SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Ing. Boris Puterka

Autoreferát dizertačnej práce

Detekcia a klasifikácia emócií v rečovom signáli

na získanie akademického titulu „doktor“ („philosophiae doctor“, v skratke „PhD.“)

v doktorandskom študijnom programe: robotika a kybernetika
v študijnom odbore: kybernetika
forma štúdia: denná
miesto, dátum: Bratislava, 2024

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Dizertačná práca bola vypracovaná na
ústave robotiky a kybernetiky FEI STU v Bratislave.

Predkladateľ: Ing. Boris Puterka
Ústav robotiky a kybernetiky FEI STU
Ilkovičova 2961/3, 841 04 Bratislava

Školiteľ: prof. Ing. Jarmila Pavlovičová, PhD.
Ústav robotiky a kybernetiky FEI STU
Ilkovičova 2961/3, 841 04 Bratislava

Oponenti: doc. Ing. Roman Jarina, PhD.
Katedra multimédií a informačno-komunikačných technológií FEIT
Univerzitná 8215/1, 010 26 Žilina

doc. Ing. Matúš Pleva, PhD.
Katedra elektroniky a multimediálnych telekomunikácií FEI
Němcovej 32, 040 01 Košice

Autoreferát bol rozoslaný: .....
Obhajoba dizertačnej práce sa bude konať dňa: ..... **o** .....**h**
na Fakulte elektrotechniky a informatiky STU, Ilkovičova 3, 841 04 Bratislava

prof. Ing. Vladimír Kutiš, PhD.
dekan FEI STU

SÚHRN

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Štúdijný program:	Robotika a kybernetika
Autor:	Ing. Boris Puterka
Dizertačná práca:	Detekcia a klasifikácia emócií v rečovom signáli
Vedúci záverečnej práce:	prof. Ing. Jarmila Pavlovičová, PhD.
Miesto a rok predloženia práce:	Bratislava, 2024

V rámci tejto dizertačnej práce sme sa venovali výskumu pokročilých metodík rozpoznávania emocionálneho stavu rečníka z rečového prejavu, so zameraním na využitie najnovších technológií v oblasti spracovania signálu a aplikácií umelej inteligencie, predovšetkým hlbokého učenia. Výskum sa venuje tomu, ako charakteristiky rečového signálu a technológie strojového učenia, najmä konvolučné neurónové siete (CNN), môžu významne prispieť k efektívnemu rozpoznávaniu a klasifikácii emocionálnych stavov vyjadrených v reči. Analyzovali sme rôzne techniky a modely, pričom sme venovali osobitnú pozornosť optimalizácii predspracovania signálu, výberu charakteristík a architektúram neurónových sietí s cieľom maximalizácie presnosti rozpoznávania. Súčasťou práce bol veľký počet experimentov, ktoré boli uskutočnené s použitím Berlínskej emočnej rečovej databázy a databázy IEMOCAP. Tieto experimenty demonštrovali, že dôkladná selekcia vlastností rečového signálu a ich kombinácii s vhodne zvolenými parametrami neurónových sietí majú zásadný vplyv na schopnosť systému rozpoznávať emočné stavy s vysokou presnosťou. Vo výsledkoch práce je tiež zdôrazňovaná dôležitosť integrácie rôznych metód predspracovania signálu a adaptácia modelov na špecifické charakteristiky daného typu rečovej emocionálnej expresie. Podstatná časť dizertačnej práce je venovaná aj preskúmaniu možných aplikácií získaných poznatkov v praxi. Diskutujeme o širokej škále možných implementácií od zlepšenia interakcií medzi človekom a počítačom, cez inovácie v oblasti zákazníckej podpory, až po pokročilé diagnostické systémy v medicíne a psychológii. Napokon, práca nastoľuje série otázok a smerov pre ďalší výskum, ktorý by mohol ďalej rozšíriť chápanie a efektivitu rozpoznávania emocionálneho stavu v reči, s dôrazom na vývoj a implementáciu nových algoritmov a technologických postupov.

Kľúčové slová: rozpoznávanie emócií, spektrálna analýza, hlboké učenie

Obsah

1

Úvod

Pri hocijakej emočnej analýze reči je dôležité pochopiť, ako je reč tvorená. Proces tvorby reči sa skladá zo štyroch krokov ktoré sú:

- *myšlienka* - čo a ako chceme povedať;
- *jazyk* - slovné vyjadrenie myšlienky;
- *prozódia* - stres, intonácia, rýchlosť hovorenia, kontext, emócie, atď.;
- *vzduchové vibrácie spôsobené rečovým ústrojenstvom.*

Ak máme fyzický signál, ktorým reč je, tak na jeho analyzovanie je nevyhnutné, aby sme poznali jeho základné charakteristiky. Takéto charakteristiky sú: nestacionarita (stacionárne intervaly v analyzovanom rozsahu od 10ms do 30ms), frekvenčný rozsah, rozloženie energie (mení sa v čase a vo frekvencii v závislosti od foném ako sú napr., samohlásky, frikatíva, atď., a rozdiel môže dosahovať až 50dB) a všetky sú skombinované excitáciou signálu a odozvou hlasového traktu. Keďže rozpoznávanie emócií v rečovom signáli je komplexný problém, existuje viacero postupov a metód pre ich detekciu. V súčasnej dobe najviac prevažujú postupy založené na strojovom učení. V posledných rokoch sa zvýšili aplikácie CNN pri spracovaní reči. Rečový signál je prirodzený spôsob komunikácie s inými ľuďmi. V oblasti automatického rozpoznávania reči (ASR) existuje mnoho aplikácií, napr. identifikácia rečníka a je neoddeliteľnou súčasťou interakcie človek-stroj. Rozpoznávanie emócií reči (SER) má tiež svoje miesto vo výskume a praktických aplikáciách. Systémy, ktoré dokážu efektívne rozpoznať emócie z reči, môžu byť použité v rôznych bezpečnostných aplikáciách napr. bezpečnosť starších obyvateľov žijúcich osamote. Inou formou aplikácií môžu byť zdravotnícki asistenti na detekciu emocionálneho stavu rečníkov. Pre ľudí s poruchami komunikácie, ako je sociálno-emocionálna agnózia alebo dokonca autizmus, sú emócie ťažko pochopiteľné, takže takýto asistent im môže pomôcť pri rehabilitácii. Niekedy je naozaj ťažké, dokonca aj pre zdravých jedincov rozpoznať emócie z reči, takže multimodálne prístupy môžu zohrávať svoju úlohu v úlohách SER.

V práci sa zameriavame preto na analýzu optimálnej dĺžky okna a rečovej vzorky pre extrakciu vlastností z rečového signálu, pretože je neoddeliteľnou súčasťou výpočtu spektrogramu. Cieľom je otestovať rôzne časové dĺžky okna a analyzovanej rečovej vzorky používané pre výpočet krátkodobej Fourierovej transformácie (STFT) s cieľom dosiahnuť najlepšiu mieru rozpoznania.

Problém extrakcie príznakov pre rozpoznávanie emócií nie je ešte úplne vyriešený. S aplikáciou strojového učenia a veľkými komplexnými databázami sa predstavili rôzne koncepty, napríklad založené na autoenkodéroch alebo end-to-end spracovania. Avšak čím sú systémy

komplexnejšie, tým ťažšie je získať cenné informácie (vedomosti) o základných charakteristikách reči relevantných pre rozpoznávanie, pretože sú skryté hlboko v klasifikačných modeloch. Preto sa v tejto práci budeme venovať aj analýze rôznych základných charakteristík reči, časovým, frekvenčným a energetickým aspektom, metódam spracovania signálu a ich nastavení používaných na extrakciu základných funkcií pomocou strojového učenia. Cieľom je nájsť relevantné metódy a nastavenia, odhaliť možné závislosti, ako konkrétne nastavenia ovplyvňujú presnosť, a rozšíriť vedomosti o tom, ktoré základné charakteristiky reči a nastavenia sú dôležité pre systémy rozpoznávania emócií a do akej miery. Také rozsiahle a podrobne vyhodnotené experimenty by mali pomôcť pri návrhu efektívneho systém, ktorý nebude vyžadovať nadmerné ladenie.

Jedným z hlavných dôvodov pre hlbšie preskúmanie podrobností vlastností rečových a zvukových signálov je uvedomenie si skutočnosti, že žiadne dve aplikácie nie sú úplne identické. Systémy rozpoznávania reči, ktoré neúnavne prekladajú hovorené slová do textu, si vyžadujú jemné vnímanie nuáns jazyka a dialektov. Z druhej strany, rozpoznávanie rečníka sa viac zameriava na unikátne fyziologické a behaviorálne charakteristiky, ktoré zanechávajú v hlasivkách jedinečné identifikátory. Rozpoznávanie emócií v reči sa ešte viac vetví do antropológie, dekóduje komplexné jemnosti, ktoré nesú váhu ľudských emócií. Systémy rozkladajú okolité zvuky, aby odhalili vzory alebo anomálie. Variabilita požiadaviek aplikácie si vyžaduje hlboké skúmanie základných fyzikálnych vlastností rečových a zvukových signálov. Preskúmaním frekvenčných pásiem, časových intervalov, presnosti kvantizácie signálu a zložitosti výpočtových modelov môžu vedci začať rozplietať skryté závislosti, ktoré preväzujú vlastnosti signálu s účinnosťou aplikácie.

Dôležitým faktorom pre rozpoznávanie emócií z reči SER je kvalita použitej emočnej rečovej databázy. Jedným z faktorov pre ohodnotenie kvality databázy je stupeň prirodzenosti rečových nahrávok. Pri použití menej kvalitnej databázy hrozí, že dostaneme nesprávne výsledky pri vyhodnocovaní emočného stavu človeka, čo môže mať negatívne následky. Preto sa kladie dôraz na správnu konštrukciu databázy. Väčšina vytvorených databáz nie je dostupná pre verejné použitie. Preto existuje iba obmedzené množstvo databáz, ktoré sú používané výskumnými skupinami v oblasti rozpoznávania emócií. Čo si však môžeme na týchto databázach všimnúť, je to, že klasifikujú rovnaké druhy emócií – hnev, radosť, smútok, potešenie, znudenie, znechutenie a bez emócie. Tieto databázy sú poväčšine zamerané na emócie dospelého človeka. Pre experimenty v tejto práci sme zvolili dve verejne dostupné databázy Berlínska emočná databáza [1] a IEMOCAP [2] vzhľadom na ich veľkosť, kvalitu zvukových nahrávok, ucelenosť dát a ich používanie v súčasnom výskume.

Berlínsku emočnú databázu [1] tvoria nahrávky od 10 hercov, 5 mužov a 5 žien, ktorí predvádzali emócie vyslovovaním 10 nemeckých viet, 5 viet bolo krátkych a 5 dlhých, ktoré sú použiteľne v každodennej komunikácii. Nahrávky boli získavané v nízko odrazovej komore za použitia s vysoko kvalitným nahrávacím zariadením. Databáza obsahuje celkovo 800 nahrávok rozdelených do 7 emočných skupín. Celková databáza bola podrobená testu a ohodnotená vzhľadom na jej rozpoznateľnosť a prirodzenosť emócií. Emócie rozpoznané s úspešnosťou vyššou ako 80% a ohodnotené viac ako 60% poslucháčmi ako prirodzené, boli zaradené do emočnej skupiny výslednej databázy. Výber emócií bol zo štandardnej skupiny, ktorá sa bežne používa vo výskumnej činnosti. Sú to tieto emócie: šťastie, hnev, strach, potešenie, znechutenie, smútok a neutrál. Databáza bola tvorená ľuďmi, ktorí sa prihlásili

na toto nahrávanie prostredníctvom reklamy v novinách. Zámer bol použiť nehercov na vytvorenie jednotlivých nahrávok emócií. Okolo 40 ľudí sa prihlásilo a boli pozvaní na nahrávacie sedenia. Skupina expertov potom vyselektovala skupinu 10 ľudí, ktorí predviedli najkvalitnejšie zahranie emócií. Napokon sa zistilo, že všetci títo 10 ľudia boli absolventmi hereckej školy. Nahrávky boli nahrávané so vzorkovacou frekvenciou 48 kHz a neskôr pod vzorkované na 16 kHz.

Databáza IEMOCAP (angl. Interactive Emotional Dyadic Motion Capture) [2] je multimodálny súbor dyadických konverzácií medzi hercami určený na výskum v oblasti rozpoznávania a analýzy emócií. Nahrávky v databáze sú od 10 hercov (5 mužov a 5 žien), ktorí boli trénovaní na improvizáciu konverzácií založených na 10 vopred definovaných scenároch. Herci boli poučení, aby sa navzájom rozprávali, a zároveň prežívali celý rad emócií, ktoré boli vyvolané rôznymi technikami, ako napríklad požiadaním hercov, aby si spomenuli na osobné zážitky, použili rekvizity alebo zahrali scenáre. Súbor obsahuje približne 12 hodín audio a video nahrávok v celkovom počte 5528, rozdelených do 5 emočných tried (hnev, šťastie, smútok, neutrál, frustrácia). Konverzácie boli nahrávané v štúdiovom prostredí s použitím vysoko kvalitných nástrojov. Vzorkovacia frekvencia audio nahrávok je 16kHz.

