Razpoznavanje čustvenih stanj govorcev in razpoznavanje varnostno sumljivih zvokov

Praktični projekt pri predmetu Govorne tehnologije

Jernej Sabadin

Univerza v Ljubljani
Fakulteta za elektrotehniko

Mentorja: izr. prof. dr. Simon Dobrišek, univ. dipl. inž. el. in univ. dipl. inž. el., As. Marija Ivanovska, mag.

`Predmet: Govorne tehnologije`
`Datum: 1. December 2023`

Kazalo vsebine

Ključne besede:

---

1 Uvod

V okviru praktičnega projekta iz Govornih tehnologij se osredotočamo na dve ključni področji: razpoznavanje čustvenih stanj govorcev in detekcijo varnostno sumljivih zvokov. Te naloge predstavljajo sofisticiran izziv, ki združuje področja strojnega učenja in obdelave zvočnih signalov. Za učinkovito klasifikacijo tako čustvenih stanj kot sumljivih zvokov uporabljamo hibridni model, ki kombinira dolgoročni kratkotrajni spomin (Long Short-Term Memory - LSTM) z konvolucijskimi nevronskimi mrežami (CNN). Ta pristop nam omogoča visoko natančnost pri razpoznavanju kompleksnih vzorcev. Dodatno, za povečanje robustnosti in generalizacije modela, razširjamo učno zbirko z vključitvijo perturbiranih vzorcev zvoka, kar prispeva k izboljšanju natančnosti in zanesljivosti sistema pri razpoznavanju.

Razpoznavanje čustvenih stanj govorcev

Čustvena inteligenca umetnih sistemov je ključnega pomena za izboljšanje interakcije med človekom in strojem. Razpoznavanje čustev iz govora omogoča napravam, da se odzivajo na človeška čustvena stanja in s tem postanejo bolj empatični in učinkoviti komunikacijski partnerji.

Razpoznavanje varnostno sumljivih zvokov

Varnost je pomembna skrb v današnji družbi. Sposobnost avtomatskega razpoznavanja sumljivih zvokov v realnem času lahko pripomore k hitrejšemu in učinkovitejšemu odzivanju na morebitne nevarnosti.

Uporabljene tehnologije

Za dosego teh ciljev uporabljamo Mel-spektrograme, ki so učinkovite značilke pridobljene iz zvokov. V kombinaciji z naprednimi tehnologijami strojnega učenja, kot so konvolucijska nevronska omrežja (CNN) in dolgoročni kratkotrajni spomin (LSTM), razvijamo modele, ki lahko prepoznajo subtilne vzorce v podatkih.

V naslednjih poglavjih bomo podrobneje opisali uporabljene metode in arhitekture ter prikazali rezultate našega raziskovanja.

2 Mel-spektrogrami

Mel-spektrogram je vizualizacija spektra frekvenc zvočnega signala, preslikanega na Mel frekvenčno lestvico. Zaradi nelinearne narave človeškega zaznavanja frekvenc, Mel-spektrogram bolje odraža zaznavo zvokov, saj Mel lestvica posnema zaznavno občutljivost človeškega ušesa, ki ni enakomerno občutljiva na celotnem frekvenčnem območju.

Nelinearno zaznavanje frekvenc: Človeško uho je bolj občutljivo na spremembe frekvence na nižjih frekvencah kot na višjih. Mel lestvica upošteva to nelinearno zaznavanje z uporabo logaritemske skale.

$$f_{\text{Mel}} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

kjer je f frekvenca v Hz in f_{Mel} je zaznana frekvenca na Mel lestvici.

Postopek izračuna Mel-spektrograma:

1. Zvočni signal razdelimo na kratke časovne segmente z oknenjem.
2. Za vsak segment signala, ki smo ga pridobili z oknenjem, izračunamo njegovo transformiranko. To storimo z DFT.

$$F(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-i \frac{2\pi n k}{N}}, \qquad 0 \leq k \leq N-1, \quad (2)$$

   kjer je N število točk, ki jih uporabimo za izračun DFT.
3. Frekvence preslikamo na Mel lestvico z uporabo melodičnih filtrov.
4. Spekter vsakega segmenta nato predstavimo v barvni lestvici, kjer intenziteta barve predstavlja amplitudo določene frekvence.

