Signalanalyse Projektbericht

Robin Brendel
708292

Melf Fritsch
691435

Jakob Horbank
697989

Christoph Schoenicke
716271

24. Oktober 2023

1 Problembeschreibung

In diesem Projekt ging es um *Klassifikationsbasierte Richtungsschätzung in Hörgeräten zur Lokalisierung eines Sprachereignisses*. Ziel war es unbekannte Sprachsignale in eine von 6 Klassen zuzuordnen wobei jede Klasse 60° von -180° bis 180° abdeckt und somit die Richtung relative genau abgeschätzt werden kann. Mithilfe von *Head-Related Impulse Responses (HRIRs)* sollten zunächst aus den Mono-Audiosignalen, Ausgangssignale mit Raumrichtung generiert werden. Aus diesen lassen sich dann Features extrahieren um den Klassifikator zu trainieren.

2 Verwendete Daten

Wir benötigen Sprachsignale für die Generierung der Trainingsdaten. Wir haben uns für den Datensatz *Device Recorded VCTK (Small subset version)*[1] entschieden. Es enthält einige hundert professionell aufgenommene Sätze von verschiedenen Sprecherinnen und Sprechern. Zusätzlich sind Aufnahmen mit anderen Geräten enthälten, die wir jedoch nicht verwenden. Alle Sprachsignale haben eine Abtastrate von 16000 Hz.

Die zweite benötigte Komponente sind die HRIRs. Diese generieren wir mit der gestellten Matlab Funktion. Mit dieser Funktion können für 72 Winkel Impulsantworten für verschiedene Mikrofonpositonen am Ohr erzeugt werden. Die HRIS haben eine Abstastrate von 48000 Hz.

Die Testdaten sind Mikrofonsignale, aufgenommen an mit 2 Mikrofonen pro Ohr. Dabei befinden sich die Mikrofone auf der gleichen Höhe und in einem Abstand von 9mm.

3 Datengenerierung

Für die Datengenerierung wählen wir Sprachsignale von einem Mann und einer Frau mit jeweils hundert aufgenommenen Sätzen aus. Für jeden der 72 Winkel werden 10 Szenarios zufällig bestimmt. Ein Szenario besteht aus einem Sprachsignal, der Distanz zwischen Sprachquelle und Ohrmikfrofon, der Höhe und der stärke des Rauschens.

Für jedes Szenario wird die HRIR mit Winkel, Höhe und Distanz erzeugt. Wir generieren die HRIRs für die Mikrofonpositionen Vorne und in der Mitte, da man so am ehesten an den bekannten Abstand der Mikrofone in den

[1] https://datashare.ed.ac.uk /handle/10283/3038

Testdaten herankommt. Bevor wir die Sprachsignale mit den HRIRs kombinieren können muss die Abtastrate angeglichen werden. Dazu interpolieren wir die Sprachsignale auf 48000 Hz.

Die Trainingssignale werden dann nach folgender Formel generiert:

$$T_m(c, \phi, n) = h_m(c, \phi, n) * s_m(n) + \eta_m(c, n) \tag{1}$$

wobei h_m die entsprechende HRIR, s_m das Sprachsignal und η_m ein additives Rauschen darstellt.

Das Rauschen fügen wir mit einer Matlab Funktion hinzu, die weißes Rauschen so hinzufügt, dass eine bestimme SNR erreicht wird. Es wird zufällig ein SNR Wert von 5, 10 oder 20 gewählt.

4 Extrahierung der Merkmalsvektoren

Jedes Beispiel aus den generierten Daten enthält 4 Mikrofonsignale. Daraus extrahieren wir 4 Features, indem wir die *Generalized Cross-Correlation* (GCC) zwischen Mikrofonsignalen berechnen. Dazu wird die `gccphat` Funktion verwendet. Es wird jeweils die Verzögerung zwischen Vorne/Mitte und Links/Rechts berechnet.

5 Klassifikator

Zur Klassifikation wird den Richtungssektoren Klassen zugewiesen. Die Zuordnung ist in Abbildung 1 abgebildet. Als Klassifikator nutzen wir einen Entscheidungsbaum `fitctree`. Für das Training verwenden wir die vorher generierten Merkmalsvektoren.

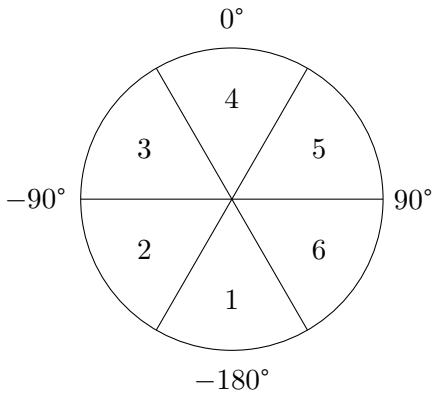

Abbildung 1: Zurordnung der Klassen zu Richtungssektoren

6 Ergebnisse

Es wurde ein Entwicklungsdatensatz mit Signalen und den Winkeln bereitgestellt. Die Genauigkeit in Abhängigkeit der Trainingbeispiele pro Winkel ist in Tabelle 1 zu sehen. Mit 50 Beispielen pro Winkel konnten wir 6/6 Signale korrekt klassifizieren. Da die Ausführung sehr lange braucht, hat es uns ausgereicht wenn mehrfach hintereinander 6/6 erreicht wurde. Bei weniger Beispielen konnten in seltenen Fällen ebenfalls 6/6 erreicht werden, meistens jedoch 5/6.

Es ist erkennbar, dass die Unterscheidung von links/rechts besser funktioniert als vorne/hinten, das Verwechslungen zwischen den Klassen 2/3 und 5/6 auftreten können. Dies erklären wir uns damit, dass die Mikrofone am Ohr nur wenige Millimeter auseinander sitzen und damit die Abstastrate nicht ausreicht um die minimalen Zeitunterschiede des eintreffenden Signales zu erkennen. Der Abstand zwischen den beiden Ohren ist deutlich größer, weshalb hier die Erkennung besser funktioniert.

Dies erklärt auch weshalb wir die größte Verbesserung durch die Interpolation auf 48000 Hz erzielen konnten. Dies entspricht einer verdreifachung der Abstastrate. Das Hinzufügen von Hall hingegen brachte keine Verbesserung. Wir vermuten, dass es zu schwer war die exakte Hallcharakteristik des Raumes, in dem die Entwicklungsdaten aufgenommen wurden, nachzuahmen.

Zur Bewertung wurde außerdem ein Evaluationsdatensatz ohne Azimuthwinkel bereitgestellt. Die vorhergesagten Azimuthwinkel sind in Tabelle 2 abbgedildet.

Traingsdaten pro Winkel	Korrekt klassifiziert
1	5/6
5	5/6
10	5/6
25	5/6
50	6/6

Tabelle 1: Klassifikation Developmentdatensatz

Name	Vorhergesagter Winkel
eval_event01	0°
eval_event02	−180°
eval_event03	60°
eval_event04	0°
eval_event05	0°
eval_event06	0°
eval_event07	60°
eval_event08	0°
eval_event09	120°
eval_event10	60°

Tabelle 2: Klassifikation Evaluationsdatensatz