1.1 Cieľe dizertačnej práce

Z rozsiahleho štúdia súčasného stavu riešenej problematiky vyplynulo, že rozpoznávanie emócií z rečového signálu je rýchlo rastúca oblasť a počet publikovaných riešení každým rokom rastie. Veľký dôraz sa kladenie najmä na výber klasifikátora do emočných tried a nie až tak na predspracovanie rečového signálu. Existujú stále nezodpovedané otázky v tejto oblasti a z tohto dôvodu si ako ciele volíme:

- Výber optimálnych parametrov a architektúry hlbokej neurónovej siete na rozpoznávanie emócií z rečového signálu. Očakávame, že tieto optimálne parametre a nastavenia vyplynú z hlbokej časovo frekvenčnej analýzy účinnosti príznakov z hľadiska:
  - frekvenčného rozsahu rečového signálu – využitie iných frekvenčných rozsahov ako pri rozpoznávaní reči
  - veľkosti a typu použitej banky filtrov – predpokladáme využitie gamma tónových filtrov pre zvýšenie úspešnosti klasifikácie
  - použitia rôznych oknových funkcií s rôznymi dĺžkami a veľkosťou prekryvu – výber optimálnych parametrov (veľkosť, prekryv, sklon) nastavenia oknových funkcií
  - typu použitých príznakov – koeficientov spektrogramu, MFCC koeficientov, LPC
  - veľkosti použitého časového úseku rečového signálu ako vstupu do klasifikátora
- Vytvorenie uceleného systému na rozpoznávanie emócií, ktorý bude tvoriť:

  - blok predspracovania rečového signálu s optimálnymi parametrami získanými časovo frekvenčnou analýzou
  - vhodná architektúra pre rozpoznávanie emócií, na základe porovnania výkonnosti s existujúcimi archichektúrami

- Vyhodnotiť účinnosť vytvoreného systému na zvolenej databáze (Berlínska emočná databáza) a IEMOCAP
- Vyhodnotiť účinnosť vytvoreného systému na ostatné aplikácie rečového signálu ako sú rozpoznávanie rečníka, rozpoznávanie reči a rozpoznávanie zvukových udalostí

2

Časová modifikácia príznakov rečového signálu na klasifikáciu emócií

V tejto kapitole prezentujeme analýzu časovej dĺžky okna a dĺžky rečového segmentu na rozpoznávanie emócií reči. Vykonali sme naše testy na spektrogramoch vypočítaných z rôznych okien rôznej dĺžky v čase a trénovali sme našu architektúru konvolučnej neurónovej siete s dvoma konvolučnými vrstvami a jednou plne pripojenou vrstvou. Cieľom bolo zistiť, aké rozlíšenie spektrogramov, či už časové alebo frekvenčné, je lepšie na úspešné zatriedenie rečového signálu do siedmich emočných tried. Podľa našich výsledkov je správna voľba dĺžky okna dôležitá pre rozpoznávanie emócií reči. Presnosť závisí od dĺžky okna a presahu po sebe nasledujúcich rámcov. Výsledky ukazujú, že predspracovanie rečového signálu v časovej oblasti hrá dôležitú úlohu pre systém SER, a to aj pri použití CNN ako klasifikátora.

2.1 Materiály a metódy

Vlastnosti reči sa s časom výrazne menia. Je to jej prirodzená a pekná vlastnosť, ale klasické použitie diskrétnej Fourierovej transformácie (DFT), nie je možné. Väčšina foném má rečové vlastnosti nemenné v krátkych časových úsekoch (približne 5-100ms, pre praktické aplikácie sa používa rozsah 10-30ms). Preto sa aplikácia časovo krátkeho okna na signál javí praxi ako relatívne úspešná technika. Väčšina metód spracovania reči preto berie ako vstup krátky časový úsek rečového signálu. Tento krátky časový úsek signálu nazývame rámec. Oknovanie je metóda, ktorá nejaký signál vynásobí oknovou funkciou konečnej dĺžky. Výsledkom je signál s konečnou dĺžkou. ktorý je modifikáciou vstupného signálu. Tvar okna v oblasti spracovania reči nehrá zvlášť veľkú úlohu, no môže pri niektorých aplikáciách jemne zlepšiť výkonnosť daného systému. Môžeme povedať, že oknová funkcia sa používa na lokalizáciu rečového signálu v čase. Vo všeobecnosti je na svojich okrajoch zúžená, aby sa zabránilo neprirodzeným prerušeniam v rečovom segmente. Pre naše potreby si zadefinujeme Hammingovu, Bartlettovu a pravouhlú oknovú funkciu.

Hammingovo okno je optimalizovaná verzia Hannovho okna. Koeficienty tejto oknovej funkcie sú optimalizované, aby sa dosiahla minimálna úroveň prvého bočného laloku. Podobne ako Hannovo okno, Hammingovo okno je jedna perióda vyvýšeného kosínusu, avšak jeho záporné hodnoty presahujú nad nulovú os. Hammingová oknová funkcia je nespojitá v amplitúde na jej koncoch (skoková zmena amplitúdy z približne 0.08 na 0). Hammingovo okno definujeme v čase ako:

$$f(t) = \begin{cases} 0.54 + 0.46cos\left(\frac{\pi t}{\tau}\right), & \text{pre } |t| \leq \tau \\ 0, & \text{inde} \end{cases} \tag{2.1}$$

s odpovedajúcou FT:

$$F(\Omega) = 1.08\left(\frac{sin(\Omega\tau)}{\Omega}\right) + 0.46\left[\frac{sin\left((\Omega + \pi/\tau)\tau\right)}{\Omega + \pi/\tau} + \frac{sin\left((\Omega - \pi/\tau)\tau\right)}{\Omega - \pi/\tau}\right] \tag{2.2}$$

kde Ω je z intervalu $(-\infty; \infty)$. Vzťah medzi (2.1) a (2.2) vieme zapísať ako:

$$\left[0.54 + 0.46cos\left(\frac{\pi t}{\tau}\right)\right]_{|t|\leq\tau} \overset{F}{\leftrightarrow} \frac{(1.08\pi^2 - 0.16\Omega^2\tau^2)\, sin(\Omega\tau)}{\Omega(\pi^2 - \Omega^2\tau^2)} \tag{2.3}$$

Asymptotický útlm Hammingovho okna je $0.16\Omega^{-1}$. Hammingovo okno sa dlhodobo používalo pri spracovaní telefónneho signálu kedy 8 bitové kodeky boli štandardom desiatky rokov. Pre kvalitnejšie spracovanie audio signálu môžu byť vyžadované kvalitnejšie okná, najmä ak tieto okná fungujú ako dolno priepustné filtre.

Bartlettovo (trojuholníkové) okno je výsledok konvolúcie dvoch pravouhlých okien s polovičnou dĺžkou, s prihliadnutím na veľkosť magnitúdy týchto dvoch pravouhlých okien. Bartlettovo okno vieme v časovej oblasti vyjadriť ako:

$$f(t) = \begin{cases} 1 - \frac{|t|}{\tau}, & \text{pre} |t| \leq \tau \\ 0, & \text{inde} \end{cases} \tag{2.4}$$

s odpovedajúcou FT:

$$F(\Omega) = \tau \left[\frac{sin\left(\frac{\Omega\tau}{2}\right)}{\left(\frac{\Omega\tau}{2}\right)} \right]^2 \tag{2.5}$$

kde Ω je z intervalu $(-\infty; \infty)$.

Vlastnosti Bartlettovho okna sú:

- je výsledkom konvolúcie dvoch pravouhlých okien dĺžky $(\tau - 1)/2$,
- hlavný lalok je dva krát širší ako pri pravouhlom okne dĺžky τ,
- často sa implicitne používa na vzorkovanie korelácií konečných javov.

Pravouhlé okno, tiež nazývané rovnomerné okno alebo „box car“ kvôli jeho tvaru, definujeme ako:

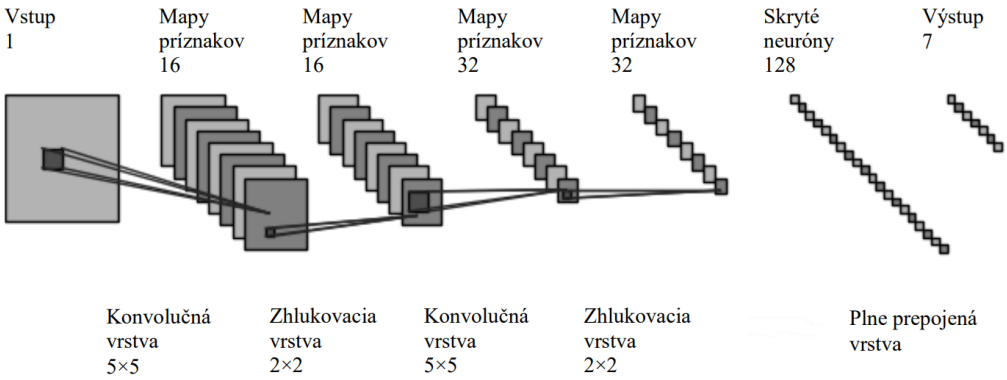
$$f(t) = \begin{cases} 1 & \text{pre} |t| \leq \tau \\ 0, & \text{inde} \end{cases} \tag{2.6}$$

kde τ je časový interval trvania jedného okna. Odpovedajúcu FT vieme vyjadriť ako:

$$F(\Omega) = \frac{2\tau sin(\Omega\tau)}{\Omega\tau} \tag{2.7}$$

kde Ω je z intervalu $(-\infty; \infty)$.

V nasledujúcich experimentoch sme použili Berlínsku databázu emócií. Ako vstup do nášho modelu sme rozdelili spektrogramy do blokov s požadovanou dĺžkou v čase pomocou posunu o 10 rámcov. Každá sekvencia bola označená triedou danej emócie, pretože celá nahrávka bola označená danou emóciou. Použili sme CNN s dvoma konvolučnými vrstvami, ktorej architektúra je zobrazená na obr. 2.1. Dávka na na trénovanie CNN bola nastavená na veľkosť 256 FV a parameter učenia pri trénovaní bol 0.01. Výkon architektúry bol vyhodnotený 10 násobnou vzájomnou validáciou. 80% údajov sa použilo na trénovanie, zvyšných 20% na testovanie. 20% trénovacích dát sme použili na validáciu počas trénovania.



Obr. 2.1: Použitá architektúra CNN pre experimenty.

2.2 Výsledky a dikusia

Ako vstup do nášho modelu sme rozdelili spektrogramy do blokov s požadovanou dĺžkou v čase pomocou posunu o 10 rámcov. Každá sekvencia bola označená triedou danej emócie, pretože celá nahrávka bola označená danou emóciou. Použili sme CNN s dvoma konvolučnými vrstvami, ktorej architektúra je zobrazená na obr. 2.1. Dávka na na trénovanie CNN bola nastavená na veľkosť 256 FV a parameter učenia pri trénovaní bol 0.01. Výkon architektúry bol vyhodnotený 10 násobnou vzájomnou validáciou. 80% údajov sa použilo na trénovanie, zvyšných 20% na testovanie. 20% trénovacích dát sme použili na validáciu počas trénovania.

V prvom experimente sme hľadali optimálnu dĺžku rámca rečového signálu pre rozpoznávanie emócií z hľadiska presnosni klasifikácie. Použili sme Hammingovo okno pre STFT na nespracovaný rečový signál segmentovaný do 256 dlhých rámcov s 50% prekrytím susedných rámcov. Aby sme našli vhodné trvanie rečového signálu na úspešné rozpoznanie emócií, trénovali sme a testovali našu sieť so segmentami rečového signálu dĺžky 0.25 s, 0.5 s, 0.75 s, 1 s, 1.25 s a 1.5 s. So vzorkovacou frekvenciou 16 kHz, oknom dlhým 256 vzoriek a 50% prekrytím predstavuje 1 snímka funkcie 16 ms dlhý rečový signál. Výsledky ukazujú, že na rozpoznanie emócií z rečového signálu reprezentovaného spektrogramom je vhodné skúmať 1.5 s dlhý interval rečového signálu. V priemere 1.5s dlhý analyzovaný rečový signál bol najlepší spomedzi testovaných dĺžok sekvencií. Môžeme pozorovať, že zvýšenie dĺžky sekvencií malo obrovský vplyv na výkon nášho modelu. Model trénovaný na dlhších sekvenciách fungoval lepšie ako trénovaný na kratších sekvenciách. Môžeme tiež pozorovať, že medzi 1.25 a 1.5 s dlhými sekvenciami je len malé zvýšenie presnosti. Pri sekvenciách dlhších ako 1.5 s môže nastať problém, pretože niektoré výroky v databáze nie sú dlhšie ako 1.5 s. Spôsobilo by to redukciu súboru údajov a nemohli by sme správne porovnávať naše výsledky. Stojí tiež za zmienku, že počet trénovacích a testovacích dát klesal s rastúcou dátovou dimenziou v priebehu času. Znamená to, že s dlhším časom trvania vstupu signálu zmenšujeme množinu údajov.

Tab. 2.1: Závislosť medzi dĺžkou analyzovaného rámca a veľkosti datasetu.

Analyzovaná dĺžka [s]	Trénovacie	Testovacie	Validačné	Spolu
0.25	9750	3048	2438	15236
0.5	8781	2745	2196	13722
0.75	7784	2433	1947	12164
1	6812	2130	1704	10646
1.25	5840	1826	1461	9127
1.5	4788	1497	1198	7483

Tab. 2.1 ukazuje závislosť medzi veľkosťou súboru údajov a dĺžkou analyzovanej sekvencie. Môžeme pozorovať, že pre sekvencie s dĺžkou 0.25s je viac ako dvakrát väčší súbor údajov ako pre sekvencie s dĺžkou 1.5 s. Systém bol hodnotený na sekvenčnej úrovni. Na základe praktického využitia môžeme využiť klasifikáciu na úrovni výpovede. Z každého výroku získame niekoľko sekvencií a vyhodnotíme ich pomocou nášho natrénovaného modelu. Výstupná trieda (kategória) pre výrok bude tá, ktorá sa najčastejšie vyskytuje v klasifikácii na úrovni sekvencie (väčšinové kritériá). S týmto prístupom môžeme dosiahnuť lepšie výsledky v rozpoznávaní emócií.

Naledujúce experimenty súviseli s typom použitého okna pre výpočet spektogramov pomocou STFT a ich prekryvmi. Najprv sme testovali vzťah medzi presnosťou a použitou segmentáciou, t. j. dĺžkami rámcov a presahmi. Dĺžky snímok riadia kompromis medzi časovým a frekvenčným rozlíšením, zatiaľ čo presahy ovlyvňujú časové rozlíšenie, množstvo dostupných tréningových dát a ich redundanciu. V oboch experimentoch je viditeľný trend veľkosti prekrytia, ktorý je skôr nezávislý od dĺžky rámca, t.j. vyššie prekrytia vedú k lepšej presnosti. To naznačuje, že vyšší počet snímok (vyššie prekrytie) v spektrogramoch, ktoré zvyšujú jeho časové rozlíšenie a redundancia, je prínosom pre klasifikáciu CNN. Na druhej strane testované dĺžky snímok neboli dominantnými faktormi pre presnosť, keďže všetky dĺžky dokázali zaznamenať rovnaký výkon na 10% úrovni významnosti, t. j. časovo-frekvenčné rozlíšenie dané dĺžkami snímok v testovanom rozsahu je vyhovujúce pre SER. Najlepšie výsledky pre 4 kHz frekvenčný rozsah dosiahli nastavenia s dĺžkou okna 10ms a 25% prekryvom (64.97 $\pm$ 5.1%), s dĺžkou okna 20ms a 50% prekryvom (64.89 $\pm$ 5.05%) a s dĺžkou okna 30ms a 25% prekryvom (65.9 $\pm$ 4.32%). Pre 8 kHz frekvenčný rozsah dosiahli nalepšie výsledky nastavenia s dĺžkou okna 10ms a 25% prekryvom (64.5 $\pm$ 4.32%), s dĺžkou okna 10ms a 50% prekryvom (65.21 $\pm$ 4.7%), s dĺžkou okna 20ms a 25% prekryvom (65.42 $\pm$ 3.69%) a s dĺžkou okna 30ms a 50% prekryvom (64.81 $\pm$ 5.2%). V oboch experimentoch je viditeľný trend veľkosti prekrytia, ktorý je skôr nezávislý od dĺžky rámca, t.j. vyššie prekrytia vedú k lepšej presnosti. To naznačuje, že vyšší počet snímok (vyššie prekrytie) v spektrogramoch, ktoré zvyšujú jeho časové rozlíšenie a redundancia, je prínosom pre klasifikáciu CNN. Na druhej strane testované dĺžky snímok neboli dominantnými faktormi pre presnosť, keďže všetky dĺžky dokázali zaznamenať rovnaký výkon na 10% úrovni významnosti, t. j. časovo-frekvenčné rozlíšenie dané dĺžkami snímok v testovanom rozsahu je vyhovujúce pre SER.