Melodični filtri: Melodični filtri se uporabljajo za preslikavo frekvenc v spektru zvoka na Mel lestvico. Ti filtri so ključni za pridobivanje značilnosti zvoka, ki so relevantne za človeško zaznavanje.

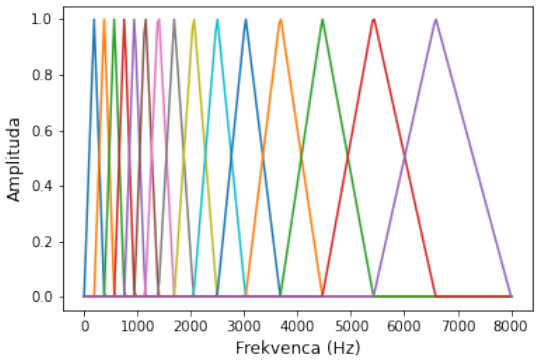

Fig. 1: Melodični filtri

Uporaba Mel-spektrogramov: Mel-spektrogrami se pogosto uporabljajo v sistemih za avtomatsko razpoznavanje govora in drugih aplikacijah za analizo zvoka, saj omogočajo bolj intuitivno razumevanje spektralnih značilnosti zvoka.

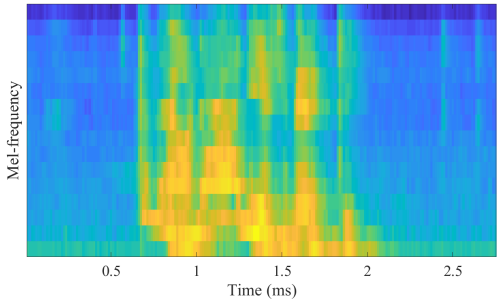

Fig. 2: Primer Mel-spektrograma iz vira [2]

3 Modeli strojnega učenja

V tem poglavju predstavimo modela strojnega učenja, ki ju uporabimo v našem raziskovalnem projektu: konvolucijskih nevronskih mrež (CNN) in dolgoročnega kratkotrajnega spomina (LSTM). Oba modela igrata pomembno vlogo pri klasifikaciji zvočnih signalov, pri čemer vsak pristopa k problemu na svoj način.

3.1 Konvolucijske Nevronske Mreže (CNN)

Konvolucijske nevronske mreže (CNN) so specializirane za obdelavo podatkov z izrazito prostorsko strukturo, kot so slike ali zvočni signali. Značilnost CNN je njihova sposobnost učenja vzorcev, neodvisno od njihove točne lokacije v vhodnem prostoru, kar dosežejo z uporabo konvolucijskih operacij.

Povezovanje Nevronov: V CNN so nevroni organizirani v plasti, kjer vsaka plast izvaja specifično funkcijo. V konvolucijskih plasteh so nevroni lokalno povezani s svojimi vhodnimi podatki, kar pomeni, da vsak nevron prejme vhod samo iz omejenega območja (imenovanega receptivno polje) prejšnje plasti ali vhodnih podatkov.

Delitev Uteži: Ključna značilnost konvolucijskih plasti je, da iste uteži (imenovane tudi filtri ali jedra) delujejo na različnih delih vhodnih podatkov, kar omogoča zaznavanje istih vzorcev na različnih mestih. To vodi do bistvenega zmanjšanja števila učljivih parametrov v primerjavi s polno povezanimi plastmi.

Arhitektura CNN: CNN običajno vključujejo več ključnih plasti:

- *Konvolucijske plasti* zaznavajo lokalne vzorce z uporabo filtrov.
- *Aktivacijske funkcije*, kot je ReLU, uvedejo nelinearnosti v model.
- *Plasti združevanja* (pooling) zmanjšujejo prostorske dimenzije.
- *Polno povezane plasti* izvajajo klasifikacijo ali regresijo.

Konvolucijski Sloj: Konvolucijske operacije so opisane z naslednjo enačbo:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) K(i-m, j-n) \quad (3)$$

kjer I predstavlja vhodno matriko, K je konvolucijski filter in $S(i,j)$ so aktivacijske vrednosti na izhodu.

Učenje v CNN: Proces učenja v konvolucijskih nevronskih mrežah (CNN) je osredotočen na optimizacijo uteži filtrov in uteži v polno povezanih plasteh. To dosežemo z minimizacijo kriterijske funkcije, ki ocenjuje razliko med napovedmi mreže in dejanskimi vrednostmi. Obstaja več različnih kriterijskih funkcij, kot so križna entropija ali srednja kvadratna napaka, ki so primerne za različne tipe problemov.

