Politechnika Śląska

Wydział Automatyki, Elektroniki i Informatyki

Systemy Interaktywne i Multimedialne
Projekt
Detekcja emocji w głosie

Natalia Stręk Jakub Kula, Paweł Wójtowicz,

Gliwice 2023

1 Analiza wyników i wnioski

1.1 Cele i zadania z okresu maczec-kwiecień

Cele wyznaczone na okres marzec-kwiecień obejmowały przede wszystkim zgłębienie wiedzy na temat przetwarzania plików audio oraz zidentyfikowanie cech, które można wykorzystać w procesie nauki modelu uczenia maszynowego. W szczególności chodziło o poznanie technik ekstrakcji cech i analizy dźwięku, aby lepiej zrozumieć, jakie elementy są istotne z punktu widzenia efektywności modelu. Następnym etapem było stworzenie odpowiedniego skryptu do przetwarzania danych, który umożliwiłby ekstrakcję wcześniej zidentyfikowanych charakterystycznych cech. Ostatnim postawionym celem na ten okres było stworzenie modelu sieci neuronowej posiadajacej minium 50% dokładności, który by służył jako model referencyjny do którego będą porónywane kolejne stworzone modele.

1.2 Opis zadań przyjętych do realizacji

Zgromadzenie i przetworzenie danych

Aby stworzyć model klasyfikacji emocji z dźwięku, wykorzystano dane z dwóch źródeł: Ravdess i TESS. Pierwotnie planowano wykorzystać wyłącznie zbiór Ravdess, jednak w celu zwiększenia różnorodności i objętości danych, do projektu włączono również zbiór TESS.

Wybór oraz ekstrakcja cech

W trakcie okresu marzec-kwiecień zostały wybrane cechy dziwięku które zostały użyte jako wejścia do sieci neuronowej. Wybrane wejścia do sieci to:

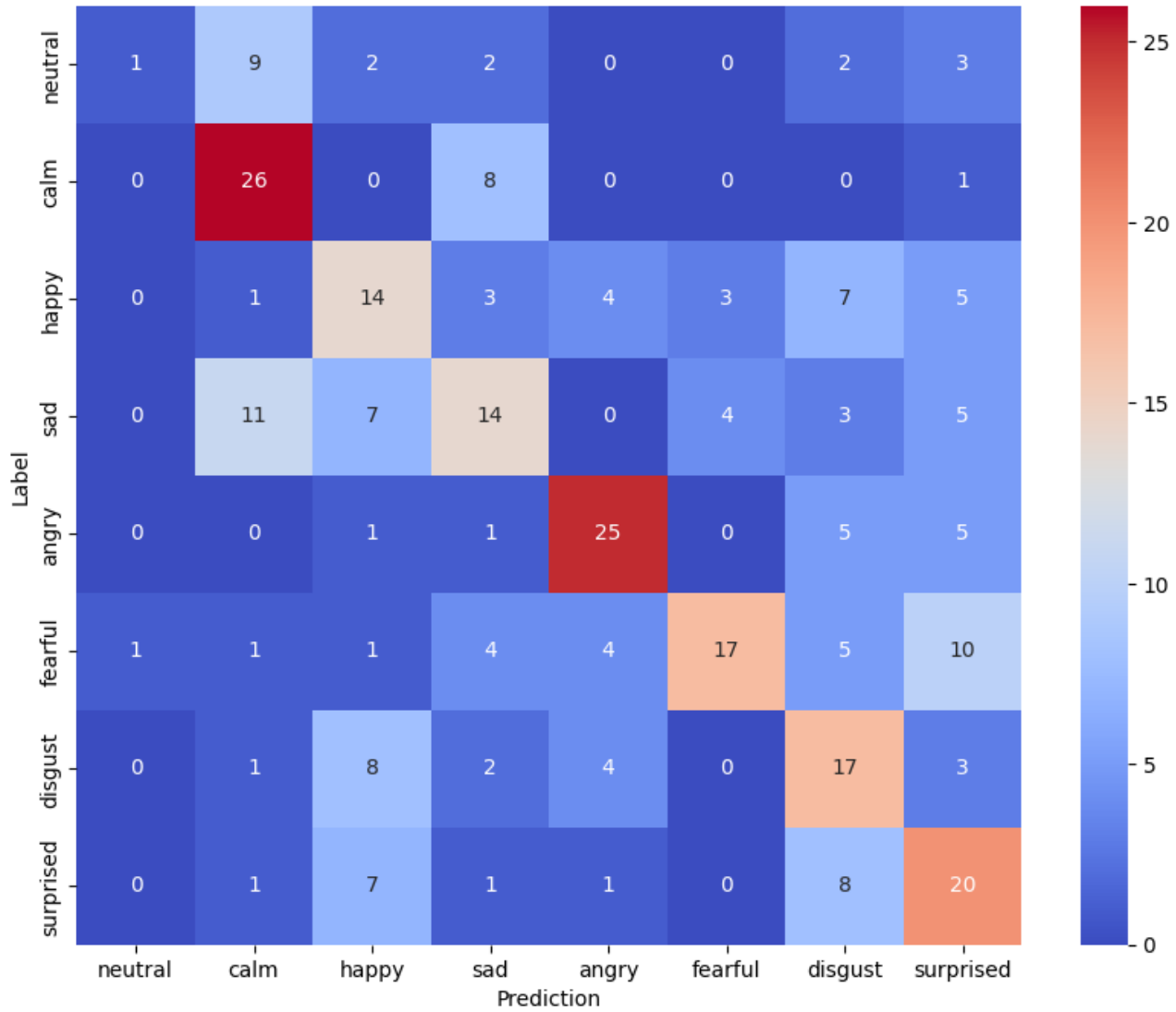
- **Chroma STFT**: Reprezentacja częstotliwości dźwięku w sposób związany z tonalnością, używana w analizie akordów i melodii.
- **MFCC**: Kompaktowa reprezentacja dźwięku, oparta na skali melowej, często stosowana w rozpoznawaniu mowy i dźwięku.
- **Root Mean Square Value**: Średnia wartość kwadratu sygnału dźwiękowego, stosowana do oceny głośności.
- **Spectral Centroid**: Średnia częstotliwość w spektrum dźwięku, służy do oceny barwy dźwięku.
- **Spectral Spread**: Mierzona rozpiętość częstotliwości w spektrum dźwięku, informuje o jego różnorodności częstotliwości.
- **Spectral Flux**: Miara zmienności spektrum dźwięku w czasie, używana w analizie dynamiki dźwięku.
- **Spectral Roll-Off**: Częstotliwość, poniżej której kumulatywna energia spektrum jest mniejsza niż określony procent całkowitej energii.
- **Chroma Vector**: Wektor reprezentujący rozkład mocy dźwięku w chromie, używany w analizie harmonicznej.
- **MelSpectrogram**: Spektrogram dźwięku, gdzie skala częstotliwości jest przekształcona na skalę melową, co odzwierciedla sposób, w jaki ludzkie ucho percepcyjnie odbiera dźwięki.