Ďalší experiment bral do úvahy dĺžky a posuny (doplnok k prekrytiu) rečových blokov, ktoré sa používajú na vytváranie spektrogramov, t.j. FV pre klasifikáciu CNN. Čím dlhší je blok,

tým viac informácií je poskytnutých, avšak extrémne dlhá časť reči môže potenciálne pokryť zmeny emócií a znížiť množstvo dostupných tréningových dát. Keď uvažujeme o posune susedných blokov, čím väčší je posun, tým menej blokov je dostupných, ale susedné bloky sú menej nadbytočné. Boli testované vplyvy dĺžok spektrogramov (0.5, 0.8, 1 a 1.2 s) v kombinácii s ich posunmi (0.1, 0.2 a 0.4 s) pre hornú frekvenciu 8 kHz. Horná hranica 1.2 s pre dĺžky spektrogramov bola daná najkratším záznamom v databáze. Najlepšie výsledky dosiahli experimenty pre spektrogramy s dĺžkou 1s a 0.1s posunom (64.95 ± 3.7%), s dĺžkou 1.2s a 0.1s posunom (65.98 ± 4.05%) a s dĺžkou 1.2s a 0.2s posunom (64.7 ± 4.1%). Z výsledkov je zrejme, že čím dlhšie bloky (spektrogramy v čase), tým lepšie presnosti, avšak nad 1s sa zdalo, že nárast sa začal nasyťovať. Veľkosť posunov medzi spektrogramami pôsobila vo všetkých prípadoch opačne, t.j. čím dlhší posun, tým horšia presnosť. To znamená, že množstvo tréningových dát prevažovalo nad zvýšenou redundanciou medzi FV.

Účinok aplikovaných okien, t. j. Boxcar, Bartlettove a Hammingove v kombinácii s rôznymi veľkosťami rámca reči, sa zvažoval v poslednej sérii experimentov zahŕňajúcich segmentáciu. Vykonali sme experimenty pre vyššie uvedené okná s dĺžkami: 10, 20 a 30 ms a maximálnou frekvenciou 4 kHz a 8kHz Pre maximálnu frekvenciu 4kHz fungovali okná pomerne rovnako (priemerné presnosti 63–66%), avšak v prípade maximálnej frekvencie 8 kHz spôsobila charakteristika pomalého poklesu frekvencie Boxcar štatisticky významný pokles presnosti bez ohľadu na dĺžku rámca (pokles priemerne takmer 2%). To môže naznačovať, že pri zvažovaní vyšších frekvenčných rozsahov je aliasing s dlhým chvostom pre aplikácie SER horší.

3

Časovo frekvenčná analýza rečového signálu pre výber optimálnych príznakov na klasifikáciu emócií

V tejto kapitole prezentujeme analýzu časovej dĺžky okna a dĺžky rečového segmentu na rozpoznávanie emócií reči. Vykonali sme naše testy na spektrogramoch vypočítaných z rôznych okien rôznej dĺžky v čase a trénovali sme našu architektúru konvolučnej neurónovej siete s dvoma konvolučnými vrstvami a jednou plne pripojenou vrstvou. Cieľom bolo zistiť, aké rozlíšenie spektrogramov, či už časové alebo frekvenčné, je lepšie na úspešné zatriedenie rečového signálu do siedmich emočných tried. Podľa našich výsledkov je správna voľba dĺžky okna dôležitá pre rozpoznávanie emócií reči. Presnosť závisí od dĺžky okna a presahu po sebe nasledujúcich rámcov. Výsledky ukazujú, že predspracovanie rečového signálu v časovej oblasti hrá dôležitú úlohu pre systém SER, a to aj pri použití CNN ako klasifikátora.

3.1 Materiály a metódy

Ako už bolo uvedené, je dôležité analyzovať reč v čase a frekvencii súčasne. Pre analýzu časovo premenlivých nestacionárnych signálov ako je napr. rečový signál, je vhodnejšie aplikovať FT na krátke časové úseky signálu ako na celý signál. Táto operácia sa nazýva krátkodobá Fourierova transformácia (ďalej len STFT) a definujeme ju ako:

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t - \tau) e^{-j\omega t} dt \tag{3.1}$$

kde $x(t)$ je analyzovaný signál a $w(t)$ je oknová funkcia so stredom v t. Oknová funkcia oreže signál v okolí t a FT bude teda lokálnym odhadom v okolí tejto časovej inštancie. Pre potreby digitálneho spracovania signálov definujeme STFT diskrétnu v čase ako:

$$X(n, \omega) = \sum_{m=-\infty}^{\infty} x[m] w[m - n] e^{-j\omega m} \tag{3.2}$$

kde $x[m]$ je navzorkovaný signál a $w[m]$ je použitá oknová funkcia. Treba si však uvedomiť, že v tomto prípade je m diskrétna premenná na rozdiel od ω, ktorá zostáva spojitá.

Štandardný spôsob počítania STFT je s použitím kladného okna $w(t)$ nejakého tvaru, ktoré je symetrické okolo nuly s energiou $\int_{-\infty}^{\infty} |w(t)|^2 \, dt = 1$. Podobne ako pre klasickú FT a jej spektrum, pre STFT vieme definovať spektrogram ako:

$$S_x(t, w) = |X(t, \omega)|^2 \tag{3.3}$$

ktorý je veľmi často používaný pre analýzu časovo premenných a nestacionárnych signálov.

Dôležitým krokom v ASR a SER systémoch je extrakcia vhodných príznakov, ktoré zvýraznia sledovaný obsah (reč, emócie a pod.) a potlačia nežiadúci obsah (napr. šum). Melove frekvenčné kepstrálne koeficienty (ďalej len MFCC) sú dnes asi najčastejšia reprezentácia signálu používaná pri rozpoznávaní reči. Dôvodom, prečo sa MFCC využívajú v SER, je to, že MFCC upravujú signál tak, ako ho približne vníma ľudské ucho. Ľudské ucho je špecifické tým, že nevníma jednotlivé frekvencie zvukového signálu lineárne – v nižších frekvenčných pásmach je citlivejšie, zatiaľ čo vo vyšších menej. Postup výpočtu MFCC môžeme rozdeliť do piatich krokov:

1. **rozdelenie vstupného signálu do krátkych časových rámcov**
2. **výpočet výkonového spektra**
3. **výpočet banky filtrov v Mel mierke**
4. **aplikovanie banky filtrov na výkonové spektrum**
5. **výpočet diskrétnej kosínusovej transformácie**

Banky filtrov (BF), zvlášť sluchovo orientované, sú veľmi populárne nástroje na extrakciu relevantnej informácie vokálneho traktu (spektrálna obálka). Ich dizajn odzrkadľuje rôzne fenomény vnímania zvuku a sluchové modely. Psychoakustiká pokrývajú viac aspektov, preto boli odvodené rôzne škály a súvisiace BF v závislosti od určitého fenoménu, na ktorý sa zameriavame. Medzi tieto škály patrí napr. Melova, Barkova a gammatónová škála.

Mel frekvenčná BF pozostáva z trojuholníkových pásmových filtrov, ktoré napodobňujú ľudský sluchový systém. BF je založená na nelineárnej frekvenčnej škále - Mel. Matematicky tento vzťah vieme vyjadriť ako:

$$f_{mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right) \tag{3.4}$$

kde f_{mel} je frekvencia v Meloch a f je lineárna frekvencia v Hertzoch. Vyššie spomenuté trojuholníkové filtre sa prekrývajú tak, že spodná hranica jedného filtra je umiestnená v strednej frekvencii predchádzajúceho filtra a horná naopak v strednej frekvencii nasledujúceho filtra. Maximálna odozva filtra, teda vrchol trojuholníkového filtra, sa nachádza na strednej frekvencii filtra a je normalizovaná na jednotkovú hodnotu. Stredné frekvencie série filtrov sú rovnomerne rozložené pozdĺž frekvenčnej stupnice Mel. Výstup m-tého filtra X_m môžeme vyjadriť nasledujúcim vzťahom:

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \, |H_m[k]| \,; \qquad 1 \leq m \leq M \tag{3.5}$$

kde $|H_m[k]|$ je frekvenčná magnitúda m-tého filtra, $S[k]$ je spektrum N-bodovej FFT okno-vaného rámca.

Barkova BF je navrhnutá tak, aby napodobňovala spôsob akým ľudské ucho spracováva zvuk a to rozdelením frekvenčného spektra do kritických pásiem. Každé pásmo má svoj pásmový filter, ktorý je rozmiestnený podľa Barkovej škály. Barkove BF sú založené na psychoakustickom výskume ktorý ukázal, že ľudské ucho je viac citlivé na frekvenčné zmeny v určitých bodoch spektra. To znamená, že príznaky extrahované Barkovou BF môžu byť obzvlášť dôležité pre rozpoznávanie emócií, pretože úzko súvisia so spôsobom, akým sa emócie prenášajú prostredníctvom zmien výšky tónu, rytmu a iných akustických prvkov.

$$f_{bark} = 13 \arctan\left(\frac{0.76f}{1000}\right) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{3.6}$$

Lichobežníkový tvar filtrov Barkovej BF je definovaný ako:

$$H_k(f) = \begin{cases} \frac{f-(f_c-B/2)}{B/2}, & f_c - B/2 \leq f < f_c \\ 1, & f_c \leq f < f_c + B/2 \\ \frac{f_c+B/2-f}{B/2}, & f_c + B/2 \leq f < f_c + B \\ 0, & \text{inde} \end{cases} \tag{3.7}$$

kde k je index filtra, f_c je stredová frekvencia a B je šírka pásma. Výstup filtra je definovaný

ako suma každého bodu FFT rečového výkonového spektra a váh filtrov.

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \, |H_m[k]| \, ; \qquad 2 \leq m \leq M-1 \tag{3.8}$$

kde $|S[k]|^2$ je spektrum N-bodovej FFT oknovaného rámca a $|H_m[k]|$ sú váhy filtrov m-tého Barkovho filtra.

Gammatonové filtre sú navrhnuté tak, aby napodobňovali správanie kochlei (taktiež známy ako slimák) v ľudskom uchu, o ktorom je známe, že je citlivejší na určité frekvenčné rozsahy. Výsledkom je, že gammatónová banka filtrov je efektívnejšia pri zachytávaní jemných spektrálnych detailov vo zvukových signáloch, vďaka čomu je efektívnejšia pre riešenie problémov, ktoré si vyžadujú presnú frekvenčnú analýzu, napríklad rozpoznávanie emócií reči. Impulzová charakteristika jednotlivých filtrov v banke je definovaná ako:

$$g(t) = at^{n-1} e^{-2\pi bt} \cos(2\pi f_{ct} + \varphi); \qquad t > 0 \tag{3.9}$$

kde n je stupeň filtra, φ je fázový posun, f_c je stredová frekvencia a b je šírka pásma. Ľudskú kochleu môžeme modelovať pomocou pravouhlých sluchových filtrov, ktorých šírky pásma nazývame ERB (ekvivalentná pravouhlá šírka pásma). Výstup filtra je definovaný ako suma rečového výkonového spektra a váhovaných FFT koeficientov.

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \, |H_m[k]| \, ; \qquad 1 \leq m \leq M \tag{3.10}$$

kde $|S[k]|^2$ je spektrum N-bodovej FFT oknovaného rámca a $|H_m[k]|$ je frekvenčná odozva m-tého gammatónového filtra.

Najrozšírenejším modelom na produkciu reči je lineárny model s časovým variantom využívajúci koeficienty lineárnej predikcie (LPC). Model kombinuje excitačný signál s parametrami vokálneho traktu na vytvorenie rečového signálu. Hlasový trakt je reprezentovaný celopólovým filtrom, ktorý priamo modeluje formantové frekvencie, zatiaľ čo modelovanie núl (nosových foném) je zvyčajne neefektívne. V dôsledku výpočtového procesu (priemerná štvorcová chyba) môžu byť vyššie formanty s nižšou energiou zanedbané. Preto je bežné vykonávať vysokopriepustnú filtráciu (preddôraz), ktorá odstraňuje efekt dolnopriepustného filtra zvuku šíriaceho sa do otvoreného priestoru. S parametrami modelu a rečovým signálom je možné odhadnúť budiaci signál pomocou inverznej filtrácie. Budiace signály väčšinou nesú prozodické informácie, napr. základnú frekvenciu, zisk, a preto môžu byť zaujímavé aj pre systémy SER. Časovo diskrétny časovo premenný lineárny systém slúži na simuláciu tunelu hlasového ústrojenstva. Generátor budenia simuluje rôzne generovania zvukov hlasového ústrojenstva. Jednotlivé vzorky rečového signálu sú potom výstupom z časovo premenného lineárneho signálu. Krátkodobá frekvenčná odpoveď lineárneho systému tvaruje hlasové ústrojenstvo. Prenosová funkcia takéhoto lineárneho systému diskrétneho v čase môže byť vyjadrená ako:

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-1}} = \frac{b_0 \prod_{k=1}^{M} (1 - d_k z^{-1})}{\prod_{k=1}^{N} (1 - c_k z^{-1})} \tag{3.11}$$

kde a_k a b_k sú parametre hlasového ústrojenstva, c_k sú póly prenosovej funkcie $H(z)$, z ktorých niektoré ležia v blízkosti jednotkovej kružnice a tak vytvárajú rezonancie na modelovanie formantov a d_k sú nuly prenosovej funkcie, ktoré je možné použiť na modelovanie nasály a frikatíva. Budiaci signál, vytváraný generátorom budenia, sa líši pre znelé a neznelé zvuky. Pre znelé sa skladá z určitého sledu impulzov, ktoré sú generované s periódou základného hlasivkového tónu, a pre neznelé je tvorený čisto šumom. Následne je ešte tento signál zosilnený a potom až privedený do lineárneho systému ktorý modeluje reč.