Pri optimizaciji nelinearnih kriterijskih funckij se običajno začne z algoritmom gradientnega spusta, ki je temelj mnogih drugih naprednih metod. V najbolj osnovni obliki je postopek posodabljanja uteži z algoritmom gradientnega spusta opisan z naslednjo enačbo:

$$W_{new} = W_{old} - \eta \frac{\partial \mathcal{L}}{\partial W} \quad (4)$$

kjer η predstavlja stopnjo učenja, $\mathcal{L}$ je kriterijska funkcija, in $\frac{\partial \mathcal{L}}{\partial W}$ je gradient kriterijske funkcije glede na uteži W.

Algoritmi, kot so stohastični gradientni spust (SGD), Adam, Adagrad, in RMSprop, gradijo na osnovnem konceptu gradientnega spusta, a uvajajo dodatne mehanizme.

Pomembno je omeniti, da je algoritem povratnega širjenja napake (backpropagation) standardna metoda za izračun gradientov v večslojnih nevronskih mrežah, kar omogoča uspešno učenje globokih arhitektur, kot je CNN.

Zaradi lokalne povezanosti in delitve uteži so CNN še posebej učinkovite pri zaznavanju vzorcev in redukciji dimenzij, kar omogoča učinkovito obdelavo kompleksnih prostorskih podatkov.

3.2 Dolgoročni Kratkoročni Pomnilnik (LSTM)

Dolgoročni kratkoročni pomnilnik (LSTM) je specializirana vrsta rekurentnih nevronskih mrež (RNN), namenjena obdelavi dolgih sekvenčnih podatkov. Te mreže so oblikovane za reševanje problema izginjajočega in eksplozivno naraščajočega gradienta, ki se pojavi pri standardnih RNN, in so učinkovite pri učenju odnosov v dolgih zaporedjih.

Struktura LSTM celice: Vsaka LSTM celica vsebuje tri regulativna "vrata" – vhodna, izhodna in vrata za pozabo. Ta vrata omogočajo celici, da modulira tok informacij skozi čas, kar povečuje njeno sposobnost ohranjanja pomembnih informacij in pozabljanja nepomembnih.

- **Vhodna vrata (Input gate):** Določajo, katere nove informacije se bodo dodale v celično stanje.
- **Pozabljiva vrata (Forget gate):** Določajo, katere informacije iz preteklega celičnega stanja se bodo zavrgle ali ohranile.
- **Izhodna vrata (Output gate):** Nadzorujejo, katere informacije iz celičnega stanja se bodo uporabile za ustvarjanje izhoda mreže.

Celično Stanje (Cell state): Je kot "nosilec" informacij skozi celotno zaporedje, ki omogoča ohranjanje informacij na dolgi rok. Celično stanje je ključno za zmožnost LSTM, da ohranja in prenaša informacije skozi dolga zaporedja brez problema izginjajočega ali eksplodirajočega gradienta.

Povezovanje celic: Celice v LSTM mreži so povezane zaporedno, kjer izhod ene celice vhod za naslednjo. To zaporedno povezovanje omogoča mreži, da prenaša informacije skozi čas in razvija notranji "spomin", ki lahko zajema kompleksne časovne odvisnosti. Ključna lastnost LSTM je, da vse celice v mreži uporabljajo isti nabor uteži (deljenje uteži), kar pomeni, da se iste uteži uporabljajo na vsakem časovnem koraku.

Matematične enačbe LSTM: Delovanje posamezne LSTM celice je opredeljeno z naslednjimi enačbami:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{7}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = o_t \odot \tanh(C_t) \tag{10}$$

kjer σ je sigmoidna funkcija, tanh je hiperbolični tangens, W in b predstavljajo uteži in pragove, f_t, i_t, in o_t so vrata za pozabo, vhodna in izhodna vrata, C_t je stanje celice, h_t je izhodni vektor, in x_t je vhod na časovnem koraku t.

