Zusammenfassung

Das Ziel dieser Arbeit ist es, eine Anwendung zu erstellen, die in der Lage ist, Emotionen in gesprochenen Aussagen zu erkennen. Dafür wird die Aussage zuerst transkribiert und das Sentiment des daraus resultierenden Textes analysiert. Außerdem wird mithilfe von Techniken der Audiosignalverarbeitung ein Modell trainiert, das Emotionen erkennt. Ein Teil des Audiomodells ist das lokale Modell, für welches das Audio zuerst in kleine Frames zerschnitten wird und spektrale Features extrahiert werden. Der andere Teil ist das globale Modell, in dem Features der Kontur der Grundfrequenz sowie Features aus den Vorhersagen des lokalen Modells kombiniert werden und eine Vorhersage erstellt wird. Text- und Audiomodell werden anschließend kombiniert, um finale Vorhersagen für die Emotionen glücklich und traurig zu erhalten. Das Modell wird auf neun Datensätzen trainiert und evaluiert. Drei davon sind Textdatensätze, fünf sind Audiodatensätze und ein Datensatz enthält Audios, in denen auch der Inhalt der jeweiligen Emotion entspricht. Zwei dieser Datensätze wurden für die Arbeit erstellt, von YouTube und über WhatsApp Sprachnachrichten. Das Modell liefert eine Vorhersagegenauigkeit von über 75% auf Audiodaten mit unbekannten Inhalten, Sprechern, Mikrophonen und Aufnahmebedingungen. Diese Genauigkeit macht es möglich, die emotionale Verfassung eines Sprechers zu bestimmen, insbesondere wenn mehrere aufeinanderfolgende Äußerungen analysiert werden. Es wurde eine Anwendung entwickelt, die das trainierte Modell verwendet, um Vorhersagen für Audiodateien und Mikrofoneingaben zu bestimmen.

Abstract

The goal of this paper is to build an application that can recognize emotions based on vocal utterances. For this, an utterance is transcribed and the sentiment of the resulting text is analyzed. Further, with the help of audio signal processing, a model is trained that recognizes emotions. One part of the audio model is the local model, for which first the audio is cut into short frames and spectral features are extracted. The other part is the global model, where features of the pitch contour of the utterance as well as features of local model predictions are combined into global prediction scores. The text and audio model are then again combined to deliver final predictions for the emotions happy and sad. The model is trained and evaluated on nine datasets, three of which are text datasets, five are audio datasets and one dataset contains emotional audio where the content corresponds to the emotion. Two of the datasets were collected for this thesis, from YouTube and via WhatsApp voice messages respectively. The model delivers an accuracy above 75% on audio data with unknown contents, speakers, microphones and recording conditions. This performance is good and makes it possible to determine the emotional state of a speaker, especially when a series of utterances is analyzed. An application was developed that utilizes the trained model to deliver predictions for audio files as well as microphone input.