Fázové spektrum sa pri analýze reči z mnohých dôvodov veľmi často nepoužíva. Konštrukcia fázových spektrogramov predstavuje dva hlavné problémy: fázovú diskontinuitu pozdĺž frekvenčných a časových osí. Na potlačenie diskontinuity pozdĺž frekvenčnej osi sa môže použiť rozbalenie fázy, čo je v diskrétnom prípade nejednoznačné. Aby to bolo robustnejšie, môže sa použiť vzorkovanie na zvýšenie frekvencie faktorom M. Ďalší zdroj fázovej variability pramení zo spracovania časovo posunutých rámcov. Cieľom je kompenzovať fázy periodických signálov tak, aby boli konštantné bez ohľadu na polohu rámca, t.j. odstrániť fázové modifikácie spôsobené spracovaním časovo posunutých rámcov, a zároveň sledovať ich prirodzený vývoj obsiahnutý v skutočných rečových signáloch. Priamym riešením je eliminácia dobre známej vlastnosti kruhového časového posunu (DFT) pre každý rámec (m) takto:

$$\varphi'_m(k) = \left[\varphi_m(k) - \frac{2\pi k S m}{N}\right]_{2\pi} \tag{3.12}$$

kde S je posun rámca, k je frekvenčný index, N je dĺžka rámca, φ_m a φ'_m sú pôvodné a kompenzované fázy a 2π je modulo operácia. Zistili sme však, že túto štandardnú metódu nemožno priamo použiť v prípade navzorkovaných frekvenčných zložiek (podľa faktora M, t.j. budú NM frekvenčné zložky vypočítané z rámca N vzoriek). V takom prípade sme odvodili frekvenčne závislé kompenzačné členy, ktoré sa majú aplikovať na pôvodný signál $x(n)$ pred DFT. Kompenzovaný signál $x_{comp\,k}$ sa vypočíta ako v 3.13.

$$\begin{aligned} x_{comp\,k} &= x(n)e^{\frac{-j2\pi kS}{NM}} \qquad n\varepsilon\{S \ldots N-1\} \\ x_{comp\,k} &= x(n)e^{\frac{-j2\pi k(S-N)}{NM}} \qquad n\varepsilon\{0 \ldots S-1\} \end{aligned} \tag{3.13}$$

Potom sa vzorkovaný DFT aplikuje na $x_{comp\,k}(n)$ dĺžky N, pričom sa vytvárajú frekvenčné zložky NM. To zabezpečuje rovnaké fázy v rámci časovo posunutých rámcov pokrývajúcich všetky frekvenčné zložky MN k. Treba poznamenať, že $x_{comp\,k}(n)$ závisí od frekvencie, a preto sa musí vypočítať pre každú frekvenciu k. Ďalšou možnosťou je rozbaliť fázy oddelene pre každý rámec a pre pevnú frekvenciu, napr. k_{ref} nastaviť konštantnú fázu (φ_{ref}) pre všetky snímky. Nakoniec sa fázy zvyšných frekvenčných zložiek (k) upravia ako v 3.14.

$$\varphi'_m(k) = \varphi_m(k) + k\left(\frac{\varphi_{ref} - \varphi_m(k)}{k_{ref}}\right) \tag{3.14}$$

3.2 Výsledky a diskusia

V prvom súbore experimentov sme používali spektrogramy založené na magnitúde pre minimálne (0^+, 150 a 300 Hz) a maximálne (4 a 8 kHz) frekvenčné limity a taktiež účinok preemfázy. Z výsledkov je možné, preemfáza skutočne zhoršila výsledky pre všetky testované rozsahy. Pozoruhodným pozorovaním bol pozitívny vplyv nízkych frekvencií pod 300 Hz, čo je v súlade so zlyhaním preemfázy. Okrem toho z výsledkov experimentov vyplýva, že zvýšenie frekvenčného rozsahu nad 4 kHz má malý dodatočný prínos, najmä v kombinácii s preemfázou. Najlepšie výsledky boli dosiahnuté pre nastavenia frekvenčné rozsahu 0–4 kHz bez preemfázy (71.1 ± 3.91%), 0–8 kHz bez preemfázy (72.4 ± 3.21%), 0–8 kHz s preemfázou (68.9 ± 4.11%) a 0–4 kHz bez preemfázou (67.7 ± 5.3%).

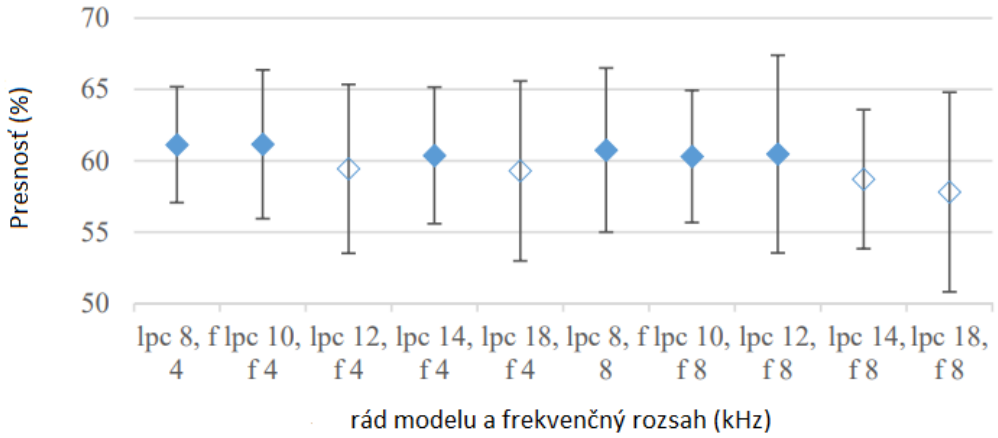
Následne sa skúmalo nasadenie nelineárnej frekvenčnej škály pomocou psychoakustickej Mel škály takisto bez a s použitím preemfázy. Rovnakým spôsobom ako v predchádzajúcom prípade preemfáza mierne zhoršila presnosť pre všetky testované nastavenia experimentov. Pozoruhodným pozorovaním je výrazné zlepšenie dosiahnuté zavedením vyššieho horného frekvenčného limitu, t.j. 8 kHz namiesto 4 kHz v oboch prípadoch. Je to však spôsobené tým, že máme viac vzoriek, ktoré (kvôli nelinearite Mel škály) poskytovali vačšie rozlíšenie pre nižšie frekvencie. Ak je však lineárny frekvenčný rozsah správne nastavený, mierne (v priemere) prevyšuje Melovu škálu. Najlepšie výsledky boli dosiahnuté pre nastavenia frekvenčné rozsahu 0–8 kHz bez preemfázy (71.5 ± 3.7%), 0.150–8 kHz bez preemfázy (70.25 ± 4.52%), 0.3–8 kHz bez preemfázy (71.75 ± 3.21%), 0-8 kHz s preemfázou (70.53 ± 2.4%), 0.150–8 kHz s preemfázou (69.79 ± 4.67%) a 0.3–8 kHz s preemfázou (69.7 ± 2.54%).

Výsledky sa veľmi líšili v závislosti od metód manipulácie s magnitúdou v experimentoch s frekvenčným rozsahom. Preto boli testované magnitúda, výkon, logaritmická magnitúda a logaritmický výkon aplikované na spektrogramy s rôznymi frekvenčnými rozsahmi. Bolo možné vidieť, modifikácia spektra je dôležitým faktorom v závislosti od frekvenčného rozsahu, ktorý je potrebné vziať do úvahy. Magnitúdy dosiahli výrazne lepšie výsledky bez ohľadu na nasadenie logaritmu. Na druhej strane výkony a logaritmické výkony boli nižšie, najmä ak sa zvažovali vyššie frekvencie (8 kHz). Najlepšie výsledky sme zaznamenali práve pre magnitúdu na 8 kHz frekvečnom rozsahu (72.3 ± 3.51%) a logaritmickú magnitúdu na 8 kHz frekvečnom rozsahu (73.12 ± 3.74%).

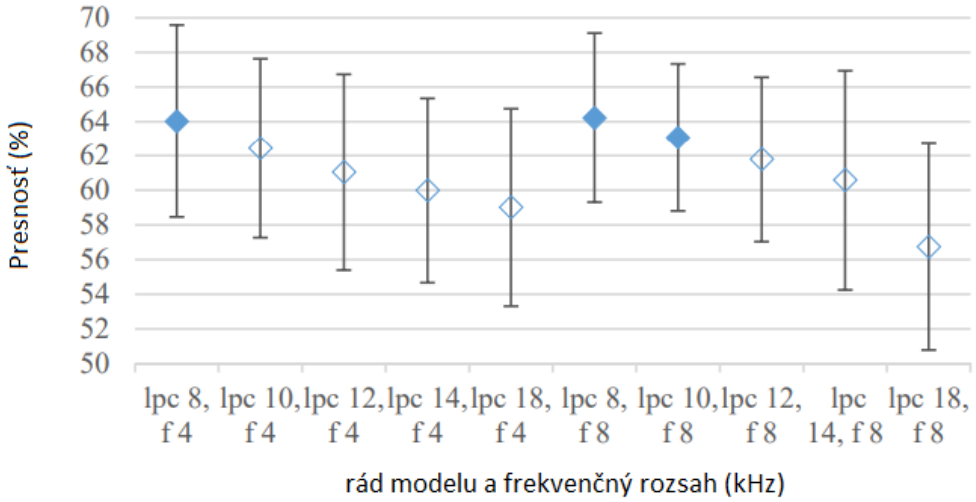
Následne boli skonštruované spektrogramy odvodené z LPC s rôznymi nastaveniami modelu. Pozorovali sme jasný trend kde sa presnosť zvyšuje s rádom modelu a maximálnym frekvenčným rozsahom. Na druhej strane zlyhalo, ako aj v iných prípadoch, použitie preemfázy, ktorá je štandardnou technikou pri vykonávaní analýzy reči LPC. T Tu však presnosti nedosahovali úrovne ako v predchádzajúcich experimentoch. Presnosť sa pohybovala na úrovni okolo 65% a najlepšie výsledky sme dosiahli pre 18. rád LPC modelu pri maximálnej frekvencii 8 kHz bez použitia preemfázy (67.05 ± 4.81%). Treba poznamenať, že uvedené výsledky odrážali iba tvar spektrálnych obalov, t.j. nebol zahrnutý zisk LPC modelu (keďže je súčasťou budenia). V ďalšom experimente bolo teda testované jeho nasadenie. Presnosť sa pre najlepšie nastavenia z prechádzajúceho experimentu zvýšila o približne 5% a najlepšie

výsledky dosiahlo nastavenie pre 20. rád LPC modelu pri maximálnej frekvencii 8 kHz so ziskom (70.09 ± 4.92%). Vo všetkých nastaveniach začlenenie zisku výrazne zvýšilo presnosť; najlepšie výsledky zaznamenali rády modelov 18 a 20 so ziskom a maximálnou frekvenciou 8 kHz.

Zvyškový signál (budenie) na kontrolu účinku glotálnych a predglotálnych znakov na SER sme tiež extrahovali z LPC modelu. Testovali sme aj rád modelu, maximálne frekvenčné limity a preemfázu. Treba ale poznamenať, že preemfáza (ak bola použitá) nebola aplikovaná na budiaci signál, ale iba na výpočtový proces odvodzujúci koeficienty LPC. Výsledky pre spektrogramy, vytvorené na základe excitačných signálov odvodených z rôznych nastavení modelu LPC, sú znázornené na obr. 3.1. Rovnaké výsledky sú zobrazené s použitím preemfázy v procese výpočtu LPC na obr. 3.2. V porovnaní so zisteniami súvisiacimi s LPC v predchádzajúcich experimentoch, v prípade budiacich signálov možno pozorovať skôr opačné výsledky, t.j. preemfáza je jednoznačne prospešná a vo všeobecnosti presnosť nerastie s rádom modelu. Môže to naznačovať, že stratégia menej presného modelu LPC funguje. Takýto filter (inverzný) aplikovaný na reč nefiltruje spektrálnu obálku, takže do budiaceho signálu uniká viac zložiek hlasového traktu. Preemfáza kladie dôraz na vyššie frekvencie, a preto inverzný LPC filter odstraňuje z reči viac vysokofrekvenčných zložiek.



Obr. 3.1: Priemerné presnosti a štandardné odchýlky pre spektrogramy skonštruované z budiacich signálov s rôznymi rádmi modelou LPC (lpc) {16, 18, 20}, maximálnymi frekvenciami (f) {4, 8 kHz} a bez použitia preemfázy. Vyplnené značky znázorňujú rovnako výkonné nastavenia.



Obr. 3.2: Priemerné presnosti a štandardné odchýlky pre spektrogramy skonštruované z budiacich signálov s rôznymi rádmi modelou LPC (lpc) {16, 18, 20}, maximálnymi frekvenciami (f) {4, 8 kHz} a s použitím preemfázy. Vyplnené značky znázorňujú rovnako výkonné nastavenia.

Doteraz boli tieto experimenty zamerane na magnitúdové spektrá eliminujúce všetky fázové informácie. Preto sa v nasledujúcom súbore experimentov zvažoval vplyv fázových signálov na SER. Experimenty boli vykonané pre nasledujúce nastavenia v prípade rozsahu 8 kHz: nespracovaná fáza, rozbalenie fázy, rozbalenie a nadzvorkovanie fázy, nastavenie referenčnej fázy, rozbalenie a nastavenie referenčnej fázy, kompenzácia posunu a kompenzácia posunu s rozbalením fázy. Rovnaké výsledky pre frekvenčný rozsah 4 kHz boli veľmi podobné a pre porovnanie výsledkov nepodstatné. Výsledky avšak dosahovali úspešnosti v rozsahu 17–33%. Aj keď presnosť bola výrazne nižšia, rozbaľovanie fázy a rozbaľovanie fázy s kompenzáciou časového posunu boli výrazne lepšie ako ostatné prístupy. Fázové experimenty boli rozšírené aj na budiace signály odvodené z inverznej LPC filtrácie. Najúspešnejšie fázové modifikácie, t. j. rozbalenie fázy, rozbalenie fázy a kompenzácia posunu a rozbalenie fazy a nastavenie referencie boli testované v kombinácii s nasledujúcimi modelmi LPC rádov: 16, 18 a 20 a s maximálnou frekvenciou 8. Najlepšie výsledky zaznamenalo rozbaľovanie fázy pre rád modelu LPC 20 (33.4 ± 4.7%) a rozbaľovanie fázy vylepšené kompenzáciou posunu rád modelu LPC 16 (31.8 ± 4.31%). Rád modelu LPC nemal v experimentoch veľký vplyv ná výsledné presnosti. V porovnaní s rečovým signálom ako celkom, dosiahli fázy založené na budiacom signáli podobné výsledky.