Postopek učenja: Učenje v LSTM mrežah vključuje prilagajanje uteži in pragov v vratih in povezavah celic za optimizacijo določene kriterijske funkcije. To se običajno doseže z uporabo algoritma povratnega širjenja skozi čas (BPTT), ki je posebej prilagojen za obdelavo časovnih zaporedij. BPTT upošteva zaporedno naravo podatkov in posodablja uteži tako, da minimizira napako v celotnem zaporedju. Ta postopek omogoča LSTM mrežam, da se učinkovito prilagajajo in izboljšujejo pri modeliranju dolgoročnih odvisnosti v podatkih.

Dinamika mreže: Dinamika mreže LSTM omogoča celicam, da ohranjajo relevantne informacije skozi daljše časovne razpone in učinkovito obdelujejo zaporedne podatke, kar je ključno pri naprednih nalogah, kot so jezikovno modeliranje, generiranje besedila, strojno prevajanje in prepoznavanje govora.

4 Uporabljena metoda 2D CNN in LSTM omrežja

Arhitektura uporabljena v tem delu združuje 2D konvolucijske nevronske mreže (CNN) z dolgoročnim spominom (LSTM) za naloge prepoznavanja čustev iz govora ali razpoznavanje varnostno sumljivih zvokov. Metoda se zgleduje po članku J. Zhaa et al. [3]. Ključne komponente vključujejo:

Blok za učenje lokalnih značilnosti (LFLB): Sestavljen iz konvolucijskih, normalizacijskih, eksponencialnih linearnih enot in slojev za maksimalno združevanje, je namenjen izvlečenju lokalnih značilnosti iz vhodnih podatkov.

Globalno učenje značilnosti preko LSTM: Po bloku LFLB se z uporabo plasti LSTM uči dolgoročne odvisnosti iz zaporedja lokalnih značilnosti.

Obdelava mel-spektrogramskih slik: Omrežje 2D CNN-LSTM zajema tako lokalne korelacije kot globalne kontekstualne informacije iz log-mel spektrogramskih slik.

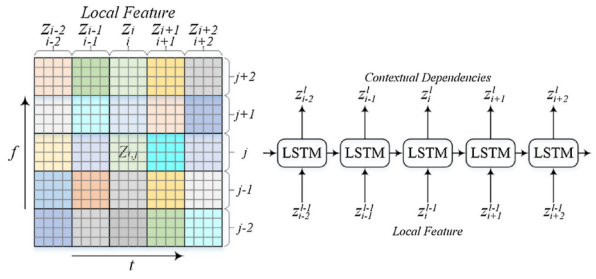

Fig. 3: Slika je vzeta iz vira [3]. Na prikazanem mel-spektrogramu vsak segment z_i predstavlja lokalno značilnost v določenem času t in frekvenci f. Te značilnosti so zaporedno vnesene v omrežje LSTM, kar modelu omogoča učenje časovnih odnosov znotraj podatkov.

5 Mere za vrednotenje sistemov razpoznavanja

Vrednotenje učinkovitosti sistemov za razpoznavanje temelji na merah, kot so natančnost, priklic, F1-score in točnost. Natančnost (Precision) meri delež pravilno identificiranih primerov med vsemi identificiranimi primeri, priklic (Recall) pa delež pravilno identificiranih primerov med vsemi relevantnimi primeri. F1-score je harmonično povprečje natančnosti in priklica, ki zagotavlja ravnovesje med obema merama in je še posebej koristen v primerih, kjer sta oba aspekta

- natančnost in priklic - enako pomembna. Visok F1-score pomeni, da sistem uspešno združuje visoko natančnost in priklic, kar je pogosto zaželeno v mnogih aplikacijah. Točnost (Accuracy) meri delež vseh pravilno klasificiranih primerov (tako pozitivnih kot negativnih) glede na vse primere. Matematično so te mere opredeljene kot:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

6 Opis podatkovne zbirke

Prvi eksperiment izvedemo na zbirki posnetkov čustvenega govora emoDB. Zbirka *emoDB* (Berlin Database of Emotional Speech) vsebuje 535 posnetkov čustvenega govora, izvedenih s strani 10 govorcev (5 moških in 5 žensk). Posnetki zajemajo sedem različnih čustvenih stanj: veselje, žalost, jeza, strah, gnus, presenečenje in nevtralno stanje.

Drugi eksperiment pa izvedemo na zbirki varnostno sumljivih zvokov. Zbirka je sestavljena iz 2524 varnostno sumljivih zvokov, ki so v povprečju dolgi okoli 5 sekund in razdeljeni v 7 razredov. Razredi so alarm, pasji lajež, eksplozija, lomljenje stekla, kričanje, streljanje in sirena.

