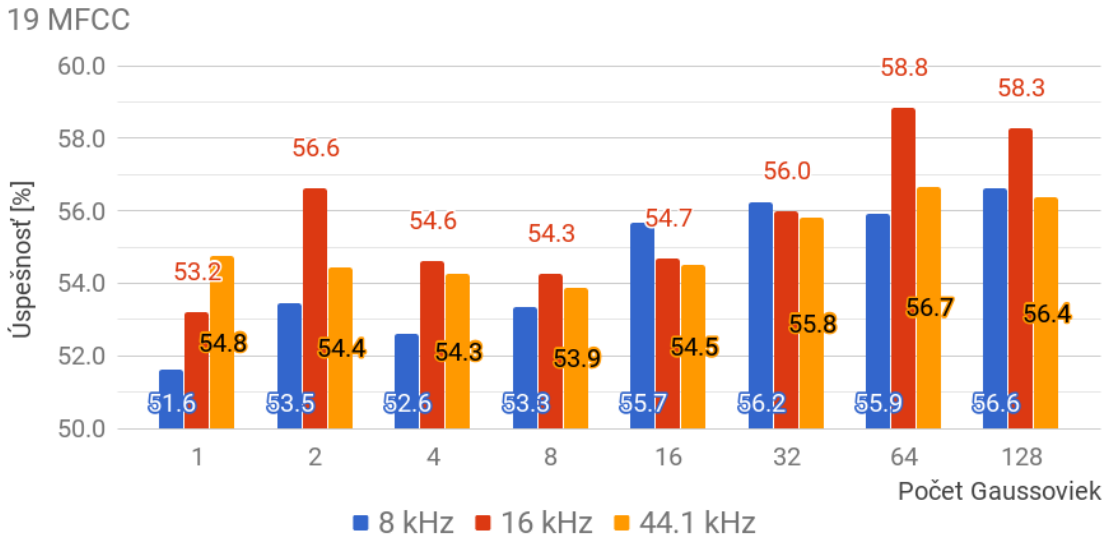
Čo sa týka frekvencií 8 kHz a 44.1 kHz, ich úspešnosť je vcelku popreplетaná. V určitých prípadoch je s malým rozdielom lepších 8 kHz, v iných prípadoch zase prevažuje 44.1 kHz. Len subjektívne však možno určiť, ktorá z týchto dvoch variant je symbolicky lepšia a ktorá horšia.

5.3.6 Audio kanál vstupných audio nahrávok

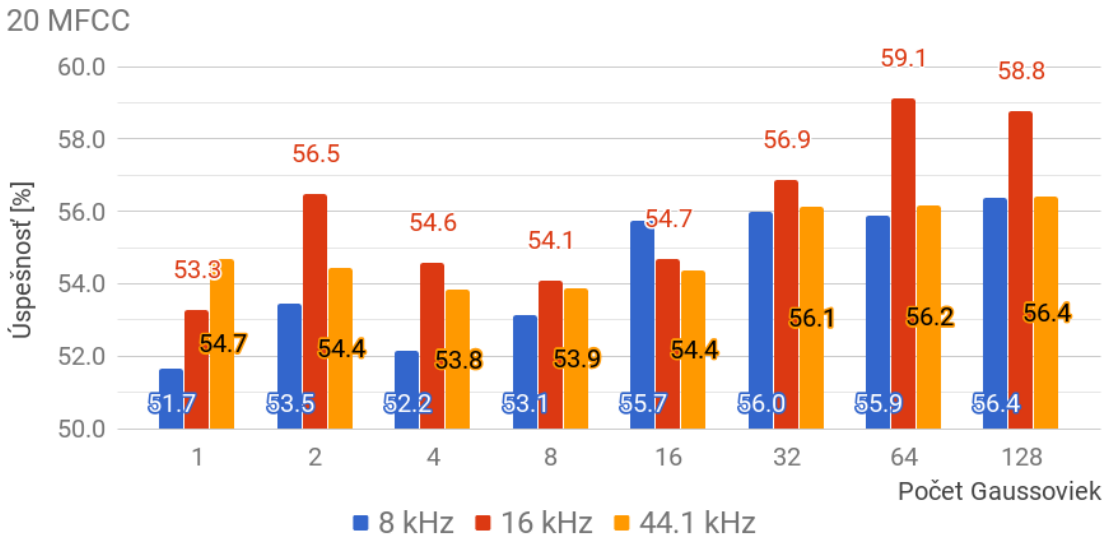
Ako bolo už spomenuté v kapitole 4, audio nahrávky poskytnuté autormi súťaže DCASE boli nahraté ako stereo nahrávky toho istého zvuku, t.j. obidva audio kanály sú si navzájom veľmi podobné, avšak nie sú úplne rovnaké. Z tohoto dôvodu som sa do svojej práce rozhodol zaradiť aj takúto kategóriu experimentov, aby bolo možné sledovať, či sú oba kanály len nepatrne odlišné, alebo či je rozdiel medzi nimi dokonca zreteľnejší. Pri experimentovaní beriem do úvahy tieto 3 možnosti:

- *priemer pravého a ľavého kanálu* – východisková hodnota predošlých experimentov
- *ľavý audio kanál*
- *pravý audio kanál*

Grafy na obrázku 5.11 ilustrujú vplyv osobitných audio kanálov na úspešnosť systému. S určitosťou môžme konštatovať, že pravý audio kanál dosahuje jednoznačne najvyššiu úspešnosť spomedzi ostatných variant. Grafy tiež napovedajú, že ľavý audio kanál nie je tou správnou voľbou, keď je treba zvýšiť úspešnosť, pretože úspešnosť takéhoto systému je príliš nízka oproti ostatným alternatívam. Tretia možnosť – priemer ľavého a pravého audio kanálu – kompenzuje úspešnosť medzi úspešnosťou systému nad ľavým kanálom a pravým kanálom.



(a) Systémy s 19 MFCC koeficientami.



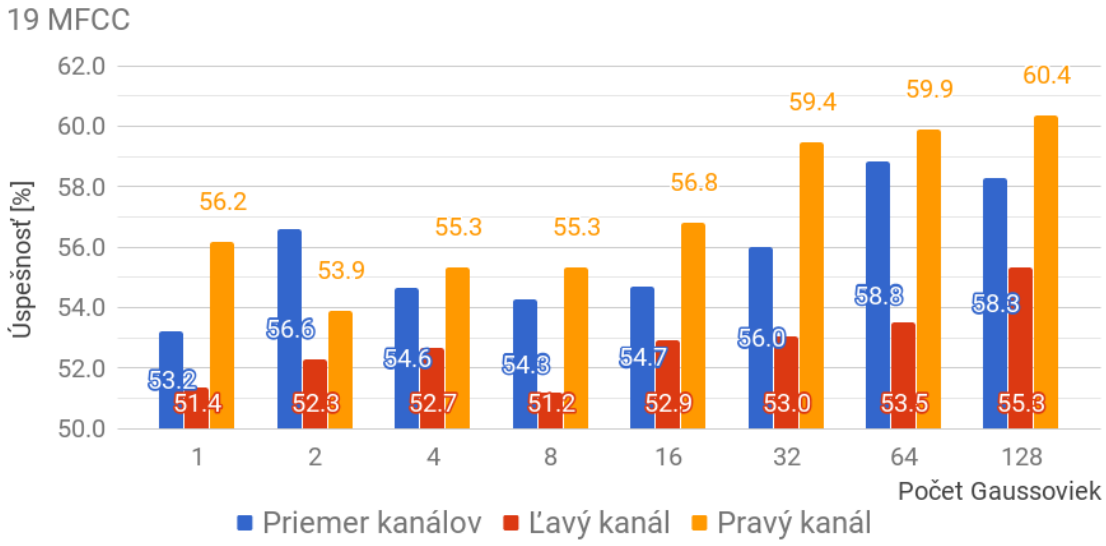
(b) Systémy s 20 MFCC koeficientami.

Obr. 5.10: Grafy znázorňujúce vplyv vzorkovacej frekvencie vstupných audio nahrávok na úspešnosť GMM systémov líšiacich sa v počte MFCC koeficientov.

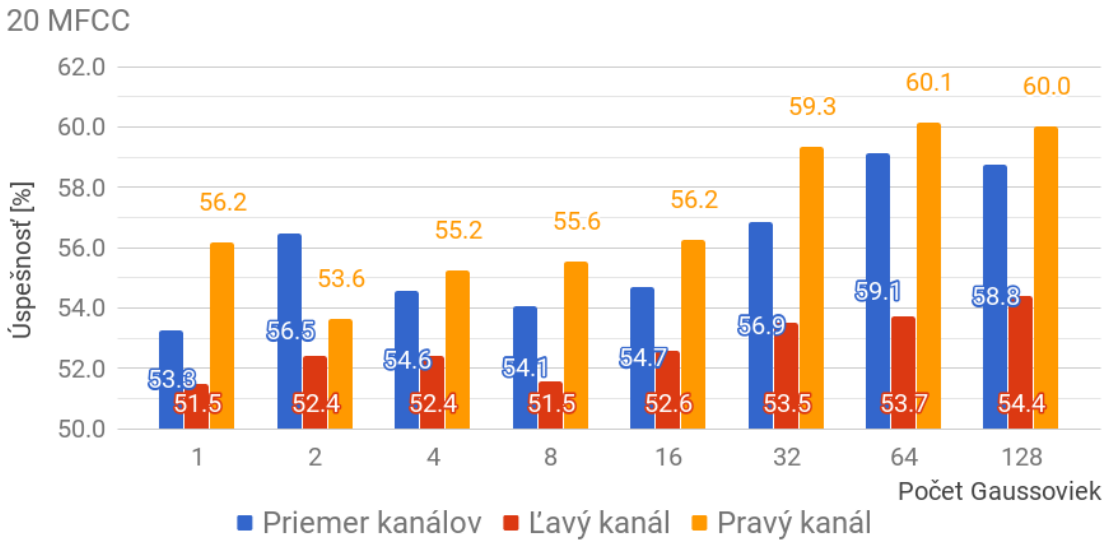
Otázkou však zostáva, prečo je medzi ľavým a pravým kanálom taký rozdiel v úspešnosti. O tom tu môžme ale jedine špekulovať. Každopádne by určite stálo za to, keby organizátori súťaže (autori dát) hlbšie zanalyzovali celý proces nahrávania týchto nahrávok. Je možné, že celý dataset bol nahratý jedným zariadením, ktorého ľavý mikrofón nefungoval úplne korektne. Na základe posluchu nahrávok z ľavého a pravého kanálu som nepostrehol rozdiel.

