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

Celami wybranymi na okres marzec-kwiecień było napisanie wykonanie odpowiedniego research na temat przetwarzania ludzkiego głosu, oraz parametrów go charakteryzujących. Kolejnym celem było skryptu przetwarzającego dane, wyciagającego opisane charakterystyczne parametry. Ostatnim celem na okres było stworzenie modelu sieci neuronowej posiadajacej minium 50% dokładności. Jest to model referencyjny do którego będą porónywane kolejne stworzone modele.

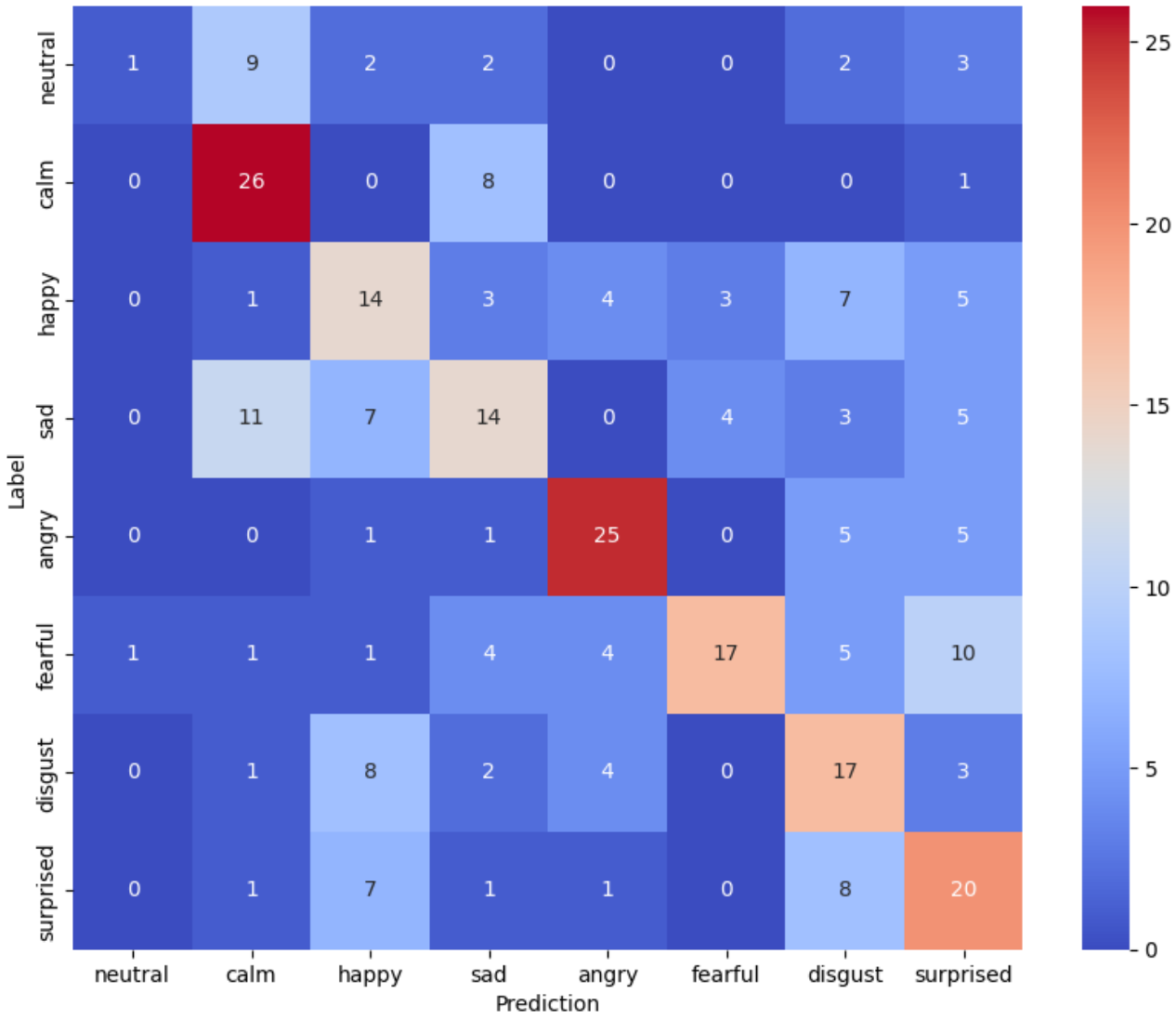
W trakcie okrsu marzec-kwiecień zostały wybrane cechy dziwięku które zostały użyte jako wejścia do sieci neuronowej.

Wybrane parametry sieci:

- **Chroma STFT**: Reprzentacja 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/ilości wejść do sieci, została wyciągnięta średnia z każdego parametru z każdego segmentu czasowego.

Została także wytrenowana sieć podstawowa. Sieć tak osiąga 68% dokładności na zbiorze uczącym oraz 45% na zbiorze testującym.



Projekt jest realizowany zgodnie z harmonogramem.

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.
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.