7 Eksperiment razpoznavanja čustvenih stanj govorcev

Iz posnetkov zbirke emoDB izluščimo mel-spektogramske slike in jih podamo modelu 2D CNN v kombinaciji z LSTM. Podatke razdelimo na učne (4/5 vseh podatkov), validacijske (1/10 vseh podatkov) in testne (1/10 vseh podatkov). Učni podatki se uporabljajo za treniranje modela, kjer se model "uči" iz podatkov. Validacijski podatki služijo za fino nastavitev parametrov modela in preprečevanje prenaučenosti, to je, da bi se model preveč prilagodil samo na učne podatke in ne bi bil generalno uporaben. Testni podatki so ločeni od učnih in validacijskih in se uporabljajo za končno oceno uspešnosti modela, saj predstavljajo neznane podatke, na katerih se preveri, kako dobro model deluje na primerih, ki jih med učenjem ni videl.

Rezultate učenja prikažemo na slednji sliki, ki prikazuje kako so se skozi iteracije spreminja vrednost kriterijske funkcije (ang. loss) in pa natančnost razppoznavanja (ang. accuracy)

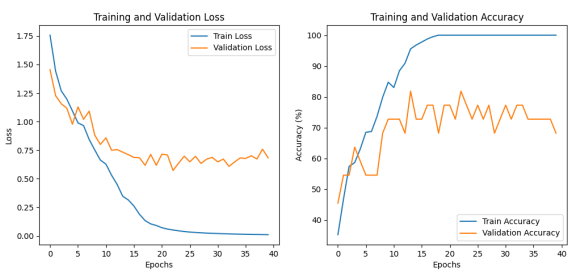

Fig. 4: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Opazna je prenaučenost modela na učnih podatkih, ki se kaže v stopnji razpoznavanja 100% na učnih podatkih, medtem pa je generalizacija na validacijske podatke veliko slabša. Znaša kar 20 % manj. Po učenju izrišemo konfuzijsko matriko testnih podatkov:

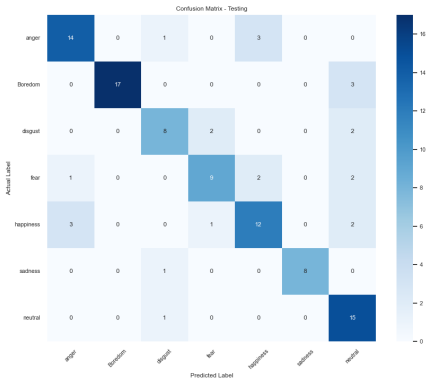

Fig. 5: Konfuzijska matrika prikazuje na y osi informacijo o dejanskih fonemih na x osi pa kako so razpoznani

Stopnja pravilnega razpoznavanja na testnih podatkih znaša 74%. Natančnost znaša 0.79, priklic 0.78 in F1 score 0.77. Rezultati testiranja modela kažejo na razmeroma dobro delovanje, s stopnjo pravilnega razpoznavanja 74%. To pomeni, da model učinkovito obravnava kompleksne naloge, čeprav še vedno obstaja prostor za izboljšave. Natančnost 0.79 in priklic 0.78 skupaj kažeta, da model ne le dobro loči med različnimi kategorijami, temveč tudi uspešno identificira relevantne primere. F1 score 0.77 dodatno potrjuje uravnoteženost med natančnostjo in priklicem, kar je ključno za aplikacije, kjer sta oba aspekta pomembna. Kljub temu ti rezultati nakazujejo na možnosti za nadaljnje izboljšave modela, zlasti v smislu zmanjševanja števila napačno klasificiranih primerov.

Ker je model pre naučen povečamo zbirko učnih podatkov tako, da originalnim učnim podaktom dodamo šum, spremenimo osnovno frekvenco ali pa jih raztegenemo pa časovni osi. Parametre perturbacij izbiramo naključno znotraj določenih intervalov vrednosti. Za učenje uporabimo poleg originalnih tudi perturbirane učne podatke. Novo nastala zbirka je tako za faktor 4 večja od prvotne. Testne in validacijske zbirke ne spreminjamo.

Po ponovnem učenju dosežemo naslednje rezultate

Fig. 6: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Po učenju na večji množici podatkov, je model sposoben bolje razvrščati tudi validacijske vzorce in sicer z stopnjo razpoznavanja 100%. Natančnost, priklic in F1-score znašajo 1.