5.3.7 Počet Gaussoviek

Záverečnou experimentálnou kategóriou je stanovenie optimálneho počtu Gaussoviek tak, aby úspešnosť daného systému bola čo najvyššia. Vzhľadom na to, že vďaka vykonaným



(a) Systémy s 19 MFCC koeficientami.



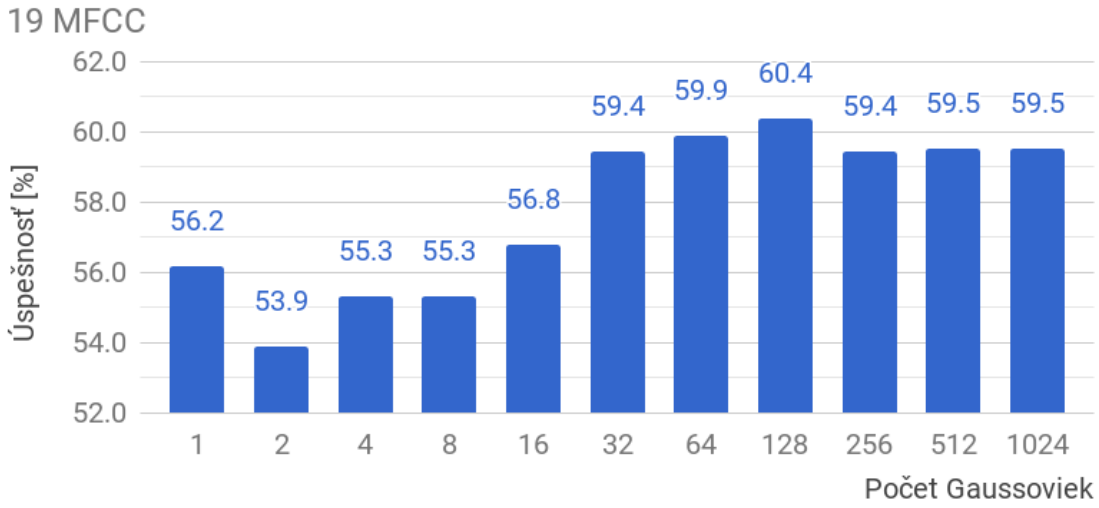
(b) Systémy s 20 MFCC koeficientami.

Obr. 5.11: Grafy znázorňujúce vplyv audio kanálu vstupných audio nahrávok na úspešnosť GMM systémov líšiacich sa v počte MFCC koeficientov.

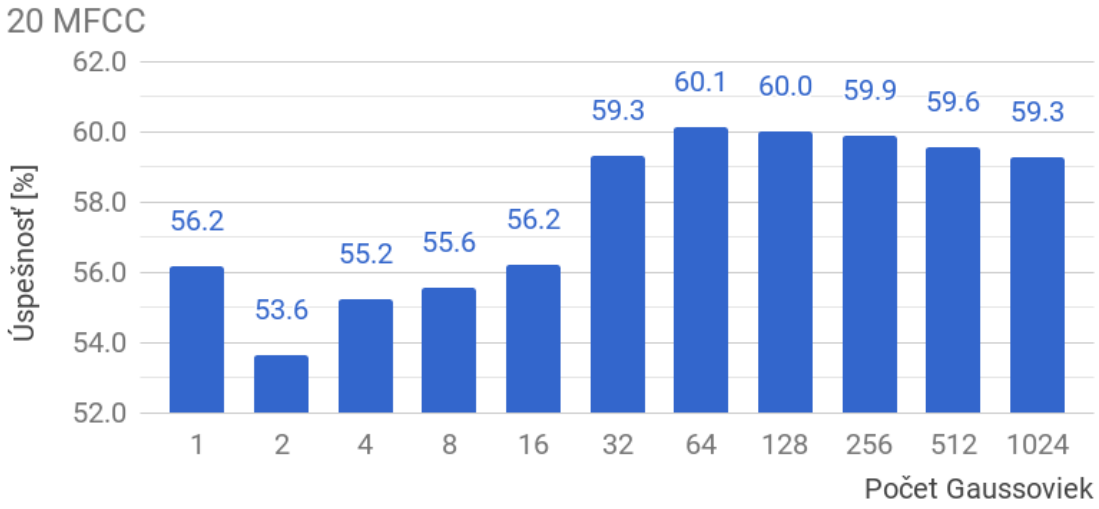
experimentom už poznáme 2 najúspešnejšie systémy, bude na zrealizovanie tohto experimentu potrebné len prekonfigurovať tieto systémy tak, aby sa spustili pre viac počtov Gaussoviek ako doteraz, t.j. napríklad pre 1 až 1024 Gaussoviek. V takom prípade bude k dispozícii širší náhľad na daný systém i na priebeh jeho úspešnosti a bude tak jednoduchšie stanoviť spomínaný optimálny počet Gaussoviek.

Grafy na obrázku 5.12 zobrazujú 2 najlepšie GMM systémy založené na MFCC koeficientoch a ich úspešnosti pri rôznych počtoch Gaussoviek. Oba grafy, tak pre systém s 19 koeficientami na obrázku 5.12a, ako aj pre systém s 20 koeficientami na obrázku 5.12b, majú veľmi podobný priebeh, vďaka čomu môžme určiť relatívne jasne, koľko Gaussoviek predstavuje ten optimálny počet. Najvyššia úspešnosť v prvom grafe sa nachádza celkom

jednoznačne pri 128 Gaussovkách, konkrétne 60.4%. V druhom grafe je to trochu menej jednoznačné, pretože najvyššia úspešnosť s minimálnymi odchýlkami sa nachádza pri 64, 128 a 256 Gaussovkách, konkrétne 60.1%, 60.0% a 59.9%. V priemere ale môžme prehlásiť aj na základe týchto údajov, že optimálny počet Gaussoviek je 128.



(a) Najlepší GMM systém s 19 MFCC koeficientami.



(b) Najlepší GMM systém s 20 MFCC koeficientami.

Obr. 5.12: Grafy znázorňujúce 2 najlepšie GMM systémy založené na MFCC koeficientoch a ich úspešnosť.

5.3.8 Systém s doubledelta koeficientami

Počas experimentov som tiež zistil, že systém s rovnakým nastavením parametrov ako ten najlepší s 20 MFCC koeficientami (na obrázku 5.12b) dosahuje tiež relatívne vysokú úspešnosť pri 128 Gaussovkách – 59.26%, keď sa k nemu pridajú aj doubledelta koeficienty. Tento systém tu spomínam hlavne z toho dôvodu, že bude figurovať pri fúzii systémov v kapitole 7.1.

5.3.9 Zhrnutie experimentov na GMM systémoch s MFCC koeficientami

Celá táto sekcia bola venovaná experimentom nad GMM systémom, ktorého príznaky reprezentujú MFCC koeficienty. Inicializačným bodom pre počiatočný experiment bola východisková konfigurácia systému, ktorá obsahovala východiskové hodnoty všetkých skúmaných parametrov. Tieto parametre boli následne s pribúdajúcimi experimentami obmieňané na základe výsledkov experimentov. Takýmto spôsobom sa experimenty postupne dostali na koniec a prostredníctvom nich bola získaná finálna konfigurácia najlepšieho systému, ktorá zahŕňa finálne hodnoty jednotlivých parametrov reflektujúce výsledky experimentov. Zoznam hodnôt osobitných parametrov najúspešnejšieho získaného GMM systému s MFCC koeficientami prehľadne odzrkadľuje tabuľka 5.4. Okrem toho sa v tabuľke nachádza aj úspešnosť tohto systému včetne počtu správne klasifikovaných nahrávok.

Experimenty okrem iného ukázali, že úspešnosť systému je priamo úmerná počtu MFCC koeficientov až do bodu, kedy nastane maximum, potom úspešnosť klesá. Normalizácia príznakov nie je správnym krokom k zvýšeniu úspešnosti systému, experimenty ukázali, že má negatívny vplyv na úspešnosť systému. Veľmi výrazne sa o zvýšenie úspešnosti pričinil nultý MFCC koeficient C0. Delta koeficienty potvrdili svoju užitočnosť, zatiaľ čo doubledelta koeficienty podľa experimentov netreba používať v záujme zachovania dostatočne vysokej úspešnosti systému. Vzorkovacia frekvencia sa najlepšie javí v prípade 16 kHz, pravý audio kanál vstupných nahrávok preukázal svoju dominanciu z hľadiska úspešnosti, čo sa o ľavom nedá povedať, pretože ten naopak úspešnosť znižoval. Záverečný experiment naznačil, že optimálny počet Gaussoviek pre takýto GMM systém je 128.

Tabuľka 5.4: Prehľad hodnôt jednotlivých parametrov najlepšieho GMM systému založeného na MFCC koeficientoch a tiež dosiahnutá úspešnosť. Hodnoty sú získané na základe vykonaných experimentov.

Parameter	**Hodnota**
Počet bánk filtrov[3]	24
Počet MFCC koeficientov	19
Normalizácia	žiadna
Nultý MFCC koeficient	áno (C0)
Delta koeficienty	áno (D)
Doubledelta koeficienty	nie (ND)
Vzorkovacia frekvencia	16 kHz
Audio kanál	pravý
Počet Gaussoviek	128
Úspešnosť [%]	60.37
Úspešnosť [nahrávky]	978 z 1620

Tabuľka 5.5 zobrazuje confusion maticu najlepšieho GMM systému, ktorého úspešnosť dosiahla 60.37%. Táto matica popisuje chovanie celého systému, ako boli rozpoznané jednotlivé nahrávky, ktoré triedy si systém navzájom najviac mýlil apod. Čísla na hlavnej diagonále predstavujú počet správne rozpoznaných nahrávok, resp. prostredí, z ktorých nahrávky pochádzajú. Analogicky, čísla mimo hlavnej diagonály predstavujú počet nesprávne rozpoznaných nahrávok danej triedy.