Príznaky založené na kepstre, ktoré sú zložitejšie, ale veľmi úspešné v mnohých aplikáciách reči, boli analyzované v poslednej sérii experimentov. Rôzne nastavenia zahrnuté vo výpočte kepstra, t.j. liftering, c_0, počet kepstrálnych koeficientov, v prípade hornej frekvencie 8 kHz a 60 bánk filtrov (najlepšie nastavenia pre Melove a gammatónové FB), boli testované na základe predchádzajúcich experimentov. Najlepšie z experimentov pre výkon MFCC s použitím 10, 13, 16 a 19 koeficientov, s a bez c_0 a lifteringu vyšlo nastavenie s 19 MFCC koeficientami, s c_0 a s použitím lifteringu (priemerná presnosť 68.2 ± 5.05%). S odvolaním sa na sľubné výsledky zahŕňajúce gammatónové FB, sme taktiež testovali kepstrálne koeficienty pomocou vytvorené pomocou gammatónovych FB (GFCC) s rovnakými nastaveniami ako

v prípade MFCC. Z experimentov s GFCC vyšlo ako najlepšie nastavenie s 19 GFCC, s c_0 a s použitím lifteringu (priemerná presnosť 71.4 ± 4.81%) a podobne ako v prípade gammatónových FB, GFCC dosiahli lepšie výsledky ako MFCC. Aj keď to nie je príliš zrejmé, ale je možné pozorovať trend zvyšovania presnosti s rastúcim počtom koeficientov, napríklad interval medzi 16 a 19 koeficientmi. Použitie c_0 sa vo väčšine prípadov ukázalo ako prospešné, najmä pri GFCC. Na druhej strane, najmenej úspešný bolo použitie lifteringu. MFCC pozostávajúce iba z 35 FB boli testované pre porovnanie výsledkov a môžeme pozorovať, že takéto nastavenie poskytuje ešte lepšie výsledky, t.j. v priemere o 2% a o 2,8% v najlepšom skóre v porovnaní s najvýkonnejšími MFB. Tento experiment nebol znázornený, pretože vyššie uvedené pozorovania zostávajú rovnaké; prirodzene pre 35 FB sa najlepší počet kepstrálnych koeficientov zodpovedajúcim spôsobom znížil na 16.

Všetky doterajšie výsledky sú založené na experimentoch na berlínskej databáze. Taktiež sme vykonali testy na inej databáze, aby sa zistilo, ako sú metódy a nastavenia konzistentné a robustné. Aby boli výsledky prehľadné, pre každý experiment sme vypočítali korelácie (Pearson) medzi presnosťami v oboch databázach. Hodnotiaca konzistentnosť metód a nastavení medzi databázami bola hodnotená aj pomocou poradovej korelácie (Spearman). Okrem toho sme vypočítali pomer zhody medzi súbormi rovnako účinných metód nájdených v každom experimente pre obe databázy. Nakoniec sme dospeli k hypotéze, že výsledky a metódy na rôznych databázach navzájom nekorelujú, vyhodnotené na 10% úrovni významnosti. Výsledky a priemerné presnosti pozorované v každom súbore experimentov sú uvedené v tab. 3.1.

Tab. 3.1: Korelácie metód a nastavení naprieč databázami pre každú sadu experimentov. T — hypotéza H_0 je pravdivá (akceptovaná), F — hypotéza H_0 je nepravdivá (zamietnutá)

Test	Korelácia priemernej presnosti	H_0 nekorelácia presností	Poradová korelácia metód	H_0 nekorelácia	Zhoda rovnako výkonných metód	Priemerná presnosť EMODB	Priemerná presnosť IEMOCAP
LPC modelovanie	0.93	F	1	F	1	0.7	0.35
Melove rozsahy mierky	0.92	F	0.89	F	0.6	0.63	0.38
MFCC a GFCC	0.82	F	0.87	F	0.67	0.67	0.39
Klasifikačný blok / dĺžka spektrogramu	0.65	F	0.7	F	0.33	0.59	0.37
Lineárne frekvenčné rozsahy	0.43	T	0.3	T	0.25	0.62	0.39
Banky filtrov	0.35	T	0.38	T	0.33	0.67	0.41
Budiace signály	0.4	T	0.3	T	0.6	0.6	0.39
Modifikácie magnitúdy	0.07	T	0.09	T	0	0.66	0.33

Výber klasifikačnej metódy (CNN) pre túto úlohu (spektrogram ako vektor príznakov) bol odôvodnený porovnaním jej výkonnosti v berlínskej databáze pre najlepšie vlastnosti

(GBF experimenty) s prístupom použitia SVM [3]. V priemere CNN prekonala SVN o 45% (presnosť: 74,7% vs. 51,5%) a pokiaľ ide o najlepšie nastavenia o 44,7% (presnosť: 75,38% vs. 52,1%). Navyše, hypotéza o rovnosti oboch metód musí byť zamietnutá na 10% úrovni významnosti.

Za účelom zhrnutia výsledkov sú metódy a nastavenia zoskupené a usporiadané do daných intervalov presnosti na 10% úrovni významnosti. Metódy a príslušné nastavenia, ktoré dosahujú presnosť nad 70% v prípade berlínskej databázy, sú uvedené v tab. 3.2. Pre prehľadnosť výsledkov je uvedených len 10 najúspešnejších metód spadajúcich do rozsahu presnosti 70%, ale v skutočnosti bolo v tomto rozsahu 28 metód. Rovnaké výsledky uvádzajúce 5 najlepších metód testovaných na databáze IEMOCAP sú uvedené v tab. 3.3.

Tab. 3.2: Metódy a nastavenia s presňosťou nad 70% na 10% úrovni významnosti zoradené zostupne (EMODB). Gammatónová BF (GFB), Melova FB (MFB), Magnitúda (mag), Logaritmická magnitúda (Log mag), Logaritmický výkon (Log pow).

Príznaky	Fmin (kHz)	Fmax (kHz)	Modifikácia magnitúdy	Frekvenčná škála	Dĺžka vektoru príznakov	Typ signálu: budenie / obálka	Presnosť (%)
GFB	0	8	mag	RBF	60	obálka	75.38
GFB	0	8	mag	RBF	45	obálka	75.21
MFB	0	8	mag	MEL	60	obálka	74.63
GFB	0	8	mag	RBF	30	obálka	73.51
MFB	0	8	mag	MEL	45	obálka	73.48
BFB	0	8	mag	BAR	30	obálka	73.3
BFB	0	8	mag	BAR	60	obálka	73.2
BFB	0	8	mag	BAR	45	obálka	72.68
Spektrogram	0	8	Log mag	Hz	160	oboje	72.45
MFB	0	8	mag	MEL	30	obálka	72
Spektrogram	0	8	mag	Hz	160	oboje	71.88
GFCC, c_0	0	8	Log pow	RBF	19	obálka	71.4
Spektrogram	0.3	8	mag	Mel	155	oboje	71.17
Spektrogram	0	4	mag	Hz	80	oboje	71.12

Tab. 3.3: 5 najlepších metód a nastavení testovaných na databáze IEMOCAP.

Príznaky	Fmin (kHz)	Fmax (kHz)	Modifikácia magnitúdy	Frekvenčná škála	Dĺžka vektoru príznakov	Typ signálu: budenie / obálka	Presnosť (%)
GFB	0	8	mag	RBF	60	obálka	44.35
GFB	0	4	mag	RBF	45	obálka	44.24
GFCC	0	8	Log pow	RBF	19	obálka	43.67
GFCC, c_0	0	8	Log pow	RBF	19	obálka	43.24
MFB	0	8	mag	BAR	60	obálka	43.02

Ako je uvedené v tab. 3.2, najúspešnejšie metódy boli odvodené z FB s použitím gammatónových, Melovych, a barkovych škál. Z toho vyplýva, že vlastnosti hlasového traktu (spektrálna obálka) sú veľmi dôležité a spôsob ich extrakcie a vyjadrenia, napr. funkcie založené na modeli LPC, sa ukázali ako štatisticky horšie pri extrakcii a reprezentovaní rovnakého druhu informácií. Aj keď je poradová korelácia iba 0.38, tieto zistenia sa väčšinou vzťahujú na databázu IEMOCAP, ako je možné vidieť v tab 3.3, t.j. najlepšie metódy sú tiež založené na GFB.

Ďalšie úspešné metódy pokrývajú spektrálne vlastnosti (spektrogramy) odvodené z celých rečových signálov, t. j. obsahujúce signály excitácie aj hlasového traktu. Na tento účel sa však používa podstatne viac parametrov ako pri FB, napr. 160 vs. 60, viď. tab. 3.2. Nastavenia sa opäť ukázali ako dôležité, najmä ak sa berú do úvahy frekvenčné rozsahy v kombinácii s metódami modifikácie magnitúdy.

Najmenej úspešné metódy boli založené na príznakoch extrahovaných z fázovej zložky signálu. Ani jedno z testovaných nastavení však nezlyhalo, t.j. všetky boli schopné extrahovať niektoré informácie relevantné pre SER, takže výkon bol výrazne lepší ako výkon poskytnutý náhodným klasifikátorom, t. j. 14.3% (sedem emočných tried). V niektorých systémoch teda môžu stále pôsobiť ako pomocné prvky.

4

Vplyv základných vlastností rečových a zvukových signálov na presnosť aplikácií spracovania reči a zvuku

Aplikácií na spracovanie reči a zvuku je veľa a ich počet rastie. Môžu pokrývať širokú škálu úloh, pričom každá má iné požiadavky na spracovávané rečové alebo zvukové signály, a teda nepriamo aj na zvukové snímače. Venujeme sa vplyvu základných fyzikálnych vlastností reči a zvukových signálov na presnosť rozpoznávania hlavných aplikácií na spracovanie reči/audia, t. j. rozpoznávanie reči, rozpoznávanie rečníka, rozpoznávanie emócií reči a rozpoznávanie zvukových udalostí. Osobitný dôraz sa kladie na frekvenčné rozsahy, časové intervaly a presnosť zobrazenia (kvantizácia) a zložitosť modelov vhodných pre každú triedu aplikácií. Pomocou doménovo špecifických dátových súborov, vhodných metód extrakcie funkcií a komplexných modelov neurónových sietí bolo možné otestovať a vyhodnotiť vplyv základných vlastností rečového a zvukového signálu na dosiahnuté presnosti pre každú skupinu aplikácií. Testy potvrdili, že základné parametre ovplyvňujú celkový výkon a navyše je tento efekt závislý od domény. Preto presné znalosti o rozsahu týchto účinkov môžu byť cenné pre systémových dizajnérov pri výbere vhodného hardvéru, senzorov, architektúry a softvéru pre konkrétnu aplikáciu, najmä v prípade obmedzených zdrojov.

4.1 Materiály a metódy

Vzhľadom na hlavný cieľ tejto kapitoly, ktorým je merať a hodnotiť vplyv základných vlastností signálov na výkon hlavných rečových/audio aplikácií s priamym vzťahom k návrhu akustických snímačov, boli vybrané nasledovné metódy extrakcie vlastností a klasifikácie. V prípade extrakcie vlastností boli vybrané spektogramy a Melove banky filtrov. Dôvodom je, že tieto základné akustické vlastnosti poskytujú cenné rozloženie v čase a frekvencii, ktoré je nevyhnutné pre analýzu nestacionárnych audio signálov, ako sú reč a environmentálne zvuky. Okrem toho sú to vlastnosti, ktoré najmenej menia signál, t.j. zavádzajú minimálne umelé úpravy signálu a zachovávajú väčšinu relevantných akustických informácií. Ich konštrukciou nám dovoľujú mať priamu kontrolu nad všetkými dôležitými fyzikálnymi parametrami, napríklad nad frekvenčnými a časovými rozsahmi, čo nemusí byť vždy možné pri použití zložitejších alebo dokonca kombinovaných vlastností, kde viac aspektov môže byť ovplyvnených jedným nastavením. Keďže spektogramy sú 2D signály prejavujúce prirodzenú variabilitu umiestnenia a rozlíšenia, t.j. obsahujúce rôzne vzory v rozličných časových a frekvenčných polohách, prirodzenou voľbou pre blokovú klasifikáciu reči/zvuku sú CNN. Blokové spracovanie nám umožňuje vyhnúť sa použitiu ďalších modelov pre dlhotrvajúce (nepretržité) modelovanie v čase, napríklad jazykové modely v rozpoznávaní reči atď., ktoré ovplyvňujú celkový výkon. Na druhej strane, takéto modely (nie akustické) nesúvisia s vlastnosťami signálu relevantnými pre akustické snímače. Nadto nám to umožňuje použiť jednotný klasifikačný rámec, ktorý je nezávislý od aplikácie. Napokon, okrem špeciálne navrhnutých CNN boli otestované známe, zložité a vopred natrénované CNN, s použitím populárneho transferového učenia. Preto nasledujúce sekcie stručne sumarizujú vyššie uvedené metódy.

Budeme pracovať so spektrogramami, bankami filtrov a kepstrálnymi koeficiantami ku ktorým doplníme ešte kvantizáciu. Kvantizácia zavádza kvantizačný šum, ktorý môže ovplyvniť výkon algoritmov. V prípade rovnomerne rozdelených signálov je stredná hodnota výkonu zavedeného šumu (bieleho) rovná:

$$P_n = \frac{1}{3}\frac{A_{max}^2}{2^{2bits}} \tag{4.1}$$

A_{max} je maximálna amplitúda a *bits* je počet bitov kvantizácie. Pre štandardnú kvalitu CD sa používa lineárna 16-bitová kvantizácia. Kvôli nelineárnemu vnímaniu, najmä v telefónii, sa používa 8-bitová nelineárna kvantizácia (a-zákon alebo μ-zákon). To znamená, že signál je pred kvantizáciou nelineárne komprimovaný nasledovne (μ-zákon):

$$X_\mu = sgn(x)\frac{\ln(1+255|x|)}{\ln(256)} \tag{4.2}$$

kde X_μ a x sú komprimované (μ-zákon) resp. pôvodné signály. Vo fáze dekódovania je signál dekomprimovaný pomocou inverzie 4.2, t.j.

$$x = sgn(X_\mu)\frac{256^{|X_\mu|} - 1}{255} \tag{4.3}$$

To efektívne pomáha znížiť bitové rýchlosti pri zachovaní prijateľnej kvality.

Keďže rečové a audio signály sú veľmi zložité, je nevyhnutné mať obrovské a špecifické dátové sady. V našich experimentoch, ktoré zahŕňali 4 rôzne aplikácie, sme použili 4 dátové sady, aby sme splnili ich špecifické vlastnosti.