W celu zmniejszenia wymiarowości danych, została wyciągnięta średnia z każdej cechy z każdego segmentu czasowego.

Wybór architektury sieci oraz trenowanie modelu

W realizowanym projekcie wykorzystano konwolucyjne sieci neuronowe (CNN), architektura ta wykorzystuje takie warstwy jak konwolucyjne do wykrywania wzorców, poolingowe do redukcji wymiarów oraz z warst gęstych do końcowej klasyfikacji.

Została także wytrenowana sieć która osiąga 68% dokładności. Jendak podczas testów na zbiorze testowym osiąga ta sieć dokładność rzędu 45%, co wskazuje na znaczącą różnicę w wydajności między zbiorem uczącym a testowym.



Do tej pory przeprowadzono wybór danych, ich przetwarzanie oraz ekstrakcję cech. Dodatkowo dokonano doboru architektury sieci neuronowej oraz stworzono modele uczenia maszynowego, które klasyfikują nagrania głosowe. Modele zostały poddane ocenie i walidacji w celu sprawdzenia ich dokładności i skuteczności w przewidywaniu oraz klasyfikacji. Ze względu na nieustanne poszukiwanie coraz doskonalszych modeli, proces ten pozostaje w ciągłej realizacji, dążenie do optymalizacji modelu wymaga ciągłego testowania oraz wprowadzania ulepszeń w celu osiągnięcia jak najskuteczniejszego modelu w klasyfikacji nagrań. Na ten moment szacujemy, że około 60% projektu zostało już zrealizowane.

Projekt jest realizowany zgodnie z harmonogramem.

1.3 Dodatkowy zbiór danych

Wprowadzono następujące zmiany w założeniach projektu: Dodano kolejny zbiór nagrań. Zbiór TESS Toronto emotional speech set posiada aż 2800 nagrań podzielonych na 6 klas. Do projektu użytko tylko nagrań audio. Klasy to: neutral, happy, sad, angry, fearful, surprise i disgust. Zbiór TESS nie posiada emocji calm który posiadał zbiór Ravdess. Wykorzystanie dodatkowego zbioru pozwala na wprowadzenie większej różnorodności dlatego też model ten może lepiej osiągąć lepszą dokładność w klasyfikacji nowych danych.
Kolejną zmianą jest zmniejszenie długości nagrań głosowych do jednej sekundy. Zostało to zastosowane ze względu na to, że wszystkie nagrania muszą mieć taką samą długość, a dane z nowego zbioru są znacząco krótsze. Dane ze zbioru Ravdess zostały użyte od 0.6s do 1.6s.