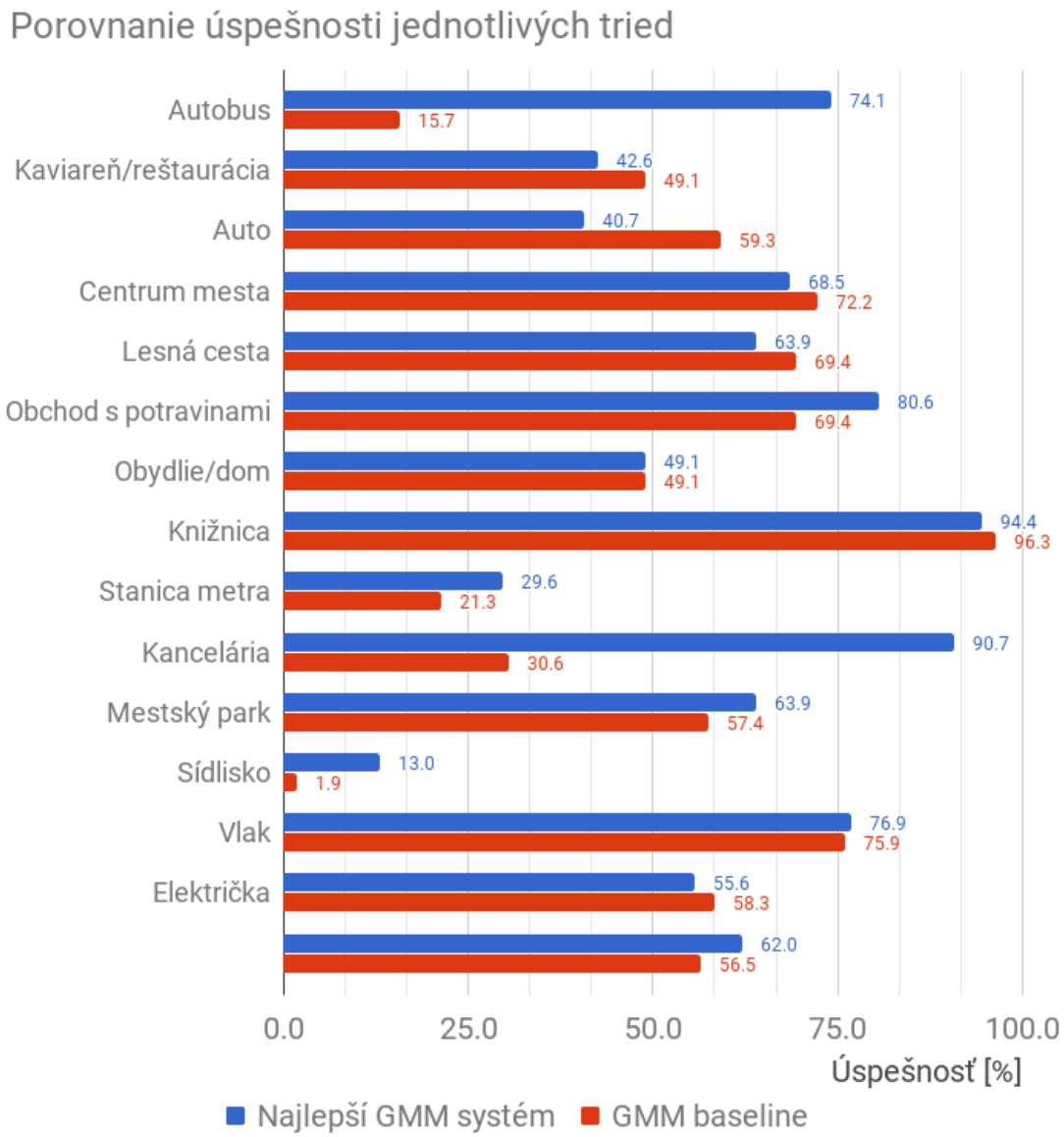
[3]Tento parameter nebol predmetom experimentovania, je uvedený pre úplnosť.

Tabuľka 5.5: Confusion matica najlepšieho GMM systému, ktorý dosiahol úspešnosť 60.37% na evaluačných dátach 2017. Riadky predstavujú skutočné návestia a stĺpce reprezentujú návestia predpovedané systémom. Čísla predstavujú počet nahrávok, na jednu triedu pripadá 108 evaluačných nahrávok.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1-Pláž	**80**	0	0	0	0	0	0	3	0	1	0	7	9	2	6
2-Autobus	0	**46**	1	**16**	0	0	**14**	2	2	**15**	0	0	0	1	**11**
3-Kaviareň	0	0	**44**	0	3	0	**10**	**17**	0	**32**	0	2	0	0	0
4-Auto	0	0	0	**74**	0	0	0	0	0	0	0	0	0	0	**34**
5-Centrum mesta	**13**	0	0	0	**69**	6	1	0	0	7	0	0	**11**	1	0
6-Lesná cesta	1	0	0	0	0	**87**	0	3	1	0	1	9	6	0	0
7-Potraviny	2	0	0	0	0	0	**53**	8	0	**45**	0	0	0	0	0
8-Obydlie/Dom	0	0	1	0	0	0	0	**102**	0	3	2	0	0	0	0
9-Knižnica	0	0	0	0	0	4	0	**48**	**32**	6	**18**	0	0	0	0
10-Stanica metra	0	0	4	0	0	2	0	0	0	**98**	4	0	0	0	0
11-Kancelária	0	0	0	0	0	0	0	**39**	0	0	**69**	0	0	0	0
12-Park	0	0	4	0	0	6	**10**	**14**	0	9	0	**14**	**51**	0	0
13-Sídlisko	0	0	0	0	**13**	7	0	1	0	2	1	1	**83**	0	0
14-Vlak	0	3	0	0	4	0	0	0	3	1	0	0	3	**60**	**34**
15-Električka	0	2	**13**	0	0	2	**13**	0	0	**11**	0	0	0	0	**67**

Z tejto matice sa dá napríklad vyčítať, že systém mal najmenšie problémy s rozpoznávaním nahrávok z *obydlia/domu,* pretože v 102 prípadoch (zo 108) rozpoznal toto akustické prostredie správne. Veľmi dobre si systém poradil aj s nahrávkami zo *stanice metra,* ktoré správne rozpoznal v 98 prípadoch. Naopak, veľké problémy robili systému nahrávky z *parku,* pretože len 14 nahrávok rozpoznal korektne. Až 51 z týchto nahrávok určil, že pochádzajú zo *sídliska,* 14 krát určil, že sa jedná o *obydlie/dom* a v 10 prípadoch klasifikoval tieto nahrávky ako nahrávky z *obchodu s potravinami.* Aj nahrávky z *knižnice* si systém často plietol. Iba 32 z nich klasifikoval správne a napríklad pri 48 nahrávkach určil, že sú nahraté v *obydlí/dome.* Ďalších 18 nahrávok zaradil, že patria do *kancelárie.* Systém si plietol vo veľkej miere taktiež nahrávky z *obchodu s potravinami,* kedy správne rozpoznal 53 nahrávok, no až pri 45 nahrávkach určil nesprávne, že sa jedná o nahrávky zo *stanice metra.*

Na záver kapitoly o GMM systémoch ešte uvádzam graf na obrázku 5.13, ktorý porovnáva úspešnosť detekcie jednotlivých tried môjho najlepšieho GMM systému a GMM baseline systému na dátach súťaže DCASE 2017. Pri viacerých triedach sú na tom oba systémy veľmi podobne, no napríklad pri triede *autobus* sa líšia značným spôsobom. Veľký rozdiel zaznamenala aj *kancelária.*



Obr. 5.13: Porovnanie úspešnosti detekcie jednotlivých tried akustických prostredí z pohľadu môjho najlepšieho GMM systému s úspešnosťou 60.37% a GMM baseline systému, ktorého priemerná úspešnosť je 52.16%. Systémy boli evaluované na evaluačných dátach súťaže z roku 2017.

Kapitola 6

Experimenty a výsledky s i-vektor extraktorom

V tejto kapitole bude detailnejšie popísaný formát experimentov, ktoré som vykonal so systémom založeným na i-vektoroch a hlavne tu budú popísané a diskutované samotné experimenty, ktorým som sa v tejto časti venoval. Kapitola obsahuje takisto výsledky jednotlivých i-vektor systémov, ktoré sa podieľali na experimentoch.

6.1 Formát i-vektor experimentov

Rovnako ako som spomenul v sekcii 5.1, aj v prípade i-vektor experimentov som sa snažil vytvoriť si jednotný systém práce, aby som čelil čím menej problémom súvisiacich s organizáciou experimentov a všetkých súborov, či zložiek potrebných k tomu. Možnosť opätovného spustenia každého jedného experimentu s istotou rovnakého výsledku je v tomto prípade taktiež veľmi dôležitá, preto som sa rozhodol vytvoriť skript, ktorý by tieto veci dokázal zabezpečiť. Tento skript, ktorý má na starosti aj trénovanie aj evaluáciu i-vektor systému, popisujem detailnejšie v nasledujúcej časti. Na rozdiel od GMM experimentov, je v tomto prípade pre každý experiment určený práve jeden skript.

6.1.1 Experimentový skript

V tejto časti je popísané chovanie shell skriptu, ktorý sa zaoberá trénovaním, ale aj evaluáciou i-vektor systému – pre zjednodušenie ho budem označovať ako *experimentový skript*. Rovnako, ako aj v prípade trénovacích a evaluačných skriptov pre GMM systém, je rozhodujúci práve názov tohto skriptu, ktorý vypovedá o tom, aký systém sa prostredníctvom daného skriptu vybuduje. Názov takéhoto experimentového skriptu pozostáva z 12 častí, ktoré sú navzájom oddelené podtržníkom:
ID_{i}_{e}_N{G}_DATA_FEA_N{it}_D_{raw}_FS{kHz}[_CH]_{ivec.sh}
Význam jednotlivých častí je nasledovný:

- **ID** – jednoznačný identifikátor každého experimentu, ktorý sa pri vytvorení ďalšieho skriptu automaticky inkrementuje. Je reprezentovaný trojciferným celým číslom počínajúc nulou, napr. *025.*
- **{i}** – konštantný textový reťazec *ivector*
- **{e}** – konštantný textový reťazec *extractor*

- **N{G}** – počet Gaussoviek trénovaného UBM GMM modelu potrebného pre i-vektor systém (napr. *512G*)
- **DATA** – špecifikácia datasetu, ktorý má byť pri trénovaní použitý, skript podporuje tieto 2 datasety:
  - *development2016* – development dataset súťaže DCASE 2016
  - *development2017* – development dataset súťaže DCASE 2017
- **FEA** – špecifikácia príznakov, ktoré majú byť pri trénovaní UBM GMM modelu extrahované z audio signálu. Skript podporuje dva varianty príznakov – banky filtrov a MFCC koeficienty, ktoré sú popísané v podsekcii 5.1.1. Pre účely i-vektor systému bude preferovaný variant MFCC koeficientov.
- **N{it}** – počet iterácií potrebných na natrénovanie UBM GMM modelu (napr. *10it*)
- **D** – požadovaná dimenzia i-vektoru. Pracoval som s variantmi *100, 200, 400, 500* a *600* rozmerných i-vektorov.
- **{raw}** – konštantný textový reťazec *raw* reprezentujúci formát audio nahrávok, s ktorým bude skript pracovať
- **FS{kHz}** – vzorkovacia frekvencia trénovacích dát v kiloherzoch, ktorá sa má pri trénovaní použiť. Skript počíta s troma variantmi: *8kHz, 16kHz* a *44.1kHz*.
- **[CH]** – voliteľná časť názvu skriptu, špecifikuje kanál audio nahrávok, ktorý má byť použitý (vzhľadom na to, že dáta sú nahrávané ako stereo nahrávky). Skript berie do úvahy 3 varianty:
  - *R* – táto možnosť špecifikuje, že sa majú použiť nahrávky z pravého audio kanálu
  - *L* – táto možnosť špecifikuje, že sa majú použiť nahrávky z ľavého audio kanálu
  - v prípade, že táto možnosť nie je v názve špecifikovaná, skript počíta s priemerom ľavého a pravého audio kanálu
- **{ivec.sh}** – konštantný textový reťazec *ivec.sh*