Po učenju izrišemo konfuzijsko matriko testnih podatkov:

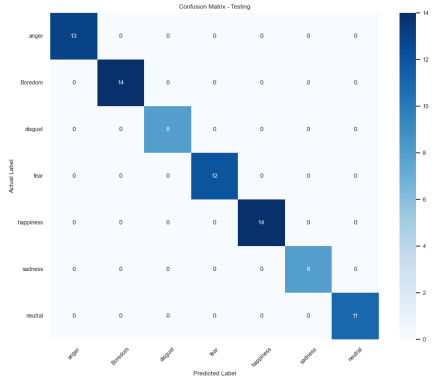

Fig. 7: Konfuzijska matrika prikazuje na y osi informacijo o dejanskih fonemih na x osi pa kako so razpoznani

8 Eksperiment razpoznavanja varnostno sumljivih zvokov

Podatke iz zbirke varnostno sumljivih zvokov razdelimo na učne (4/5 vseh podatkov), validacijske (1/10 vseh podatkov) in testne (1/10 vseh podatkov). Učno množico razširimo z prej omenjenimi perturbacijami.

Iz posnetkov izluščimo mel-spektogramske slike in jih podamo modelu 2D CNN v kombinaciji z LSTM. Po učenju dosežemo naslednje rezultate

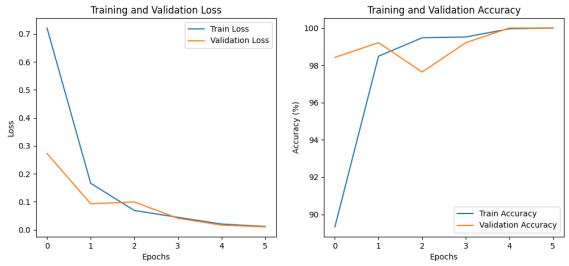

Fig. 8: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Model sposoben razvrščati tudi validacijske vzorce z 100% stopnjo razpoznavanja.

Po učenju izrišemo konfuzijsko matriko testnih podatkov:

Stopnja razpoznavnja na testnih podatkih znaša 100%. Natančnost, priklic in F1-score znašajo prav tako 1.

Dobljeni rezultati so boljši od uporabe CNN na spektogramksih slikah izluščenih iz zvočnih signalov, kjer dosežemo stopnjo razpoznavanja testnih podatkov 94%.

9 Zaključek

V zaključku lahko povzamemo, da predstavljeni model, ki združuje konvolucijsko nevronsko mrežo in LSTM, kaže obetavne rezultate v razpoznavanju čustvenih stanj govorcev in varnostno sumljivih zvokov. Modelova sposobnost prilagajanja različnim aplikacijam poudarja njegovo prenosljivost in potencial za široko uporabo. Dodatno, raziskava pokaže, da količina in kakovost učnih podatkov ključno vplivata na

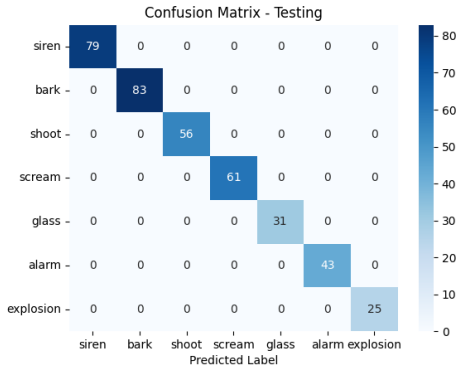

Fig. 9: Konfuzijska matrika prikazuje na y osi informacijo o dejanskih fonemih na x osi pa kako so razpoznani

natančnost in zanesljivost modela, kar nakazuje potrebo po obsežnih in raznolikih podatkovnih zbirkah za nadaljnje izboljšanje modelov.

10 Reference

1 D. M. Low, K. H. Bentley, S.S. Gosh, Automated assessment of psychiatric disorders using speech: A systematic review, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC704265` (Datum ogleda: 20. 10. 2023)

2 T. Bäckström, Cepstrum and MFCC, `https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC`

3 J. Zhao a, X. Mao a, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, `https://www.sciencedirect.com/science/article/pii/S1746809418302337`. (Datum ogleda: 15. 12. 2023)