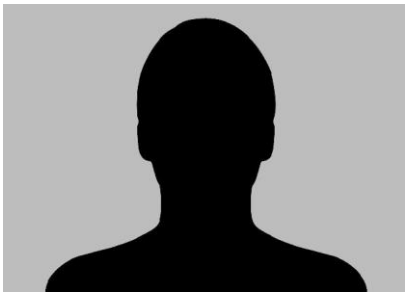
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

ISC Institute for Signal Processing and Wireless Communications

# Deep Learning based Pitch Detection



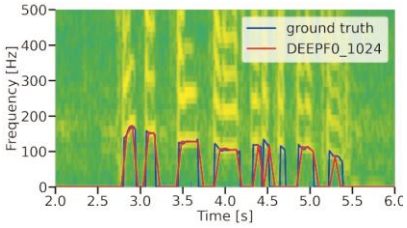
Diplomierende
Luca Di Lanzo
Kaspar Reto Wolfisberg

Dozent
Sigisbert Wyrsch

Ein wichtiges Feature für Sprachkompression und -codierung, Speaker Recognition und Speech Enhancement ist der Pitch oder genauer die Fundamentalfrequenz von Sprache. Eine robuste Detektion des Pitches der z. B. durch Hintergrundgeräusche verrauschten Sprache ist jedoch anspruchsvoll. In der Praxis werden zur Pitch-Detektion klassische Signalverarbeitungsalgorithmen angewandt.

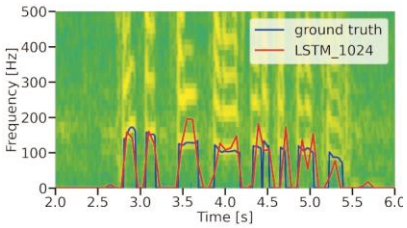
Die vorliegende Arbeit evaluiert die Tauglichkeit von Deep-Learning-basierten Methoden zur Pitch-Detektion verrauschter Sprache. Für viele praktische Anwendungen ist es wichtig, dass der Pitch mit einer niedrigen Latenz und geringem Rechenaufwand bestimmt werden kann. Deshalb werden in dieser Arbeit neben der Qualität der Pitch-Detektion auch die Laufzeiten der Methoden untersucht.

Die berücksichtigten Methoden umfassen CREPE und DeepF0, zwei Deep-Learning-basierte Pitch-Detection-Modelle aus der aktuellen Literatur. Zudem wurde ein eigenes LSTM-basiertes Modell implementiert. Die Modelle wurden mit verschiedenen Inputs trainiert und die Qualität sowie die Laufzeiten der Pitch-Bestimmungen evaluiert. Zusätzlich wurde untersucht, wie die Modelle mit sprachlosen Audiosegmenten umgehen. Die Ergebnisse wurden anschliessend mit den RAPT- und SWIPE-Pitch-Bestimmungsalgorithmen aus der klassischen Signalverarbeitung verglichen. Es hat sich gezeigt, dass von den Deep-Learning-Methoden das DeepF0-Modell die beste Balance zwischen Qualität und Laufzeit bietet. Anschliessend wurde durch Reduktion der Netzgrösse versucht, die Laufzeiten des DeepF0-Modells zu reduzieren.




Spectrogramm eines verrauschten Sprachsignals inklusive Pitch Ground Truth (blaue Linie) und Pitch Predictions (rote Linie) des DeepF0-Modells mit Input von 1024 Samples.

Die Ergebnisse zeigen, dass der eigene LSTM-Ansatz bezüglich der Pitch-Bestimmungsqualität vergleichbar mit den RAPT- und SWIPE-Algorithmen abschneidet. Die CREPE- und DeepF0-Modelle erzielen sehr gute Qualität. Die untersuchten Deep-Learning-Ansätze können hinsichtlich der Laufzeiten nicht mit den klassischen Signalverarbeitungsalgorithmen mithalten. Die Reduktion der Netzgrösse des DeepF0 führt kaum zu einer Qualitätseinbusse und vermochte die Laufzeiten zu verringern. Diese bleiben jedoch eine Grössenordnung über den RAPT- und SWIPE-Algorithmen.




Spectrogramm eines verrauschten Sprachsignals inklusive Pitch Ground Truth (blaue Linie) und Pitch Predictions (rote Linie) des LSTM-Modells mit Input von 1024 Samples.

Schliesslich hat sich gezeigt, dass die untersuchten Deep-Learning-Modelle sich kaum für den Echtzeiteinsatz eignen. Jedoch ist die Qualität der Pitch-Bestimmung vielversprechend und liegt für den verwendeten Datensatz über derjenigen der klassischen Signalverarbeitungsalgorithmen RAPT und SWIPE.