Pre úplnosť uvádzam príklad názvu konkrétneho experimentového skriptu:
`057_ivector_extractor_4G_development2017_MFCC24-C0-D-NDD-32-1_10it_100_raw_16kHz_R_ivec.sh`
Experimentový skript č. 57 natrénuje UBM GMM model so 4 Gaussovkami (a s 10 iteráciami) na development dátovej sade súťaže DCASE 2017, pričom k tomu využije 24 priamych MFCC koeficientov rozšírených o nultý koeficient a delta koeficienty (získané s kontextom 1), ktoré extrahuje prostredníctvom 32 bánk filtrov. V celom i-vektor systéme budú použité raw nahrávky so vzorkovacou frekvenciou 16 kHz z pravého audio kanálu. I-vektor extraktor následne extrahuje UBM GMM supervektory na 100 rozmerné i-vektory, ktoré budú vstupom do klasifikátoru.

Funkcionalita experimentového skriptu

V zložke, v ktorej sú umiestnené experimentové skripty, sa nachádza aj podzložka s názvom `workdir/`. Do tejto podzložky vytvorí na začiatku svojho vykonávania každý experimentový skript novú zložku, ktorej názov je rovnaký ako názov skriptu, líši sa jedine v koncovke `.dir`. Pre jednoduchosť nazvime túto zložku ako *výstupná zložka experimentu.* Výstupná zložka experimentu slúži na uchovávanie všetkých potrebných výstupov z experimentového skriptu, čo umožňuje spätné dohľadanie potrebných informácií o každom experimente. V tejto zložke skript vytvorí ďalšie 4 podzložky:

- `stats/` - slúži na uloženie UBM GMM štatistík tak pre trénovacie dáta, ako aj pre evaluačné dáta
- `ivec/` - slúži na uloženie i-vektorov pre trénovacie a aj evaluačné nahrávky
- `iXtractor/` - slúži na uloženie natrénovaného i-vektor extraktoru
- `models/` - slúži na uloženie natrénovaných UBM GMM modelov. Tak ako v prípade GMM modelov v predošlej kapitole, tak aj v tejto kapitole sa ukladá jeden model na jeden počet Gaussoviek, čiže napr. pri počte 128 Gaussoviek sa uložia modely aj pre všetky menšie počty Gaussoviek, t.j. 1, 2, 4, 8, 16, 32 a 64.

Po tom, čo skript vytvorí tieto zložky, dochádza k trénovaniu UBM GMM modelu pomocou pythonového skriptu, ktorý je na to určený. Následne, keď je UBM model natrénovaný, spustí experimentový skript iný pythonový skript, ktorého úlohou je vygenerovať UBM GMM štatistiky pre všetky trénovacie a evaluačné audio nahrávky. Akonáhle sú vygenerované tieto štatistiky, experimentový skript deleguje vykonávanie na matlabový skript, ktorý prostredníctvom vygenerovaných štatistík natrénuje i-vektor extraktor.

I-vektor extraktor je nevyhnutne nutný pre vygenerovanie očakávaných i-vektorov. I-vektory sa generujú tak pre trénovacie nahrávky ako aj pre evaluačné. Z vygenerovaných i-vektorov potom experimentový skript vytvorí spoločnú maticu všetkých trénovacích i-vektorov a spoločnú maticu všetkých evaluačných i-vektorov.

Tieto matice sú napokon vstupom do lineárneho Gaussovského klasifikátora, ktorý v mojej práci reprezentuje ďalší matlabový skript. Ten uloží svoj výstup do textového súboru v rámci výstupnej zložky experimentu, v ktorom sa nachádza priebeh i finálne výsledky klasifikácie. Po ukončení klasifikácie sa vykonávanie experimentového skriptu ukončí a tým sa ukončí aj celý experiment.

6.2 Systém založený na i-vektoroch

Zvyšok tejto kapitoly je venovaný podrobnejšiemu popisu konkrétnych experimentov, ktoré som v rámci tejto práce vykonal so systémom založeným na i-vektoroch. Cieľom vykonaných experimentov je aj v tomto prípade nájdenie optimálnej kombinácie hodnôt jednotlivých parametrov, ktoré sa podieľajú na tvorbe systému, aby výsledný systém dosahoval čo najlepšiu úspešnosť detekcie akustického prostredia. Podobne ako tomu bolo v sekciách 5.2 a 5.3, aj v tejto sekcii budú prezentované výsledky, či už v podobe grafu alebo tabuliek, na základe ktorých bude možné pozorovať vplyv jednotlivých parametrov na úspešnosť celého i-vektor systému.

Zoznam parametrov, s ktorými som v tejto časti experimentoval, a zároveň aj ich východiskové hodnoty zobrazuje tabuľka 6.1. Jedným z hlavných parametrov, ktorý je špecifický

Tabuľka 6.1: Prehľad východiskových hodnôt jednotlivých parametrov i-vektor systému.

Parameter	**Východisková hodnota**
Počet Gaussoviek	128
MFCC konfigurácia	iba priame koeficienty
Nultý MFCC koeficient	E (energia)
Vzorkovacia frekvencia	16 kHz
Audio kanál	priemer pravého a ľavého kanálu
Rozmer i-vektoru	200

pre i-vektor systém, je práve rozmer i-vektorov použitých v systéme. Ich vplyv na úspešnosť takéhoto systému bude taktiež súčasťou experimentovania. Hodnoty niektorých parametrov, hlavne počet MFCC koeficientov a bánk filtrov, neboli v prípade i-vektor systému skúmané príliš do detailov, a preto je v experimentoch uvedených len pár hodnôt, medzi ktorými sa rozhodovalo. Úmyslom bolo zamerať sa viac na ostatné parametre systému.

Ak nie je explicitne uvedené inak, sú UBM GMM modely všetkých i-vektor systémov spomenutých v ďalšej časti natrénované na development aj evaluačnej dátovej sade zo súťaže DCASE 2017, i-vektor extraktor týchto systémov je natrénovaný na development dátovej sade DCASE 2017 a celý systém je evaluovaný na evaluačnej dátovej sade súťaže DCASE 2017.

6.2.1 Počet MFCC koeficientov a bánk filtrov

Úvodným experimentom zistíme, koľko priamych MFCC koeficientov a bánk filtrov najlepšie vyhovuje i-vektor systému z hľadiska úspešnosti. Systém, ktorého ostatné parametre majú východiskové hodnoty z tabuľky 6.1, bol testovaný s niekoľkými kombináciami počtu MFCC koeficientov a bánk filtrov. Tieto kombinácie boli zvolené predovšetkým na základe získaných skúseností a poznatkov z experimentov na GMM systémoch.

Tabuľka 6.2: Tabuľka znázorňujúca vplyv počtu MFCC koeficientov a bánk filtrov na úspešnosť i-vektor systému pri použití východiskových hodnôt ostatných parametrov systému.

Počet koeficientov	**Počet bánk filtrov**	**Úspešnosť [%]**
20	40	54.56
20	32	53.57
24	**32**	**57.96**
20	24	54.56
19	24	55.12

Tabuľka 6.2 ukazuje, ako osobitné kombinácie vplývajú na úspešnosť celého systému. Najúspešnejšie sa javí použitie 24 priamych MFCC koeficientov v kombinácii s 32 bankami filtrov. Takýto systém dosiahol najvyššiu úspešnosť 57.96% a preto jeho nastavenie použijem ako základ pre ďalšie experimenty.

6.2.2 Audio kanál vstupných nahrávok

Ako bolo už spomenuté aj v minulých kapitolách, nahrávky z datasetu súťaže DCASE boli nahraté ako stereo nahrávky, a preto ďalšia kategória experimentov skúmala práve vplyv použitia nahrávok z rôznych audio kanálov na úspešnosť i-vektor systému.

Tabuľka 6.3: Tabuľka zobrazujúca úspešnosť i-vektor systémov líšiacich sa v nahrávkach z rôznych audio kanálov.

Audio kanál	**Úspešnosť [%]**
Priemer pravého a ľavého kanálu	57.96
Pravý	**60.25**
Ľavý	51.85

Údaje v tabuľke 6.3 potvrdzujú to, čo už ukázali aj experimenty s GMM systémami. Najviac sa darí detekovať akustické prostredie systému, ktorý má na vstupe nahrávky z pravého audio kanálu. V tomto prípade sa jedná o zlepšenie úspešnosti na úroveň 60.25%. Vplyv ostatných variantov kanálov má taktiež známy priebeh, ako ukázali aj GMM experimenty – ľavý audio kanál získal najmenšiu úspešnosť, ktorú kompenzuje priemer oboch kanálov. Priemer získal úspešnosť približujúcu sa k úspešnosti pravého kanálu so stratou necelých 3%. Aj tieto výsledky podporujú domnienku, že všetky nahrávky boli nahrávané zariadením, ktorého ľavý mikrofón mohol byť poškodený, alebo iným spôsobom znevýhodnený oproti pravému mikrofónu.

6.2.3 Počet Gaussoviek

Ďalšou časťou experimentovania bolo zistiť, pri akom počte Gaussoviek sa natrénuje UBM GMM model tak, že i-vektor systém následne dosahuje čo najlepšie výsledky. V tomto prípade sa zatiaľ jedná o systém, ktorý je zložený, čo sa týka MFCC konfigurácie, len z priamych MFCC koeficientov rozšírených o nultý koeficient v podobe energie.

Tabuľka 6.4: Počet Gaussoviek trénovaného UBM GMM modelu v závislosti na úspešnosti systému, ktorý je tvorený len priamymi MFCC koeficientami a nultým koeficientom, a 200 rozmernými i-vektormi.