V úlohe rozpoznávania rečníkov sme sa rozhodli použiť dataset LibriSpeech [4], ktorý je open source a ponúka niekoľko podmnožín. Vybrali sme podmnožinu Train-clean-100, ktorá obsahuje približne 100 hodín nahrávok reči 251 rôznych rečníkov (125 žien, 126 mužov), ktorí čítajú rôzne audioknihy v "čistom" prostredí. Každý rečník nahrával priemerne 25 minút vhodnej reči, ktorá bola segmentovaná do nahrávok. Nahrávky mali rôzne dĺžky v rozmedzí od 2 do 40 sekúnd. Nahrávky boli vzorkované so vzorkovacou frekvenciou 16 kHz použitím 16-bitovej kvantizácie.

Keďže naším cieľom bolo testovať základné fyzikálne (akustické) charakteristiky ako frekvencia, čas a kvantizácia, nebolo potrebné zahrnúť špecifický jazykový model. Preto by systém rozpoznávania izolovaných slov pokrývajúci široký rozsah rečníkov a podmienok obsahujúci akusticky bohaté vzorky mal byť dostačujúci. Za týmto účelom bol vybraný dataset Speech Commands Dataset v0.01 [5]. Obsahuje 30 rôznych slov (príkazov) viackrát vyslovených väčšinou účastníkov. Celkovo má 58 000 jeden sekundu dlhých záznamov vyprodukovaných tisíckami rečníkov. Databáza taktiež obsahuje reálne aj syntetizované zvuky. Nahrávky boli uskutočnené skôr v nekontrolovanom, ale interiérovom prostredí po celom svete. Nahrávky boli získané v rôznych formátoch (použitím aj kompresných techník), ktoré nakoniec boli konvertované na 16-bitové, 16 kHz PCM vzorky. Je potrebné poznamenať, že nahrávky kratšie ako 1s neboli zvážené pre ďalšie spracovanie, keďže sa pozorovalo, že mnohé kratšie záznamy obsahujú len šum, hoci boli označené ako skutočné slová.

V našich experimentoch AER sme sa z niekoľkých dôvodov rozhodli pre dataset ESC-50 [6]. Obsahuje širokú variabilitu zvukov zaradených do 5 širokých tried, a to sú: zvieratá, príroda a voda, človek (bez reči), domáce prostredie a mestský hluk, pričom každá má 10 podtried tvoriacich dohromady 50 odlišných kategórií. Každá kategória obsahuje 40 nahrávok, čo výsledne dáva 2000 záznamov. Každý záznam trvá 5 sekúnd a obsahuje len jednu udalosť s určitým množstvom pozadového šumu, čo nám umožňuje aplikovať blokové spracovanie. Nahrávky sú uložené ako PCM vzorky kvantované na 16 bitov so vzorkovacou frekvenciou 44,1 kHz. Je potrebné poznamenať, že z dôvodov uskutočniteľnosti (čas vykonania) bolo zvážených len 8 odlišných kategórií nasledovne: štekajúci pes, oheň, tečúca voda, plačúce dieťa, klopanie na dvere, rozbíjanie skla, ohňostroj a sirény.

Vo väčšine experimentov boli použité jednotlivé siete pre testované domény. Ak nie je uvedené inak, siete sú nasledujúce. V prípade testov SER (rozpoznávania emócií z reči) sa používala sieť so 2 konvolučnými vrstvami s 16 a 32 jadrami o veľkosti 5 × 5, 2 max pooling vrstvami o veľkosti 2 × 2, hustou vrstvou (128 neurónov), dropout vrstvou a softmax vrstvou s celkovo približne 170 tisíc parametrami. V experimentoch na rozpoznávanie rečníka sa použila 5-vrstvová CNN (3 konvolučné a 2 husté vrstvy) s 96 jadrami o veľkosti

2 × 2 a s približne 50 tisíc parametrami. Pre aplikáciu rozpoznávania reči sme nasadili 6-vrstvovú sieť (4 konvolučné a 2 husté vrstvy) s 40 jadrami, 3 max pooling vrstvami a jednou dropout vrstvou, ktorá mala celkovo približne 55 tisíc parametrov. Napokon, v testoch AER (rozpoznávania zvukových udalostí) sa použila 6-vrstvová sieť (3 konvolučné a 3 husté) s 168 jadrami o veľkosti 3 × 3 a 2 max pooling vrstvami s celkovo približne 1 miliónom parametrov. Na trénovanie sietí na dátových súboroch, ktoré boli rozdelené na trénovaciu (80% dát), testovaciu (20% dát) a validačnú (20% trénovacích dát) sady, každá majúca rovnaký počet vzoriek na triedu, bol zvolený trénovací algoritmus ADAM. Aplikovalo sa kritérium predčasného ukončenia, aby sa zabránilo pretrénovaniu.

4.2 Výsledky a diskusia

Prvý súbor experimentov je zameraný na testovanie nastavení segmentácie reči/zvuku, t.j. dĺžok rámcov (okien), ich posunov, ako aj dĺžok spektrogramov a ich posunov. Tieto parametre kontrolujú časové—frekvenčné rozlíšenie, počet dostupných dát, veľkosť FV a stupeň redundancie. Sú preto priamo spojené s požiadavkami na pamäť, oneskoreniami spracovania a výpočtovou záťažou.

Pri experimentoch pre rozpoznávanie rečníka boli testované rámce s dĺžkami 10, 20 a 30 ms s 50% posunmi pre 1 sekundu dlhé spektrogramy s frekvenčným rozsahom 0–8 kHz. Ako najúspešnejšie nastavenie sa ukázala dĺžka rámca 10ms, kde sme dosiahli priemernú presnosť na úrovni 98%, v porovnaní s dĺžkou rámca 30ms, kde sme dosiahli priemernú presnosť o približne 4% nižšiu. Je zrejmé, že kratšie rámce (s vyšším časovým rozlíšením) dosiahli lepšie výsledky. Je to čiastočne pripísané vyššiemu počtu rámcov na jeden spektrogram a tým pádom dlhším FV. Pri teste časového rozsahu boli testované Mel spektrogramy (s frekvenčným rozsahom 0.1–8 kHz) s nasledujúcimi trvaniami: 0.5, 1, 2, 3, 4 a 5 sekúnd. Ako najlepšie sa ukázali experimenty s dĺžkou trvania spektrogramu 3s. Priemerná presnosť tu dosahovala úroveň takmer 86% v porovnaní s 0.5s spektrogramom, kde priemerná presnosť bola iba 78.5%. Pri 4s a 5s spektrogramoch sme už pozorovali pokles presnosti o približne 1.5–2%.

V experimentoch pre rozpoznávanie reči sme vykonávali pre dĺžky rámcov 15, 20 a 25 ms s posunmi 90, 75 a 50% medzi susednými rámcami pre logaritmické spektrogramy. Kratšie rámce, napríklad 15 ms, miernie prevýšili dlhšie rámce. Uprednostnili sa spektrogramy s väčším počtom rámcov, a to aj na úkor jemnejšieho frekvenčného rozlíšenia. Okrem toho sa uprednostňovali kratšie posuny, čo ešte viac zvýšilo počet použiteľných rámcov za cenu vyššej redundancie. Najúspešnejšie nastavenie bolo s dĺžkou rámca 15ms a prekryvom 75%, kde priemerná presnosť bola okolo 81% v porovnaní s dĺžkou rámca 25ms a prekryvom 90%, kde priemerná presnosť bola na úrovni 68%.

V experimentoch pre rozpoznávanie zvukových udalostí, v prípade všeobecných zvukov z prostredia, ktoré majú neznáme vlastnosti (v čase a frekvencii), boli testované niektoré spracovateľské parametre v rozsahoch, ktoré nie sú bežné u signálov reči. Experimenty boli vykonané pre dĺžky rámcov 20, 30, 40 a 50 ms s 50% posunom. Po prvé, rozdiely v

presnosti boli pomerne nevýznamné naprieč testovaným rozsahom dĺžok rámcov (20–50 ms), to znamená, že presnosť nie je veľmi citlivá na tento parameter. Po druhé, okrem rámcu o dĺžke 40 ms (najúspešnejšie nastavenie - priemerná presnosť 75.2%), bola zaznamenaná mierne klesajúca tendencia v presnosti, čo naznačuje, že štandardná dĺžka rámcu pre reč (20 ms) je tiež vhodná. Odchýlka pri 40 ms najpravdepodobnejšie súvisí s náhodnou povahou tréningu (5-násobná validácia), keďže nebola žiadna zrejmá fyzická príčina, prečo by k tomu malo dôjsť, okrem toho, že existujú veľmi špeciálne triedy zvukov, ktoré vyžadujú práve toto časovo-frekvenčné rozlíšenie. Aj tak by to však neumožňovalo všeobecnú aplikovateľnosť. Ďalej boli testované vhodné dĺžky signálov 500, 750, 1000 a 1500 ms a ich 100, 75, 50 a 25% posuny. Pre lepšiu reprezentáciu sú tieto uvedené v tab. 4.1, spolu s priemernými presnosťami pre dĺžky signálu a posuny zvlášť, aby bolo možné zaznamenať potenciálne tendencie.

Tab. 4.1: Vplyv dĺžok a posunov signálu na presnosť (rozpznávanie zvukových udalostí). Sú tiež uvedené priemerné hodnoty pre dĺžky a posuny signálu.

Posuny [%]	Dĺžka 500 ms	Dĺžka 750 ms	Dĺžka 1000 ms	Dĺžka 1500 ms	Priemer [%]
100	70.74	72.08	71.83	61.66	69.07
75	70.12	69.27	74.3	70	70.92
50	72.71	75	74.35	73.25	73.82
25	72.9	74.09	73.92	72.24	73.28
Priemer [%]	71.61	72.61	73.6	69.2875	-

Je vidieť, že kratšie posuny, teda väčšie množstvo dát (redundancia), sú výhodné, zatiaľ čo optimálna dĺžka analyzovaného signálu sa nachádza niekde medzi 750 a 1000 ms. Ide o kompromis medzi schopnosťou zachytiť celú udalosť a neinterferovať s inou udalosťou alebo pozadovým hlukom. Teda toto môže byť ovplyvnené sadou udalostí, ktoré majú byť rozpoznané.

Druhý súbor experimentov je zameraný na frekvenčné rozsahy.Frekvenčné rozsahy sú dôležitými parametrami, pretože môžu seriózne ovplyvniť mieru rozpoznávania, množstvo spracovaných a uložených dát, kvalitu a náklady na audio vstupné senzory atď. Okrem toho sa môžu líšiť v závislosti od aplikácie.
Ako aj v prvom súbore experimentov, aj tu sme najskôr testovali aplikáciu pre rozpoznávanie rečníka. Testovali sme vplyv maximálnych frekvenčných limitov pri fixne nastavenej minimálnej frekvencii (0 Hz), a minimálne frekvenčné limity, pričom maximálna frekvencia je nastavená na 8 kHz; použité vektory príznakov boli 1 sekundu dlhé spektrogramy. Testy jasne ukázali, že pre zabezpečenie najlepšieho rozpoznávania sú potrebné oba limity: maximálna horná a dolná frekvencia. Najlepšie výsledky sme dosiali pre frekvenčný rozsah 0–8 kHz (priemerná presnosť 98.9%) a najhoršie pre frekvenčný rozsah 0.3–4 kHz (priemerná presnosť

96%).
V prípade rozpoznávanie reči sú výsledky experimentov pre minimálne frekvencie 0, 150 a 300 Hz kombinované s maximálnymi frekvenciami 4 a 8 kHz a pri použití magnitúdových spektrogramov. Z výsledkov je zrejmé, že frekvenčný rozsah 4 kHz je nevyhnutný, zatiaľ čo prínos horného frekvenčného limitu 8 kHz je pomerne obmedzený, čo je v súlade s všeobecnými poznatkami (telefónia alebo zrozumiteľnosť reči). Avšak tu sa zaznamenané zlepšenia dokonca aj pri nižších frekvenciách (pod 300 Hz). Najlepšie výsledky dosiahli práve testy pre frekvenčný rozsah 0–4 kHz, kde priemerná presnosť dosiahla úroveň 82%, narozdiel od frekvenčného rozsahu 0.15–8 kHz, kde priemerná presnosť dosiahla úroveň 62%.
Pre aplikácie rozpoznávanie zvukových udalostí boli testované aj ďalšie frekvencie pre horné aj dolné limity a sú spísané v tab. 4.2 s priemernými presnosťami pre dolné a horné frekvenčné limity, aby boli potenciálne trendy lepšie viditeľné.

Tab. 4.2: Presnosti (rozpoznávanie zvukových udalostí) pre rôzne dolné a horné frekvenčné limity. Uvedené sú tiež priemerné hodnoty pre oba limity.

F_{min} [Hz]	F_{max} 4 kHz	F_{max} 8 kHz	F_{max} 12.5 kHz	F_{max} 17.5 kHz	F_{max} 22.05 kHz	Priemer [%]
0	75.72	79.61	81.22	79.61	81.55	79.5
100	75.08	85.37	84.78	85.11	86.76	83.2
200	75.72	82.52	71.52	80.25	81.87	78.34
300	74.75	72.49	78.64	83.81	79.54	77.84
Priemer [%]	75.91	79.99	79.04	82.2	82.18	-

Na základe priemerných hodnôt boli preferované horné a dolné frekvenčné limity približne od 100 Hz do cca 17.5 kHz s obmedzeným potenciálom až do 22.05 kHz. Presné limity však môžu závisieť od špecifických tried zvukov, ktoré sa majú oddeliť.