Počet Gaussoviek	**Úspešnosť [%]**
32	57.28
64	57.67
128	60.25
256	60.45
512	61.05
1024	60.65
2048	**62.30**
4096	61.31
8192	61.84

Z tabuľky 6.4 môžme usúdiť, že úspešnosť takéhoto systému priamo úmerne závisí na počte Gaussoviek, t.j. čím viac Gaussoviek, tým lepšia úspešnosť. Najvyšší bod v tomto prípade

je 2048 Gaussoviek, kedy úspešnosť ešte stúpala, pri väčšom počte dochádza už ale k pretrénovaniu systému a úspešnosť klesá. V nasledujúcom experimente bude preto použitých práve 2048 Gaussoviek.

6.2.4 MFCC konfigurácia

Experimenty tejto časti skúmajú, aká kombinácia priamych a odvodených MFCC koeficientov spolu s nultým koeficientom je najlepšia z hľadiska úspešnosti i-vektor systému. Podrobnejšie výsledky tohto experimentu sa nachádzajú v tabuľke 6.5.

Tabuľka 6.5: Vplyv rôznych kombinácií priamych MFCC koeficientov s nultým a odvodenými koeficientami na úspešnosť i-vektor systému *(NC0=bez nultého koeficientu, C0=štandardný nultý MFCC koeficient, E=energia použitá ako nultý koeficient).*

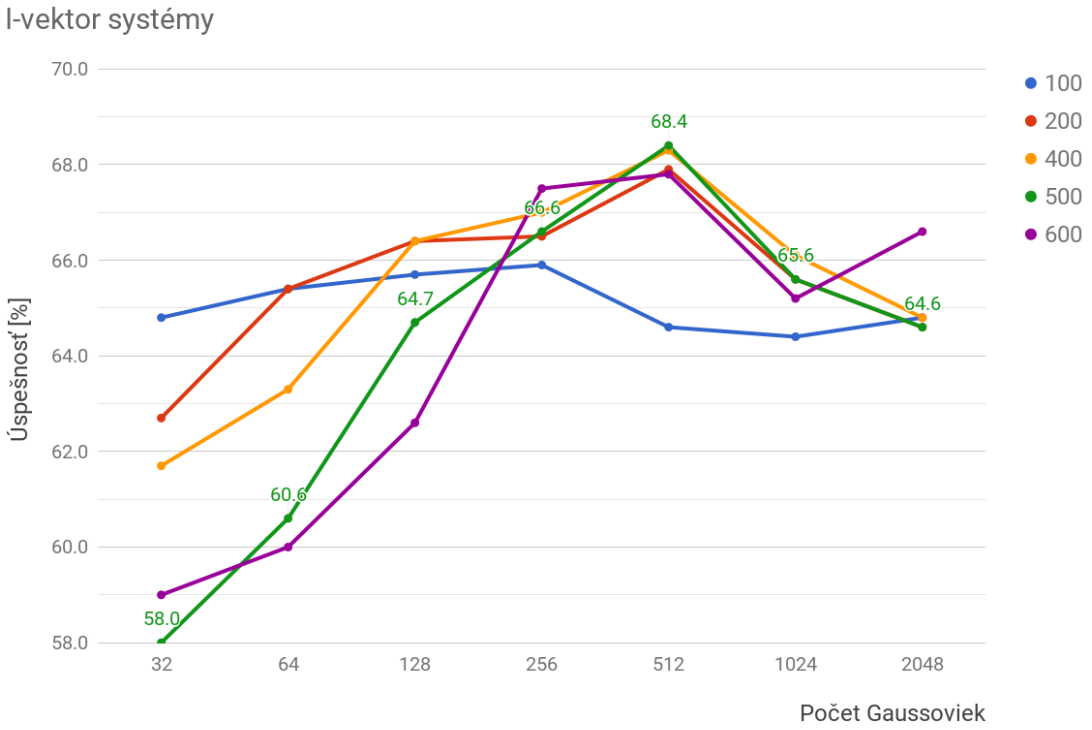
MFCC konfigurácia		
Priame a odvodené koeficienty	**Nultý koeficient**	**Úspešnosť [%]**
Priame koeficienty	NC0	61.84
	C0	62.30
	E	62.30
Priame + delta koeficienty	NC0	62.17
	C0	**64.63**
	E	61.48
Priame + delta + doubledelta koeficienty	NC0	58.60
	C0	63.49
	E	57.94

Z tejto tabuľky možno vyčítať, že celkom jednoznačne patrí v tejto kategórii víťazstvo priamym MFCC koeficientom rozšíreným o štandardný nultý MFCC koeficient C0 v spoločnej kombinácii s delta koeficientami. Pri takomto nastavení systému bola dosiahnutá úspešnosť až 64.63%. Na základe údajov z tabuľky môžme tiež konštatovať, že prítomnosť štandardného nultého MFCC koeficientu v každej kombinácii s priamymi a odvodenými koeficientami dosahuje všeobecne najlepšie výsledky, zatiaľ čo koeficient E (energia) sa zväčša podieľa na úspešnosti najmenej. Poznatok z tohto experimentu bude teda zohľadnený v systémoch ďalších experimentov.

6.2.5 Rozmer i-vektoru

Hlavnou kategóriou experimentov v celej tejto kapitole je práve kategória, ktorá skúma dimenziu i-vektorov použitých pre klasifikáciu prostredia v závislosti na úspešnosti systému. Tieto experimenty som sa rozhodol vykonať v spojení s ďalším parametrom – počtom Gaussoviek, aby bolo názornejšie a komplexnejšie vidieť, čo sa deje s úspešnosťou celého i-vektor systému pri zmenách oboch týchto parametrov. Experimentoval som s nasledujúcimi hodnotami rozmerov i-vektorov: 100, 200, 400, 500 a 600. Počty Gaussoviek som volil v rozsahu od 32 do 2048.

Obrázok 6.1 graficky zobrazuje úspešnosť i-vektor systémov v závislosti na oboch spomínaných parametroch. Z grafu je jasne vidieť, že pre takéto i-vektor systémy sa najviac hodí 512 Gaussoviek. Čo sa týka rozmeru i-vektorov, najlepšiu úspešnosť dosahujú systémy



Obr. 6.1: Grafické znázornenie úspešnosti i-vektor systému v závislosti na počte Gaussoviek a dimenzionalite i-vektorov.

so 400 a 500 rozmernými i-vektormi, pričom 500 rozmerné i-vektory majú o trochu vyššiu úspešnosť. Systém, ktorý dosiahol najlepšiu úspešnosť – konkrétne 68.40% – má teda 500 rozmerné i-vektory a jeho UBM GMM model bol natrénovaný pomocou 512 Gaussoviek. Tabuľka 6.6 obsahuje detailné výsledky jednotlivých systémov zobrazených na obrázku 6.1.

Tabuľka 6.6: Úspešnosť i-vektor systému [%] vzhľadom na počet Gaussoviek a dimenzionalitu i-vektorov.

	Rozmer i-vektoru				
Počet Gaussoviek	100	200	400	**500**	600
32	64.75	62.72	61.73	57.96	58.95
64	65.43	65.43	63.33	60.56	60.00
128	65.68	66.42	66.42	64.69	62.59
256	**65.93**	66.54	66.98	66.61	67.53
512	64.63	**67.90**	**68.27**	**68.40**	**67.84**
1024	64.44	65.62	66.11	65.56	65.25
2048	64.82	64.63	64.75	64.63	66.61

Okrem týchto hlavných poznatkov tu môžme pozorovať ešte jednu vec v súvislosti s tabuľkou 6.4, ktorá zobrazuje vplyv počtu Gaussoviek na i-vektor systém tvorený len pria-

mymi MFCC koeficientami a energiou ako nultým koeficientom. Ten experiment ukázal, že sa najlepšie javí systém s 2048 Gaussovkami, zatiaľ čo tento experiment celkom jednoznačne ukázal, že sa najlepšie javí 512 Gaussoviek pre i-vektor systém tvorený priamymi a delta MFCC koeficientami spolu so štandardným nultým MFCC koeficientom C0. To znamená, že počet Gaussoviek je veľmi závislý na konkrétnej MFCC konfigurácii, na základe ktorej môže mať graf úspešnosti i-vektor systému odlišný priebeh.

6.2.6 Vzorkovacia frekvencia audio nahrávok

Záverečná kategória experimentov tejto kapitoly sa venuje vzorkovacej frekvencii vstupných audio nahrávok. GMM experimenty ukázali aj v prípade bánk filtrov, aj v prípade MFCC koeficientov, že najlepšie výsledky dosahujú systémy, ktoré používali nahrávky so vzorkovacou frekvenciou 16 kHz. Tabuľka 6.7 ukazuje, ako tomu bolo v prípade i-vektor systému.

Tabuľka 6.7: Vplyv vzorkovacej frekvencie vstupných audio nahrávok na i-vektor systém.

Vzorkovacia frekvencia	**Úspešnosť [%]**
8 kHz	63.03
16 kHz	**68.40**
44.1 kHz	66.54

Aj v prípade i-vektor systému sa potvrdilo, že najlepšie výsledky dosahuje systém, ktorý využíva audio nahrávky so vzorkovacou frekvenciou 16 kHz. Trochu nižšiu úspešnosť má systém s nahrávkami, ktorých vzorkovacia frekvencia je 44.1 kHz a najmenej úspešné sa zdajú byť 8 kHz nahrávky.

6.2.7 Zhrnutie experimentov na i-vektor systémoch

Experimenty s i-vektor systémom boli vykonané podobným spôsobom ako v prípade GMM systémov, t.j. spočiatku bola použitá východisková konfigurácia všetkých uvažovaných parametrov systému, ktorú postupne nahradzovali iné hodnoty, ktoré sa v osobitných experimentoch ukázali ako úspešnejšie.

Tabuľka 6.8: Prehľad hodnôt parametrov najlepšieho i-vektor systému a dosiahnutej úspešnosti. Hodnoty sú získané na základe vykonaných experimentov.

Parameter	**Hodnota**
MFCC konfigurácia	priame + delta koeficienty
Nultý MFCC koeficient	C0
Počet MFCC koeficientov	24
Počet bánk filtrov	32
Počet Gaussoviek	512
Vzorkovacia frekvencia	16 kHz
Audio kanál	pravý
Rozmer i-vektoru	500
Úspešnosť [%]	68.40
Úspešnosť [nahrávky]	1108 z 1620

Prehľad finálnych hodnôt jednotlivých parametrov najlepšieho i-vektor systému zobrazuje tabuľka 6.8. Postup, ktorým som sa k týmto hodnotám dostal, je popísaný krok po kroku prostredníctvom experimentov analyzovaných v predošlej časti práce.

Experimenty ukázali, že najvhodnejším nastavením MFCC koeficientov pre natrénovanie UBM GMM modelu je 24 priamych koeficientov rozšírených o štandardný nultý koeficient spolu s delta koeficientami a 32 bankami filtrov. Úspešnosť je najvyššia v prípade použitia 512 Gaussoviek pri trénovaní UBM GMM modelu. Tak ako ukázali aj experimenty s GMM systémom, najlepšie sa vzorkovacia frekvencia audio nahrávok javí ako 16 kHz. Okrem toho sa s GMM experimentami zhoduje tiež výsledok experimentov zaoberajúcich sa audio kanálom nahrávok. Najlepšie výsledky dosahovali systémy, ktorých nahrávky pochádzali z pravého audio kanálu. Experimenty, ktoré skúmali, aký veľký i-vektor je optimálny pre systém detekcie akustického prostredia, ukázali, že 500 rozmerné i-vektory sú tou správnou voľbou. Pri zohľadnení všetkých týchto poznatkov dosiahol najlepší i-vektor systém úspešnosť 68.40%.

Tabuľka 6.9: Confusion matica najlepšieho i-vektor systému, ktorý dosiahol úspešnosť 68.4% na evaluačných dátach 2017. Riadky predstavujú skutočné návestia a stĺpce reprezentujú návestia predpovedané systémom. Čísla predstavujú počet nahrávok, na jednu triedu pripadá 108 evaluačných nahrávok.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1-Pláž	**100**	0	1	0	0	1	0	0	0	0	0	3	3	0	0
2-Autobus	0	**60**	2	5	0	0	**27**	0	0	**11**	0	0	1	1	1
3-Kaviareň	0	0	**83**	0	0	2	2	2	0	**19**	0	0	0	0	0
4-Auto	0	2	0	**80**	0	0	1	0	0	0	0	1	0	5	**19**
5-Centrum mesta	3	0	0	0	**81**	0	3	0	0	2	0	**15**	4	0	0
6-Lesná cesta	3	0	0	0	0	**97**	0	0	0	0	0	1	7	0	0
7-Potraviny	0	0	5	0	0	0	**66**	2	3	**32**	0	0	0	0	0
8-Obydlie/Dom	1	0	1	0	0	0	2	**70**	3	0	**26**	1	3	1	0
9-Knižnica	0	0	0	0	0	8	1	**26**	**49**	5	**19**	0	0	0	0
10-Stanica metra	0	0	4	0	0	2	4	0	1	**92**	4	1	0	0	0
11-Kancelária	0	0	0	0	0	0	0	**10**	0	0	**98**	0	0	0	0
12-Park	0	0	0	0	4	4	**13**	1	0	7	0	**40**	**39**	0	0
13-Sídlisko	0	0	1	0	**11**	**21**	0	0	0	2	0	3	**70**	0	0
14-Vlak	0	4	0	0	2	0	0	0	4	2	0	0	0	**75**	**21**
15-Električka	0	4	8	0	6	0	**16**	0	1	**20**	0	0	0	6	**47**

Tabuľka 6.9 zobrazuje confusion maticu najlepšieho i-vektor systému. Matica popisuje chovanie celého i-vektor systému detekcie akustického prostredia, zobrazuje, ktoré triedy si systém navzájom pletie medzi sebou a zároveň, koľko nahrávok bolo rozpoznaných správne. Jednotlivé čísla predstavujú počty audio nahrávok. Čísla na hlavnej diagonále predstavujú počet korektne rozpoznaných audio nahrávok, čísla ležiace mimo hlavnej diagonále reprezentujú počty nesprávne rozpoznaných nahrávok.

Na základe matice môžme skonštatovať viaceré skutočnosti. Najúspešnejšie boli rozpoznané nahrávky z *pláže*, systém rozpoznal 100 evaluačných nahrávok korektne, čo znamená, že len zvyšných 8 nahrávok si poplietol s triedami iných prostredí. Veľmi úspešne boli tiež rozpoznané nahrávky z *kancelárie* a *lesnej cesty*. Systém si najviac plietol nahrávky z *parku,* ktoré správne rozpoznal iba v 40 prípadoch, a skoro rovnako veľa nahrávok (39) rozpoznal

nesprávne ako *sídlisko.* Tieto nahrávky si tiež relatívne často plietol aj s *obchodom s potravinami.* Ďalšou problematickou triedou nahrávok bola *električka.* Nahrávky z *električky* systém v 20 prípadoch rozpoznal ako nahrávky zo *stanice metra* a v 16 prípadoch ako *obchod s potravinami.* Pri rozpoznávaní mal systém problém aj s nahrávkami z *knižnice.* Confusion matica ukazuje, že v 26 prípadoch si ich systém pomýlil s nahrávkami z *obydlia/domu* a v 19 prípadoch s *kanceláriou.*

Nahrávky z *obchodu s potravinami* boli často (32 krát) rozpoznané ako nahrávky zo *stanice metra,* nahrávky z *autobusu* boli zase viackrát (27) rozpoznané ako nahrávky z *obchodu s potravinami.* Nahrávky z *obydlia/domu* boli 26 krát rozpoznané ako *kancelária,* problém systému robili tiež nahrávky z *vlaku,* ktoré v 21 prípadoch vyhodnotil ako nahrávky z *električky.*

Kapitola 7

Fúzia a zhrnutie výsledkov

Po experimentoch s GMM systémami a i-vektor systémami som sa rozhodol vyskúšať fúziu viacerých systémov a zistiť, do akej miery sa to prejaví na úspešnosti. Články viacerých autorov, ako som uviedol aj v sekciách 2.3 a 2.4, hovorili o tom, že použitím fúzie dvoch, prípadne viacerých systémov, dosiahli zvýšenie úspešnosti nového systému, ktorý je tvorený viacerými podsystémami. Pri čítaní týchto článkov som nadobudol pocit, že fúzia sa v poslednej dobe v tejto oblasti začala využívať naozaj často a prináša pozitívne výsledky. To bol dôvod a zároveň moja motivácia, prečo som chcel vyskúšať tiež fúziu aspoň dvoch systémov a pozorovať, ako sa jej použitie prejaví na finálnej úspešnosti.

Na fúziu systémov som používal Lineárny Gaussovský klasifikátor, ktorý som ale neimplementoval sám, pretože som základnú fungujúcu verziu dostal od môjho školiteľa. Túto verziu mi stačilo akurát prispôsobiť k môjmu riešeniu a rozhraniu a mohol som začať fúzovať jednotlivé systémy. V nasledujúcej sekcii budú popísané konkrétne kombinácie systémov, ktorých fúziu som vykonal a v záverečnej časti tejto kapitoly bude uvedené porovnanie mnou implementovaných systémov s najlepšou úspešnosťou so systémami zo súťaže DCASE 2017.

7.1 Fúzia

V rámci tejto práce som experimentoval s fúziou viacerých kombinácií systémov, preto pre prehľadnosť uvediem na začiatku zoznam všetkých použitých systémov a ich označenie, pomocou ktorého ich budem v tejto kapitole volať.

GMM systémy:

- **GMM1** – najlepší GMM systém (priame + delta koeficienty, 128 Gaussoviek, obrázok 5.12a)
- **GMM2** – druhý najlepší GMM systém (64 Gaussoviek, obrázok 5.12b)
- **GMM3** – GMM systém s doubledelta koeficientami spomínaný v kapitole 5.3.8 (128 Gaussoviek)

I-vektor systémy[1]:

- **IVEC1** – najlepší i-vektor systém (500 rozmerné i-vektory, pravý audio kanál)
- **IVEC1_L** – IVEC1 systém s ľavým audio kanálom

[1] Systémy IVEC1, IVEC2, IVEC3 a IVEC4 boli vybraté na základe tabuľky 6.6.

- **IVEC1_AVG** – IVEC1 systém, ktorého audio nahrávky sú spriemerované z ľavého a pravého audio kanálu
- **IVEC2** – druhý najúspešnejší i-vektor systém (400 rozmerné i-vektory, pravý audio kanál)
- **IVEC2_L** – IVEC2 s ľavým audio kanálom
- **IVEC2_AVG** – IVEC2 s audio kanálom, ktorým je priemer ľavého a pravého kanálu
- **IVEC3** – tretí najúspešnejší i-vektor systém (200 rozmerné i-vektory, pravý audio kanál)
- **IVEC4** – štvrtý najúspešnejší i-vektor systém (600 rozmerné i-vektory, pravý audio kanál)

7.1.1 Fúzia GMM a i-vektor systému

Tabuľka 7.1 zahŕňa 5 podtabuliek, ktoré zobrazujú výsledky fúzie najlepších GMM systémov s i-vektor systémami.

Tabuľka 7.1: Výsledky fúzie GMM systémov s i-vektor systémami.

Tabuľka 7.2

Systém	**Úspešnosť [%]**
GMM1	60.37
IVEC1	**68.40**
Fúzia	65.74

Tabuľka 7.3

Systém	**Úspešnosť [%]**
GMM1	60.37
IVEC2	**68.27**
Fúzia	65.68

Tabuľka 7.4

Systém	**Úspešnosť [%]**
GMM2	60.12
IVEC1	**68.40**
Fúzia	66.79

Tabuľka 7.5

Systém	**Úspešnosť [%]**
GMM3	59.26
IVEC2	68.27
Fúzia	**69.01**

Tabuľka 7.6

Systém	**Úspešnosť [%]**
GMM3	59.26
IVEC1	68.40
Fúzia	**69.32**

V tejto kategórii bolo vykonaných 5 rôznych fúzií, pričom zlepšenie výsledkov finálneho systému nastalo v 2 prípadoch, ktoré znázorňujú tabuľky 7.5 a 7.6. Dôvodom, pre ktorý nenastalo zlepšenie úspešnosti pre niektoré vykonané fúzie, môže byť napríklad fakt, že

oba fúzované systémy boli veľmi podobne natrénované a pomer ich úspešností pre jednotlivé triedy akustického prostredia je veľmi podobný a teda sa tieto dva systémy nemohli navzájom výraznejšie "doplniť" a vylepšiť tak úspešnosť.

Na základe výsledkov z tabuľky 7.1 možno usúdiť, že systém GMM3 mal všeobecne pozitívny vplyv pri fúzii, obe jeho fúzie sa podieľali na zvýšení úspešnosti. To sa ale nedá povedať o dvoch najlepších GMM systémoch, GMM1 a GMM2, ktoré ukázali, že v kombinácii s danými i-vektor systémami sa nepodieľajú na zvýšení úspešnosti.