Okrem základného zobrazovania signálov reči/zvuku prostredníctvom spektrogramov sme skúmali aj viac spracované príznaky zamerané na charakteristiky hlasového traktu. V tomto súbore experimentov sme otestovali najbežnejšie akustické príznaky, a to MFB a MFCC.
Pre aplikácie rozpoznávanie rečníka sme testovali výkonnosť MFB (30–80) a MFCC (13–16). Pozorovali sme pozitívny trend pre vyšší počet FB až do približne 70 (priemerná presnosť 94 %, pre 80 bol pokles presnosti približne o 2 %). Na druhej strane, MFCC poskytli stabilnejšie výsledky v celom testovanom rozsahu s malým vrcholom okolo 15 koeficientov (priemerná presnosť okolo 93 %). Toto ukazuje schopnosť MFCC komprimovať akustické informácie.
Výsledky charakteristík hlasového traktu, v aplikáciach pre rozpoznávanie reči, sme testovali pre 26, 30, 35, 40 a 45 MFB a 17, 21 a 26 MFCC. Počty MFCC boli 50, 60 a 75 % z 35 MFB (najlepší skórer), z ktorých boli MFCC vypočítané. Viac spracované príznaky (MFCC) v kombinácii s CNN nepriniesli želané zlepšenie, čo naznačuje, že CNN môžu stále nájsť lepšiu reprezentáciu. Vhodný počet MFB bol okolo 35 (priemerná presnosť na úrovni 83 %), pre

ktoré MFB miernie predčili spektrogram založený na magnitúde približne o 1 percentuálny bod.
Výsledky pre rôzne nastavenia príznakov MFB a MFCC, v aplikáciach pre rozpoznávanie zvukových udalostí, sme testovali pre frekvenčný rozsah 100 Hz–17.5 kHz. V prípade MFB bol pozorovaný vrchol pri 64 FB (priemerná presnosť 87 %). Tieto príznaky navyše mierne prekonali spektrogramy založené na magnitúde. Na druhej strane, viac spracované príznaky MFCC nezaznamenali žiadne zlepšenia v kombinácii s CNN.

V ďalšej sade experimentov bol testovaný vplyv presnosti reprezentácie signálu. V rámci všetkých aplikácií boli použité nasledujúce schémy kvantizácie: 16 bitová lineárna (pôvodná reprezentácia), 12 bitová lineárna, 8 bitová lineárna a 8 bitová μ-zákon (nelineárna). Treba poznamenať, že pred kvantizáciou boli signály normalizované na interval (-1, 1).
Experimenty pre SER boli vykonané s 1 sekundu dlhými spektrogramami. Hoci pôvodný signál zaznamenal najlepšie skóre, rozdiely nie sú významné, a to najmä v prípade 8 bitovej kvantizácie podľa μ-zákona (priemerná presnosť 64.3 %). To naznačuje, že výkon bol ovplyvnený len minimálnym spôsobom, čo znamená, že v tejto aplikácii môže byť ušetrené významné množstvo dát.
Experimenty pre testované schémy kvantizácie, pre rozpoznávanie rečníka, boli taktiež vykonané na 1 sekundu dlhých spektrogramoch s frekvenčným rozsahom 0–8 kHz. Je zaujímavé, že aj v tomto prípade bol vplyv kvantizácie v testovanom rozsahu skôr zanedbateľný. Napriek tomu 8-bitové reprezentácie poskytli mierne horšie výsledky. Najlepšie výsledky sme dosiahli pre 12 bitovú lineárnu kvantizáciu (priemerná presnosť 99.1 %).
Vplyv kvantizácie (rozpoznávanie reči) bol testovaný pomocou spektrogramov aj MFB. Metódy kvantizácie 12-bitová lineárna a 8-bitová podľa μ-zákona stále poskytujú konkurencieschopné výsledky, pričom zároveň ušetria významný priestor kvôli nespracovanej reprezentácii. Tieto pozorovania sú pritom platné ako pre spektrogramy tak aj pre MFB. Na druhej strane, 8-bitová lineárna kvantizácia sa ukázala byť nedostatočná, keďže viedla k poklesu presnosti o 7.6 % (spektrogram) a 12 % (MFB).

Doteraz bol v rámci všetkých pokusov pre konkrétnu aplikáciu použitý jeden model CNN. Takéto modely boli navrhnuté na základe už publikovaných článkov a prispôsobené s použitím obmedzeného množstva dát. Ďalej, pre najlepšie nastavenia (príznaky), je hodnotený vplyv zložitosti modelu testovaním ďalších modelov. To môže byť dôležité pre hardvérové požiadavky, obzvlášť v prípade samostatných aplikácií s obmedzenými zdrojmi.
Zložitosť modelu pre SER bola testovaná použitím troch vlastne navrhnutých sietí. Prvá bola pôvodná, označená ako CNN 2. Druhou je CNN 3, ktorá bola 5-vrstvová CNN sieť s 3 konvolučnými a 2 hustými vrstvami s 3 vrstvami max pooling, 112 jadrami veľkosti 5 × 5 a približne 150 tisíc parametrami. Nakoniec, CNN 4 bola 6-vrstvová CNN sieť so 4 konvolučnými a 2 hustými vrstvami, 4 vrstvami max pooling, 240 jadrami veľkosti 5 × 5 a približne 310 tisíc parametrami. Vstupom do CNN boli 1 sekundu dlhé spektrogrami vo frekvenčnom rozsahu 0–8 kHz. Z výsledkov bolo možné vidieť, že zvýšený počet vrstiev nebol veľmi prínosný a dokonca sieť CNN s dvoma konvolučnými vrstvami poskytla najlepšie výsledky z testovaných sietí a s približne iba polovicou parametrov (presnosť 68 %).
V experimtoch rozpoznávania rečníka bol v systéme Keras nasadený nástroj na ladenie hyperparametrov, aby sa našla "optimálna" štruktúra CNN pre použité príznaky (MFB a MFCC osobitne). Systém bol inštruovaný hľadať 3 až 6 konvolučných vrstiev, každá s 16,

20, 32, 48 a 64 jadrami o rozmeroch 2 × 2, 3 × 3, 4 × 4, a pooling vrstvami. Najlepšia sieť pre MFB mala 5 konvolučných vrstiev s 20, 32, 64, 32 a 20 jadrami o veľkosti 2 × 2, 5 max pooling vrstiev 2 × 2 a jednu hustú vrstvu s 251 neurónmi, čiže celkovo približne 39 tisíc parametrov. V prípade MFCC bola nájdená 3-vrstvová sieť s 32, 48 a 64 jadrami o veľkostiach 3 × 3, 3 × 3 a 2 × 2, s rovnakým počtom a veľkosťami pooling vrstiev a 2 hustými vrstvami s 208 a 251 neurónmi, celkovo s približne 59 tisíc parametrami. Ako zložitejšia sieť pre oba príznaky bola vybraná Alexnet [7].
Okrem vlastne navrhnutej 4 vrstvovej CNN boli testované aj zložité predtrénované siete Inception V3 [8] a Alexnet pre aplikácie rozpoznávania reči. Obidve boli dotrénované s použitím 1 sekundu dlhých logaritmických spektrogramov. Hoci dosiahnuté presnosti sa mierne líšili v rozmedzí od 84 do 90 %, zložitosti sietí boli veľmi rozdielne, napríklad 0.091 % (CNN 4), 39,9 % (Inception V3) a 100 % (Alexnet). Najlepšie výsledky boli dosiahnuté s Inception V3, hoci tá mala iba približne 40 % trénovateľných parametrov Alexnet. To ukazuje, že aj štruktúra siete je dôležitá a relatívne jednoduchá sieť môže tiež poskytnúť konkurencieschopné výsledky. Nakoniec treba poznamenať, že výsledok sa môže líšiť podľa veľkosti podporovaného slovníka.

V posledných experimentoch (rozpoznávanie zvukových udalostí) boli testované niekoľko malých, vlastne navrhnutých modelov, ako aj predtrénované siete, t.j. Inception V3 [8], VGG19 [9] a ResNet50v2 [10]. CNN 6 bol model so 6 konvolučnými vrstvami, 4 hustými vrstvami, 3 max pooling vrstvami, 224 jadrami a približne 277 tisíc parametrami, a CNN 8 predstavuje CNN so 8 konvolučnými vrstvami, 4 hustými vrstvami, 4 max. pooling vrstvami, 576 jadrami a približne 198 tisíc parametrami. Presnosti modelov boli v rozsahu od 85 do 93 %, kde najlepší výsledok poskytla najzložitejšia sieť VGG19. Predtrénované modely zaznamenali lepšie výsledky ako vlastné malé modely, a to o 7.3 % (v najlepších prípadoch). Na druhej strane boli medzi relatívnymi zložitosťami veľké rozdiely, t.j. 0.19 % (CNN 6), 0.18 % (CNN 8), 16.6 % (Inception), 17.8 % (ResNet) a 100 % (VGG19). To ukazuje, že je možné v prípade obmedzených zdrojov používať malé modely za cenu miernych poklesov presnosti.

5

Zhodnotenie

V tejto práci sme sa zaoberali komplexným skúmaním a hodnotením rôznych aspektov rozpoznávania emócií v reči pomocou moderných metód spracovania rečového signálu a hĺbkového učenia. Zameriavajúc sa na detailnú analýzu významu charakteristík rečového signálu a efektivity konvolučných neurónových sietí (CNN), táto práca poskytuje cenné poznatky a príspevky k pochopeniu a zlepšeniu rozpoznania emócií v reči (SER). Na základe starostlivo navrhnutých experimentov s využitím Berlínskej databázy emócií sme dokázali identifikovať kľúčové faktory, ktoré majú priamy vplyv na výkon systémov SER. Naše výskumné zistenia ukazujú, že optimalizácia dĺžok rámcov a posunov medzi spektrogramami je kriticky dôležitá pre maximalizáciu presnosti systémov SER. Experimenty odhalili, že existuje určitá optimálna konfigurácia, pri ktorej systém dosahuje najvyššiu presnosť. V tejto práci sme tiež poukázali na fakt, že aplikácia rôznych technik predspracovania signálu a výber relevantných charakteristík môže mať významný vplyv na úspešnosť rozpoznávania emócií, čo zvýrazňuje potrebu dôkladnej analýzy a experimentovania s rôznymi prístupmi k predspracovaniu reči.

Okrem toho sme demonštrovali, že použitie CNN s dôkladne navrhnutou architektúrou môže výrazne prispieť k lepšej identifikácii emócií zo zvukových nahrávok. Architektonické rozhodnutia, ako napríklad hĺbka siete, veľkosť a krok konvolučných filtrov, ako aj stratégia trénovania, hrajú podstatnú úlohu vo výkonnosti konečného modelu. Z praktického hľadiska ponúka táto práca podrobnejší pohľad na to, ako rôzne metódy spracovania signálu a štrukturálny dizajn siete môžu ovplyvniť schopnosť systémov SER presne rozpoznať špecifické emócie. Niektoré z našich experimentov poskytli príklady významných zlepšení v presnosti, čo naznačuje opodstatnenosť cielenej optimalizácie a množstvu experimentov v rámci dizajnu systému.

5.1 Prínosy

- **Analýza rečového signálu v časovej oblasti**
  - nájdenie optimálnych nastavení rámcov a posunov na predspracovanie signálu v časovej oblasti
  - nájdenie optimálneho pomeru medzi dĺžkou analyzovaného rámca a veľkosti trénovacieho datasetu

  - porovnanie rôznych typov okien pri tvorbe spektrogramu ako vstupu do CNN na rozpoznávanie emócií

- **Časovo frekvenčná analýza rečového signálu**
  - optimalizácia príznakov na rozpoznávanie emócií pomocou úprav vstupného signálu vo frekvenčnej oblasti
  - realizácia a aplikácia gammatónovej BF a GFCC, ktoré simulujú správanie kochlei v ľudskom uchu na rozpoznávanie emócií
  - porovnanie rôznych psycho-akustických škál a nájdenie optimálnej pre SER
  - vyhodotenie úspesňosti viacerých rádov LPC modelu pre SER a príznakov vytvorených z budiaceho signálu LPC modelu
  - analýza a vyhodnotenie závislosti parametrov rečového signálu od zvolenej databázy na rozpoznávanie emócií
  - vyhodnotenie úspešnosti fázovej informácie rečového signálu pre SER
- **Aplikácie rečového signálu**
  - návrh optimalizácia CNN architektúry a porovnanie s inými architektúrami CNN pre SER
  - porovnanie s nezvukovými existujúcimi modelmi CNN dotrénovaných na aplikácie používajúce rečové signály
  - aplikácia a vyhodnotenie nájdených vhodných parametrov predspracovania rečového signálu aj na oblasti ako rozpoznávanie rečníka, rozpoznávanie reči a rozpoznávanie zvukových udalostí

6

Publikácie a riešené projekty autora

Publikácie

- KAČUR, Juraj - PUTERKA, Boris - PAVLOVIČOVÁ, Jarmila - ORAVEC, Miloš. On the speech properties and feature extraction methods in speech emotion recognition. In Sensors. Vol. 21, iss. 5 (2021), Art. no. 1888 [27] s. ISSN 1424-8220 (2021: 3.847 - IF, Q2 - JCR Best Q, 0.803 - SJR, Q1 - SJR Best Q). V databáze: SCOPUS: 2-s2.0-85102031118 ; WOS: 000628669900001 ; CC: 000628669900001 ; DOI: 10.3390/s21051888. (15 citations in Scopus)
- KAČUR, Juraj - PUTERKA, Boris - PAVLOVIČOVÁ, Jarmila - ORAVEC, Miloš. Frequency, time, representation and modeling aspects for major speech and audio processing applications. In Sensors. Vol. 22, iss. 16 (2022), Art. no. 6304 [26] s. ISSN 1424-8220 (2022: 3.900 - IF, Q2 - JCR Best Q, 0.764 - SJR, Q1 - SJR Best Q). V databáze: DOI: 10.3390/s22166304 ; SCOPUS: 2-s2.0-85136733062 ; WOS: 000845515800001 ; CC: 000845515800001. Typ výstupu: článok; Výstup: zahraničný; Kategória publikácie do 2021: ADC, (2 citations in Scopus)
- PUTERKA, Boris - KAČUR, Juraj. Time window analysis for automatic speech emotion recognition. In Proceedings ELMAR-2018 : 60th International symposium. Zadar, Croatia. September 16-19, 2018. 1. ed. Zagreb : University of Zagreb, 2018, S. 143-146. ISSN 1334-2630. ISBN 978-953-184-244-0. V databáze: IEEE: 8534630 ; WOS: 000454262700033 ; SCOPUS: 2-s2.0-85058708605. (6 citations in Scopus)
- PUTERKA, Boris - KAČUR, Juraj - PAVLOVIČOVÁ, Jarmila. Windowing for speech emotion recognition. In Proceedings ELMAR-2019 : 61st International symposium. Zadar, Croatia. September 23-25, 2019. 1. ed. Zagreb : University of Zagreb, 2019, S. 147-150. ISSN 1334-2630. ISBN 978-1-7281-2182-6. V databáze: IEEE: 8918885209/URK/2019 ; WOS: 000534133200033 ; SCOPUS: 2-s2.0-85077554070. (3 citations in Scopus)
- STAROŇ, Matej - PUTERKA, Boris - KAČUR, Juraj. Vowel recognition based on formant frequencies. In Redžúr 2017 : 11th International workshop on multimedia information and communication technologies. Bratislava, Slovakia. May 19, 2017. 1. vyd. Bratislava : Vydavateľstvo Spektrum STU, 2017, S. 19-22. ISBN 978-80-227-4691-5.