7.1.2 Fúzia i-vektor systémov

Tabuľka 7.7 vo svojich 6 podtabuľkách zobrazuje výsledky fúzií, na ktorých sa podieľali len i-vektor systémy v rôznych kombináciách. V tomto prípade teda bolo vykonaných 6 fúzií i-vektor systémov a podľa výsledkov možno konštatovať, že až v 5 prípadoch sa fúzia prejavila pozitívne a úspešnosť systémov sa zdvihla.

Tabuľka 7.7: Výsledky fúzie i-vektor systémov.

Tabuľka 7.8

Systém	**Úspešnosť** [%]
IVEC2	68.27
IVEC2_L	59.94
IVEC2_AVG	64.26
Fúzia	**68.46**

Tabuľka 7.9

Systém	**Úspešnosť** [%]
IVEC1	68.40
IVEC1_L	58.64
IVEC1_AVG	63.77
Fúzia	**68.58**

Tabuľka 7.10

Systém	**Úspešnosť** [%]
IVEC2	68.27
IVEC3	67.90
Fúzia	**68.33**

Tabuľka 7.11

Systém	**Úspešnosť** [%]
IVEC1	**68.40**
IVEC2	68.27
Fúzia	68.27

Tabuľka 7.12

Systém	**Úspešnosť** [%]
IVEC1	68.40
IVEC2	68.27
IVEC3	67.90
Fúzia	**68.70**

Tabuľka 7.13

Systém	**Úspešnosť** [%]
IVEC1	68.40
IVEC2	68.27
IVEC3	67.90
IVEC4	67.84
Fúzia	**68.46**

Tabuľky 7.8 a 7.9 ukazujú fúziu 3 systémov pre 2 najlepšie i-vektor systémy. Na fúzii sa podieľali systémy, ktorých nahrávky sú aj z pravého aj z ľavého audio kanálu a tiež priemer oboch kanálov. Takto postavená fúzia sa prejavila v oboch prípadoch pozitívne.

Tabuľky 7.10 a 7.11 predstavujú výsledky fúzie 2 systémov, ktorými boli v prvom prípade druhý a tretí najlepší i-vektor systém a v druhom prípade 2 najlepšie i-vektor systémy. Prekvapivo sa úspešnosť zvýšila len v prvom prípade a v druhom prípade ostala nezme-

nená, čo pravdepodobne tiež vypovedá o tom, že prvé dva najlepšie i-vektor systémy sú natrénované veľmi podobne a líšia sa v jednotlivých triedach len minimálne.

Tabuľky 7.12 a 7.13 prezentujú zase výsledky prvých troch a prvých štyroch najlepších i-vektor systémov. V prípade troch fúzovaných systémov nastalo zvýšenie úspešnosti zo 68.40% na 68.70%, v prípade fúzie 4 systémov sa finálna úspešnosť síce trochu zvýšila, no oproti 3 fúzovaným systémom klesla na úroveň 68.46%.

7.1.3 Sumarizácia výsledkov fúzie

Spolu bolo vykonaných 11 fúzií rôznych kombinácií GMM a i-vektor systémov a práve 7 z nich viedlo k zvýšeniu úspešnosti. Pre prehľadnosť uvádzam zoznam týchto 7 fúzií v tabuľke 7.14 zoradených podľa dosiahnutej úspešnosti.

Tabuľka 7.14: Prehľad vykonaných fúzií, ktoré sa podieľali na zvýšení úspešnosti finálneho systému.

Fúzia systémov	**Úspešnosť [%]**
GMM3 + IVEC1	**69.32**
GMM3 + IVEC2	69.01
IVEC1 + IVEC2 + IVEC3	68.70
IVEC1 + IVEC1_L + IVEC1_AVG	68.58
IVEC2 + IVEC2_L + IVEC2_AVG	68.46
IVEC1 + IVEC2 + IVEC3 + IVEC4	68.46
IVEC2 + IVEC3	68.33

Tabuľka 7.15: Confusion matica najlepšieho fúzovaného systému, ktorý dosiahol úspešnosť 69.32% na evaluačných dátach 2017. Riadky predstavujú skutočné návestia a stĺpce reprezentujú návestia predpovedané systémom. Čísla predstavujú počet nahrávok, na jednu triedu pripadá 108 evaluačných nahrávok.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1-Pláž	**93**	0	1	0	0	0	0	0	0	0	0	5	9	0	0
2-Autobus	0	**54**	1	6	0	0	**26**	1	0	**13**	0	4	0	2	1
3-Kaviareň	0	0	**86**	0	0	0	2	1	0	**19**	0	0	0	0	0
4-Auto	0	2	0	**82**	0	0	1	0	0	0	0	0	0	0	**23**
5-Centrum mesta	3	0	0	0	**90**	0	0	0	0	2	0	**10**	3	0	0
6-Lesná cesta	1	0	0	0	0	**92**	0	0	0	0	0	7	8	0	0
7-Potraviny	0	0	3	0	1	0	**70**	4	0	**30**	0	0	0	0	0
8-Obydlie/Dom	0	0	0	0	0	2	0	**98**	0	2	1	1	4	0	0
9-Knižnica	0	0	0	0	0	**11**	1	**25**	**51**	**17**	3	0	0	0	0
10-Stanica metra	0	0	4	0	0	2	0	0	0	**97**	3	2	0	0	0
11-Kancelária	0	0	0	0	0	**17**	0	**28**	0	0	**63**	0	0	0	0
12-Park	0	0	2	0	5	4	**17**	1	0	5	0	**42**	**32**	0	0
13-Sídlisko	0	0	1	0	**11**	9	0	0	0	2	0	2	**83**	0	0
14-Vlak	0	3	0	0	3	0	0	0	3	5	0	0	0	**69**	**25**
15-Električka	0	2	9	0	5	0	**23**	0	0	**16**	0	0	0	0	**53**

Najúspešnejším systémom sa teda javí byť fúzia systémov GMM3 a IVEC1. Chovanie tohto systému detailne stvárňuje confusion matica v rámci tabuľky 7.15. Tento systém mal najmenší problém pri rozpoznávaní nahrávok z prostredia *obydlia/domu.* Podarilo sa mu správne rozpoznať 98 nahrávok, takže len 10 ďalších nahrávok rozpoznal nesprávne. Veľmi podobne úspešne systém klasifikoval nahrávky zo *stanice metra,* kedy správne rozpoznal 97 nahrávok. Naopak, najmenej sa systému darilo rozpoznať nahrávky z *parku,* pri ktorých správne určil len 42 nahrávok. Často (až 32 krát) si ich splietol so *sídliskom,* v 17 prípadoch určil, že sa jedná o nahrávky z *obchodu s potravinami.* Problematickou triedou pri rozpoznávaní bola aj *knižnica.* Systém správne identifikoval 51 nahrávok, ďalších 25 si pomýlil s triedou *obydlia/domu,* iných 17 nahrávok určil ako *stanicu metra.*

Na základe confusion matice tiež môžme pozorovať, že si systém často (30 krát) plietol nahrávky z *obchodu s potravinami* so *stanicou metra.* Taktiež nahrávky z *kancelárie* boli viackrát (28) klasifikované ako nahrávky z *obydlia/domu.* Nahrávky z *vlaku* systém v 25 prípadoch rozpoznal ako nahrávky z *električky.* Nahrávky z *električky* boli zase 23 krát identifikované ako nahrávky z *auta.* Môžme teda konštatovať, že nahrávky z dopravných prostriedkov si systém často plietol navzájom, čo je aj celkom pochopiteľné, pretože aj ľudia majú často problém rozpoznať, v akom dopravnom prostriedku sa nachádza človek, s ktorým hovoríme v telefóne, apod.

7.2 Zhrnutie výsledkov

Na záver by som rád zhrnul dosiahnuté výsledky tejto práce a porovnal ich s niektorými systémami zo súťaže DCASE Challenge 2017. Tabuľka 7.16 zobrazuje informácie o najlepších systémoch implementovaných v rámci tejto práce a tiež o kľúčových systémoch súťaže DCASE 2017.

Tabuľka 7.16: Prehľad dosiahnutých výsledkov v tejto práci a porovnanie s kľúčovými systémami súťaže DCASE 2017. Všetky systémy sú evaluované na evaluačnej dátovej sade súťaže DCASE 2017.

Systém	**Úspešnosť [%]**
najlepší GMM systém (banky filtrov)	54.63
najlepší GMM systém (MFCC koeficienty)	60.37
najlepší i-vektor systém	68.40
najlepšia fúzia (GMM + i-vektor)	69.32
GMM baseline 2017 systém[2]	52.16
NN baseline 2017 systém[3]	61.00
najlepší i-vektor systém súťaže DCASE 2017 [16]	68.70
víťazný systém súťaže DCASE 2017 [18]	83.30

Na základe údajov z tabuľky môžme konštatovať, že všetky implementované systémy dosahujú úspešnosť vyššiu ako GMM baseline súťaže DCASE. Okrem toho je tiež vidieť, že najlepší GMM systém s MFCC koeficientami dosahuje bez pár desatín percenta úspešnosť baseline systému súťaže DCASE založeného na neurónových sieťach.

[2] Jedná sa o GMM baseline systém, ktorý bol poskytnutý pre ročník 2016 súťaže DCASE [9], ktorý som evaluoval na dátach z roku 2017.

[3] Baseline systém založený na neurónových sieťach (NN) [8].

V prípade implementovaného i-vektor systému sa môj najúspešnejší variant dostal s úspešnosťou takmer na úroveň najlepšieho i-vektor systému súťaže DCASE 2017, ktorého autorom je tím ľudí z Rakúska, ktorí sa stali víťazmi súťaže DCASE v roku 2016. Rozdiel medzi týmito i-vektor systémami je len 0.30%.

Fúzia tvorená GMM systémom a i-vektor systémom dosahuje najlepšiu úspešnosť spomedzi všetkých mojich implementovaných systémov. Úspešnosť fúzie 69.32% na evaluačných dátach súťaže DCASE 2017 by v čase súťaže v roku 2017 znamenala umiestnenie na 20.mieste v celkovom hodnotení všetkých odovzdaných systémov z celého sveta, ktorých by v prípade aj mojej účasti bolo dohromady 98. V prípade tímového hodnotenia by to znamenalo 12.miesto spomedzi 42 medzinárodných tímov.

Ďalšou motiváciou pre zlepšovanie úspešnosti implementovaných systémov môže byť napríklad aj víťazný systém súťaže, ktorý dosiahol úspešnosť až 83.30%.

Kapitola 8

Záver

Táto práca sa zaoberá detekciou akustického prostredia z audio nahrávok, pričom sa berie do úvahy 15 rôznych tried akustických prostredí, v ktorých sa ľudia bežne pohybujú, ako napríklad mestský park, kancelária, vlak, kaviareň, či knižnica. Je to zároveň téma, ktorej sa aktívne venuje aj medzinárodná súťaž s názvom DCASE Challenge a preto táto práca veľmi úzko súvisí práve s touto súťažou. Súťaž je organizovaná skupinou nadšencov o spracovanie reči a zvuku a zúčastniť sa jej môže ktokoľvek, koho táto problematika zaujíma. Výhodou súťaže DCASE Challenge je, že má priamočiare zadanie, jasné pravidlá a dokonca organizátori pripravili aj audio dataset, pomocou ktorého majú súťažiaci budovať svoje systémy.

Na základe poznatkov zo systémov, ktoré boli vytvorené v rámci súťaže v rokoch 2016 a 2017 som sa rozhodol, že túto úlohu budem riešiť dvoma rôznymi spôsobmi. Prvým spôsobom je GMM systém, ktorý som stavial nad dvoma typmi príznakov – nad bankami filtrov a MFCC koeficientami. Druhým spôsobom je systém, ktorý je založený na i-vektoroch získaných z i-vektor extraktoru. V práci sú popísané sady experimentov, ktoré som vykonal s danými systémami, aby som našiel optimálnu kombináciu všetkých skúmaných parametrov systémov s cieľom získať čo najvyššiu úspešnosť detekcie.

Výsledky ukázali, že implementované metódy naozaj fungujú. Najlepší GMM systém založený na bankách filtrov dosiahol úspešnosť 54.63% na evaluačných dátach súťaže DCASE 2017, čo je v porovnaní s GMM baseline systémom súťaže viac, pretože ten dosiahol na rovnakých dátach úspešnosť 52.16%. GMM systém založený na MFCC koeficientoch dosiahol všeobecne lepšiu úspešnosť ako GMM systém s bankami filtrov, čím sa potvrdili všeobecne známe poznatky. Najlepší GMM systém s MFCC koeficientami dosiahol úspešnosť 60.37%, čo je skoro o 6 percentuálnych bodov viac ako v prípade systému s bankami filtrov.

V prípade systému založeného na i-vektoroch sa celková úspešnosť detekcie akustického prostredia zvýšila oproti úspešnosti GMM systémov. Najlepší i-vektor systém dosiahol na evaluačných dátach súťaže DCASE 2017 úspešnosť 68.40%, čo znamená, že použitie i-vektor systému dokázalo zlepšiť úspešnosť detekcie GMM systému s MFCC koeficientami o viac ako 13% relatívne. Okrem toho, najlepší i-vektor systém, ktorý sa zúčastnil súťaže DCASE 2017, patril rakúskemu tímu, ktorý vyhral túto súťaž v roku 2016. Úspešnosť tohto najlepšieho súťažného i-vektor systému bola 68.70%, čo znamená, že i-vektor systém implementovaný v tejto práci sa dostal takmer na úroveň najlepšieho i-vektor systému súťaže DCASE 2017.

Vzhľadom na to, že autori súťažných systémov často využívali fúziu systémov, ktoré vybudovali, pretože sa pozitívne prejavila na finálnej úspešnosti celého systému, som sa napokon aj ja rozhodol zakomponovať do práce fúziu, či už GMM systému s i-vektor systé-

mom, alebo fúziu viacerých i-vektor systémov. Fúzia sa v závislosti na rôznych systémoch prejavila rôzne na finálnej úspešnosti. V istých prípadoch sa nepodieľala zlepšením úspešnosti, v iných prípadoch dokázala zlepšiť úspešnosť. Najlepší výsledok priniesla fúzia jedného GMM systému s najlepším i-vektor systémom. Jej úspešnosť sa dostala na úroveň 69.32%, čo predstavuje zároveň najúspešnejší systém, ktorý som v rámci tejto práce vytvoril.

V čase súťaže DCASE 2017 by systém s takouto úspešnosťou obsadil 20.miesto v celkovom hodnotení všetkých odovzdaných systémov z celého sveta, t.j. 20.miesto spomedzi 98 systémov z celého sveta. V prípade tímového hodnotenia by sa jednalo o 12.miesto zo 42 medzinárodných tímov.

V závere by som ešte spomenul jeden zaujímavý experiment, ktorý vykonali autori víťazného systému súťaže DCASE 2016 [16]. Vo svetle tohto experimentu možno nadobúdajú uvedené čísla jednotlivých úspešností trochu iný rozmer. Títo autori, pôsobiaci na univerzite v rakúskom Linzi, dali klasifikovať evaluačné audio nahrávky zo súťaže DCASE 2017 svojim študentom, aby zistili aspoň približne, s akou úspešnosťou dokážu ľudia rozpoznať jednotlivé akustické prostredia z nahrávok. Výsledkom bolo, že študenti dokázali správne rozpoznať v priemere zhruba 50% audio nahrávok.

Čo sa týka budúceho vývoja, ktorým by sa práca mohla uberať, určite by stálo za to skúsiť vytvoriť systém založený na neurónových sieťach. Prvých 5 tímov, ktoré vyhrali súťaž DCASE 2017, vo svojom riešení totiž nejakým spôsobom mali zahrnuté neurónové siete, takže ich pozitívny vplyv na detekciu akustického prostredia je nespochybniteľný. Okrem toho by bolo vhodné zapracovať ešte viac na samotnej fúzii systémov, možno vyskúšať inú metódu fúzie ako pomocou Lineárneho Gaussovského klasifikátoru. Značným problémom je tiež málo trénovacích dát, takže by bolo určite prospešné skúsiť rozšíriť dataset napríklad duplikáciou dát, apod. Ďalším možným budúcim experimentom by mohlo byť použitie iných druhov príznakov, napr. PLP, PNCC koeficienty, apod. Za zmienku stojí tiež aplikovanie techniky *voice activity detection,* ktorou by došlo k odfiltrovaniu určitých nerelevantných častí vstupného signálu.

Literatúra

[1] Bishop, C.: *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer New York, 2016, ISBN 9781493938438.
URL https://books.google.cz/books?id=kOXDtAEACAAJ

[2] Bisot, V.; Serizel, R.; Essid, S.; aj.: Supervised Nonnegative Matrix Factorization for Acoustic Scene Classification. Technická správa, DCASE2016 Challenge, September 2016.

[3] Brümmer, N.: The EM algorithm and minimum divergence. *Online http://niko. brummer. googlepages. Agnitio Labs Technical Report*, 2009.

[4] Eghbal-Zadeh, H.; Lehner, B.; Dorfer, M.; aj.: CP-JKU Submissions for DCASE-2016: a Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks. Technická správa, DCASE2016 Challenge, September 2016.

[5] Gerhard, D.: *Audio signal classification: History and current techniques*. Citeseer, 2003.

[6] Glembek, O.; Burget, L.; Kenny, P.; aj.: Simplification and optimization of I-Vector Extraction. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, IEEE Signal Processing Society, 2011, ISBN 978-1-4577-0537-3, s. 4516–4519.
URL http://www.fit.vutbr.cz/research/view_pub.php.cs?id=9655

[7] Han, Y.; Park, J.: Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification. Technická správa, DCASE2017 Challenge, September 2017.

[8] Heittola, T.; Mesaros, A.: DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System. Technická správa, DCASE2017 Challenge, September 2017.

[9] Heittola, T.; Mesaros, A.; Virtanen, T.: DCASE2016 Baseline System. Technická správa, DCASE2016 Challenge, September 2016.

[10] Heittola, T.; aj.: *Detection and Classification of Acoustic Scenes and Events 2016*. [Online; navštívené 15.03.2018].
URL http://www.cs.tut.fi/sgn/arg/dcase2016/

[11] Heittola, T.; aj.: *Detection and Classification of Acoustic Scenes and Events 2017*. [Online; navštívené 20.03.2018].
URL http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/

[12] Heittola, T.; aj.: *Detection and Classification of Acoustic Scenes and Events 2018*. [Online; navštívené 15.05.2018].
URL http://dcase.community/challenge2018/

[13] Hyder, R.; Ghaffarzadegan, S.; Feng, Z.; aj.: BUET Bosch Consortium (B2C) Acoustic Scene Classification Systems for DCASE 2017. Technická správa, DCASE2017 Challenge, September 2017.

[14] Karafiát, M.; Burget, L.; Matějka, P.; aj.: iVector-Based Discriminative Adaptation for Automatic Speech Recognition. In *Proceedings of ASRU 2011*, IEEE Signal Processing Society, 2011, ISBN 978-1-4673-0366-8, s. 152–157.
URL http://www.fit.vutbr.cz/research/view_pub.php.cs?id=9762

[15] Kenny, P.; Boulianne, G.; Oullet, P.; aj.: Joint Factor Analysis versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 15, č. 7, 2007: s. 2072–2084, ISSN 1558-7916.

[16] Lehner, B.; Eghbal-Zadeh, H.; Dorfer, M.; aj.: Classifying Short Acoustic Scenes with I-Vectors and CNNs: Challenges and Optimisations for the 2017 DCASE ASC Task. Technická správa, DCASE2017 Challenge, September 2017.

[17] Martínez, D. G.; Plchot, O.; Burget, L.; aj.: Language Recognition in iVectors Space. In *Proceedings of Interspeech 2011*, 2011, ISBN 978-1-61839-270-1, ISSN 1990-9772, s. 861–864.

[18] Mun, S.; Park, S.; Han, D.; aj.: Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane. Technická správa, DCASE2017 Challenge, September 2017.

[19] Omar, A. H.: *Audio segmentation and classification*. Diplomová práca, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2005.

[20] Park, S.; Mun, S.; Lee, Y.; aj.: Score Fusion of Classification Systems for Acoustic Scene Classification. Technická správa, DCASE2016 Challenge, September 2016.

[21] Silovský, J.: Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích. Technická správa, Technická univerzita v Liberci, November 2011.

[22] Tax, D. M.; Van Breukelen, M.; Duin, R. P.; aj.: Combining multiple classifiers by averaging or by multiplying? *Pattern recognition*, ročník 33, č. 9, 2000: s. 1475–1485.

[23] Černocký, H.: Zpracování řečových signálu — studijní opora.
URL http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf

[24] Weiping, Z.; Jiantao, Y.; Xiaotao, X.; aj.: Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion. Technická správa, DCASE2017 Challenge, September 2017.

[25] Zabidi, A.; Mansor, W.; Khuan, L. Y.; aj.: Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing Its Applications*, March 2009, s. 204–208, doi:10.1109/CSPA.2009.5069217.