Ostatné publikácie

- FARKAŠ, Peter - PUTERKA, Boris - STAROŇ, Matej - RUŽICKÝ, Eugen. Distributed local topology determination technique for topological interfence alignment in partially connected ad hoc networks. In IN-TECH 2015 : Proceedings of the International conference on innovative technologies. Dubrovnik, Croatia. 09.-11.09.2015. Rijeka : Engineering University of Rijeka, 2015, S. 327-330. ISSN 1849-0662.
- CSÓKA, Filip - PUTERKA, Boris - POLEC, Jaroslav. Real-time recognition of Slovak sign language. In ELITECH´17 [elektronický zdroj] : 19th Conference of doctoral students. Bratislava, Slovakia. May 24, 2017. 1. ed. Bratislava : Spektrum STU, 2017, CD-ROM, [5] p. ISBN 978-80-227-4686-1.
- CSÓKA, Filip - PUTERKA, Boris - POLEC, Jaroslav. SSL finger spelling recognition using Gabor wavelet and chamfer distance. In Redžúr 2017 : 11th International workshop on multimedia information and communication technologies. Bratislava, Slovakia. May 19, 2017. 1. vyd. Bratislava : Vydavateľstvo Spektrum STU, 2017, S. 11-14. ISBN 978-80-227-4691-5.

Riešené projekty

- MLbiomedia – Advanced methods of machine learning to design biometric and medical diagnostic systems. VEGA 1/0867/17
- Operačný program VÝSKUM A INOVÁCIE: Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov – II. etapa, Kód ITMS: 313021W404, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja, 2019-2023

Zoznam použitej literatúry

[1] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of german emotional speech. volume 5, pages 1517–1520, 09 2005.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, Dec 2008.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[5] Inc Google. Google's speech commands dataset. `https://pyroomacoustics.readthedocs.io/en/pypi-release/pyroomacoustics.datasets.google_speech_commands.html` [Accessed on July 2022].

[6] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA, 2015. Association for Computing Machinery.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Malcolm Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. 1997.

[12] Roy D. Patterson, Ken Robinson, John L. Holdsworth, D. McKeown, C. Q. Zhang, and Michael Allerhand. Complex sounds and auditory images. 1992.

[13] Karen Martin, Alexandra Bremner, Jo Salmon, Michael Rosenberg, and Billie Giles-Corti. Physical, policy, and sociocultural characteristics of the primary school environment are positively associated with children's physical activity during class time. *J. Phys. Act. Health*, 11(3):553–563, March 2014.

[14] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin, Heidelberg, 1999.

[15] W.W. Daniel. *Applied Nonparametric Statistics*. Duxbury advanced series in statistics and decision sciences. PWS-KENT Pub., 1990.

[16] J.A. Rice. *Mathematical Statistics and Data Analysis*. Advanced series. Cengage Learning, 2006.

[17] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016.

[18] Malcolm Slaney and Gerald McRoberts. Babyears: A recognition system for affective vocalizations. *Speech Communication*, 39:367–384, 02 2003.

[19] Cynthia Breazeal and Lijin Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12(1):83–104, Jan 2002.

[20] Nick Campbell. Databases of emotional speech. 01 2000.

[21] John H. L. Hansen and Sahar E. Bou-Ghazale. Getting started with susas: a speech under simulated and actual stress database. *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997.

[22] Shadi Langari, Hossein Marvi, and Morteza Zahedi. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*, 20:100424, 2020.

[23] S. Haq and P.J.B. Jackson. *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA, Aug. 2010.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface' 05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8, 2006.

[25] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, January 1997.

[26] Guo-Min Li, Na Liu, and Jun-Ao Zhang. Speech emotion recognition based on modified relieff. *Sensors*, 22(21), 2022.

[27] Jesus Alcala-Fdez, Alberto Fernández, Julián Luengo, J. Derrac, S Garc'ia, Luciano Sanchez, and Francisco Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17:255–287, 01 2010.

[28] Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer. *Glot international*, 5:341–345, 01 2001.

[29] Asma Mansour, Farah Chenchah, and Zied Lachiri. Emotional speaker recognition in real life conditions using multiple descriptors and i-vector speaker modeling technique. *Multimedia Tools and Applications*, 78(6):6441–6458, Mar 2019.

[30] Turgut Özseven. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. *Applied Acoustics*, 142:70–77, 2018.

[31] Wei Jiang, Zheng Wang, Jesse S. Jin, Xianfeng Han, and Chunguang Li. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors*, 19(12), 2019.

[32] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[33] Fatemeh Daneshfar and Seyed Jahanshah Kabudian. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimedia Tools and Applications*, 79(1):1261–1289, Jan 2020.

[34] Wu Chenjian, Huang Chengwei, and Chen Hong. Text-independent speech emotion recognition using frequency adaptive features. *Multimedia Tools and Applications*, 77:60–68, 2018.

[35] Eivind Kvedalen. Signal processing using the teager energy operator and other nonlinear operators. 2003.

[36] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017. Advances in Cognitive Engineering Using Neural Networks.

[37] Noam Amir, Samuel Ron, and Nathaniel Laor. Analysis of an emotional speech corpus in hebrew based on objective criteria. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[38] Shinichi Tokuno, Gentaro Tsumatori, Satoshi Shono, Eriko Takei, Taisuke Yamamoto, Go Suzuki, Shunnji Mituyoshi, and Makoto Shimura. Usage of emotion recognition in military health care. In *2011 defense science research conference and expo (DSR)*, pages 1–5. IEEE, 2011.

[39] Shunji Mitsuyoshi, Mitsuteru Nakamura, Yasuhiro Omiya, Shuji Shinohara, Naoki Hagiwara, and Shinichi Tokuno. Mental status assessment of disaster relief personnel by vocal affect display based on voice emotion recognition. *Disaster and military medicine*, 3:1–9, 2017.

[40] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.

[41] Amin Jalili, Sadid Sahami, Chong-Yung Chi, and Rassoul Amirfattahi. Speech emotion recognition using cyclostationary spectral analysis. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.

[42] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu. Efficient speech emotion recognition using multi-scale CNN and attention. *CoRR*, abs/2106.04133, 2021.

[43] Javier de Lope and Manuel Graña. An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11, 2023.

[44] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018.

[45] Andry Chowanda, Irene Anindaputri Iswanto, and Esther Widhi Andangsari. Exploring deep learning algorithm to model emotions recognition from speech. *Procedia Computer Science*, 216:706–713, 2023. 7th International Conference on Computer Science and Computational Intelligence 2022.

[46] J.L. Flanagan. *Speech Analysis, Synthesis and Perception.* Communication and Cybernetics. Springer Berlin Heidelberg, 2013.

[47] Lawrence R. Rabiner and Ronald W. Schafer. Introduction to digital speech processing. *Found. Trends Signal Process.*, 1:1–194, 2007.

[48] Gunnar Fant. *Acoustic Theory of Speech Production.* De Gruyter Mouton, Berlin, Boston, 1971.

[49] Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka. Speaker identification and verification by combining mfcc and phase information. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1085–1095, 2012.

[50] Popescu Theodor. Time-frequency analysis, by l. cohen, prentice hall signal processing series, prentice hall, englewood cliffs, new jersey, 1995 - book review. *Control Engineering Practice*, 5:292–294, 02 1997.

[51] Charles K Chui. *An introduction to wavelets*, volume 1. Academic press, 1992.

[52] P. M. J. Roberts. *Signals and Systems: Analysis Using Transform Methods & MATLAB.* McGraw-Hill Education, 2011.

[53] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages I–401, 2003.

[54] Prajakta P. Dahake, Kailash Shaw, and P. Malathi. Speaker dependent speech emotion recognition using mfcc and support vector machine. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pages 1080–1084, 2016.

[55] Pavol Harár, Radim Burget, and Malay Kishore Dutta. Speech emotion recognition with deep learning. pages 137–140, 02 2017.

[56] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[57] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, 1997. Available at www.dspguide.com.

[58] Pooja Mohindru, Rajesh Khanna, and Satvinder Singh Bhatia. Spectral analysis of generalized triangular and welch window functions using fractional fourier transform. *Automatika*, 57:221 – 229, 2016.

[59] Bassem R. Mahafza. Radar systems analysis and design using matlab. 2000.

[60] Juraj Kacur, Mario Varga, and Gregor Rozinaj. Speaker identification in a multimodal interface. In *Proceedings ELMAR-2013*, pages 191–194, 2013.

[61] Mustaqeem and Soonil Kwon. Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network. *Mathematics*, 8(12), 2020.

[62] Corina Albu, Eugen Lupu, and Radu Arsinte. Emotion recognition from speech signal in multilingual experiments. In Simona Vlad and Nicolae Marius Roman, editors, *6th International Conference on Advancements of Medicine and Health Care through Technology; 17–20 October 2018, Cluj-Napoca, Romania*, pages 157–161, Singapore, 2019. Springer Singapore.

[63] Margaret Lech, Melissa Stolar, Robert Bolia, and Michael Skinner. Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images. *Advances in Science, Technology and Engineering Systems Journal*, 3:363–371, 08 2018.

[64] Minjie Ren, Weizhi Nie, Anan Liu, and Yuting Su. Multi-modal correlated network for emotion recognition in speech. *Visual Informatics*, 3(3):150–155, 2019.

[65] Tursunov Anvarjon, Mustaqeem, and Soonil Kwon. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors*, 20(18), 2020.

[66] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875, 2020.

[67] Saikat Basu, Jaybrata Chakraborty, and Md. Aftabuddin. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, pages 333–336, 2017.

[68] Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan. Speaker diarization using latent space clustering in generative adversarial network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6504–6508, 2020.

[69] Finnian Kelly, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander. Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors. 2019.

[70] Alexandru-Lucian Georgescu and Horia Cucu. Gmm-ubm modeling for speaker recognition on a romanian large speech corpora. In *2018 International Conference on Communications (COMM)*, pages 547–551, 2018.

[71] YuJuan Xing, Ping Tan, and Xin Wang. Speaker verification normalization sequence kernel based on gaussian mixture model super-vector and bhattacharyya distance. *Journal of Low Frequency Noise, Vibration and Active Control*, 40:146134841988074, 12 2019.

[72] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[73] Yong Zhao, Tianyan Zhou, Zhuo Chen, and Jian Wu. Improving deep cnn networks with long temporal context for text-independent speaker verification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6834–6838, 2020.

[74] Zhongxin Bai and Xiao-Lei Zhang. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99, 2021.

[75] Sarthak Yadav and Atul Rai. Frequency and temporal convolutional attention for text-independent speaker recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6794–6798, 2020.

[76] Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang. Multi-resolution multi-head attention in deep speaker embedding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6464–6468, 2020.

[77] Qian-Bei Hong, Chung-Hsien Wu, Hsin-Min Wang, and Chien-Lin Huang. Statistics pooling time delay neural network based on x-vector for speaker verification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6849–6853, 2020.

[78] Ismael Kazheen and Adnan Mohsin Abdulazeez. Deep learning convolutional neural network for speech recognition: A review. 01 2021.

[79] Yanxi Tang, Jianzong Wang, Xiaoyang Qu, and Jing Xiao. Contrastive learning for improving end-to-end speaker verification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2021.

[80] Alakbar Valizada, Natavan Akhundova, and Samir Rustamov. Development of speech recognition systems in emergency call centers. *Symmetry*, 13(4), 2021.

[81] Wei Zhou, Wilfried Michel, Kazuki Irie, Markus Kitza, Ralf Schlüter, and Hermann Ney. The rwth asr system for ted-lium release 2: Improving hybrid hmm with specaugment, 2020.

[82] Mohammad Zeineldeen, Jingjing Xu, Christoph Lüscher, Wilfried Michel, Alexander Gerstenberger, Ralf Schlüter, and Hermann Ney. Conformer-based hybrid asr system for switchboard dataset, 11 2021.

[83] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, 2018.

[84] Jinyu Li. Recent advances in end-to-end automatic speech recognition, 2022.

[85] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457, Mar 2021.

[86] Peter Smit, Sami Virpioja, and Mikko Kurimo. Advances in subword-based hmm-dnn speech recognition across languages. *Computer Speech Language*, 66:101158, 2021.

[87] William Renda and Charlie H. Zhang. Comparative analysis of firearm discharge recorded by gunshot detection technology and calls for service in louisville, kentucky. *ISPRS International Journal of Geo-Information*, 8(6), 2019.

[88] Hanne Lyngholm Larsen, Cino Pertoldi, Niels Madsen, Ettore Randi, Astrid Vik Stronen, Holly Root-Gutteridge, and Sussie Pagh. Bioacoustic detection of wolves: Identifying subspecies and individuals by howls. *Animals*, 12(5), 2022.

[89] Justin Salamon, Juan Bello, Claudio Silva, Oded Nov, R. DuBois, Anish Arora, Charlie Mydlarz, and Harish Doraiswamy. Sonyc: A system for the monitoring, analysis and mitigation of urban noise pollution. *Communications of the ACM*, 05 2018.

[90] René Grzeszick, Axel Plinge, and Gernot A. Fink. Bag-of-features methods for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1242–1252, 2017.

[91] Sajjad Abdoli, Patrick Cardinal, and Alessandro Lameiras Koerich. End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications*, 136:252–263, 2019.

[92] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society.

[93] Sungho Shin, Jongwon Kim, Yeonguk Yu, Seongju Lee, and Kyoobin Lee. Self-supervised transfer learning from natural images for sound classification. *Applied Sciences*, 11(7), 2021.

[94] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993.

[95] David Gerhard. Audio signal classification: History and current techniques. 2003.

[96] Vishal H. Shah and Mahesh Chandra. Speech recognition using spectrogram-based visual features. In Srikanta Patnaik, Xin-She Yang, and Ishwar K. Sethi, editors, *Advances in Machine Learning and Computational Intelligence*, pages 695–704, Singapore, 2021. Springer Singapore.

[97] Tomás Arias-Vergara, Tomás Arias-Vergara, Tomás Arias-Vergara, Philipp Klumpp, J. C. Vásquez-Correa, J. C. Vásquez-Correa, Elmar Nöth, Juan Rafael Orozco-Arroyave, Juan Rafael Orozco-Arroyave, and Maria Schuster. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, 24:423 – 431, 2020.

[98] Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam, Ankur Dumka, Rajesh Singh, Mamoon Rashid, Anita Gehlot, Sultan S. Alshamrani, and Ahmed Saeed AlGhamdi. Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network. *Applied Sciences*, 12(12), 2022.

[99] Kyu J. Han, Jing Pan, Venkata Krishna Naveen Tadala, Tao Ma, and Dan Povey. Multistream cnn for robust acoustic modeling, 2021.

[100] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.