UNIVERSITÄT ZU LÜBECK
Institut für Signalverarbeitung
Prof. Dr.-Ing. Alfred Mertins
Fabrice Katzberg
Marco Maass

SoSe23

Projektpraktikum Signalanalyse

Klassifikationsbasierte Richtungsschätzung in Hörgeräten zur Lokalisierung eines Sprachereignisses

Organisatorisches:

- Die Projektaufgabe wird in Gruppen bestehend aus 2 bis 4 Personen bearbeitet.
- Als Programmiersprache wird MATLAB verwendet. Es dürfen alle in MATLAB integrierten Toolboxen, Klassen und Funktionen verwendet werden. Für die Extraktion von Datenmerkmalen dürfen auch externe Funktionen (z.B. von Mathworks) verwendet werden.
- Laden Sie Ihre finalen Lösungen (10 Richtungsschätzungen für den Evaluationsdatensatz) und die zugehörigen Programmcodes als ein `zip` oder `tar.gz` Archiv mit dem Namen Ihres Teams über das Moodle-System hoch. Es soll auch eine Datei mit den Namen der beteiligten Personen beigelegt werden. Die für das Training verwendeten Audiodatensätze sollen nicht hochgeladen werden, jedoch müssen diese benannt und mit entprechender URL referenziert werden. Achten Sie darauf, dass Ihr Code lesbar geschrieben und verständlich kommentiert ist. Jeder Schritt muss reproduzierbar sein. Ihre Skripte und Funktionen müssen lauffähig sein. Programmcode, der zu Fehlermeldungen führt, wird nicht akzeptiert.
- Stellen Sie außerdem einen kurzen Bericht (2-4 Seiten) im `pdf` Format zur Verfügung, der Ihr Vorgehen bezüglich Datenvorverarbeitung, Merkmalsextraktion, Klassifikationsverfahren und Entscheidungs-Pooling dokumentiert. Berichten Sie außerdem über den Fehler, den Ihre Methode bezüglich der bereitgestellten Ground-Truth-Daten erzielt (6 Richtungsschätzungen für den Entwicklungsdatensatz).

Projektaufgabe

Ziel dieses Projekts ist es, eigenständig eine Klassifikationsmethode zu entwickeln, die in der Lage ist, den Richtungswinkel (Azimutwinkel) eines eintreffenden Sprachsignals auf einen Hörgeräteträger sinnvoll zu schätzen. Dies soll anhand der gemessenen Signale an den 4 Mikrofonen der Hörgeräte geschehen (jeweils 2 auf jeder Seite). Dabei können Sie davon ausgehen, dass jede Aufnahme nur eine einzige statische Sprachquelle beinhaltet.

Zur finalen Bewertung ihres Algorithmus, soll seine Leistungsfähigkeit an den 10 einzelnen Sprachereignissen des Evaluationsdatensatzes `evaluation_data.mat` unter Beweis gestellt werden. Für Testzwecke stehen Ihnen im Entwicklungsdatensatz `development_data.mat` 6 weitere Szenarien samt tatsächlicher Richtungswinkel zur Verfügung.

a) Ihre erste Aufgabe ist es, umfangreiche und realitätsnahe Trainingsdaten mit richtungsabhängigen Sprachereignissen zu erzeugen, auf Basis derer Ihr Klassifikationsalgorithmus trainiert werden soll. Zur Simulation der Richtungsabhängigkeit wird Ihnen ein Datensatz kopfbezogener Impulsantworten, *Head-Related Impulse Responses (HRIRs)*, zur Verfügung gestellt.

b) Das ursprüngliche Regressionsproblem der Richtungsschätzung soll zunächst als ein *Multiclass*-Klassifikationsproblem dargestellt werden. Dazu wird die gesamte $360°$-Spanne des Azimutwinkels durch mehrere kleine disjunkte Richtungssektoren abgedeckt. Jeder Sektor repräsentiert eine Klasse und erstreckt sich über ein schmales Winkelintervall. Im einfachsten Fall ist dann der zentrale Winkel des vom Klassifikator gewählten Sektors die Richtungsschätzung. Je schmaler das Sektorintervall, desto höher die Winkelauflösung der Richtungsschätzung und desto komplexer das Klassifikationsproblem.

   Implementieren Sie ein Klassifikationssystem, das eine Richtungsauflösung von $60°$ hat und folglich zwischen $360°/60° = 6$ Klassen unterscheidet. Entwickeln Sie hierzu auf Grundlage Ihrer Trainingsdaten

   i) etwaige Vorverarbeitungsschritte für die Mikrofonsignale,

   ii) einen geeigneten Merkmalsvektor

   iii) sowie einen *Multiclass*-Klassifikator, der sinnvolle Richtungsschätzungen liefert.

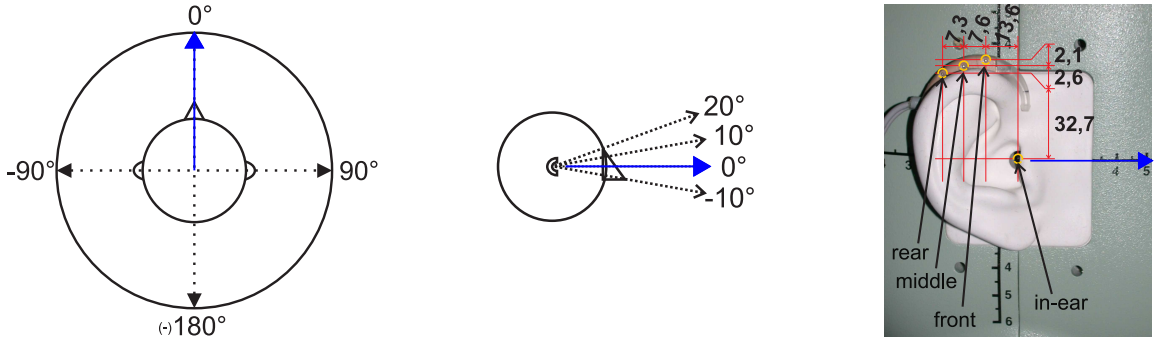
   Die Signalvorverarbeitung ist optional und kann z.B. das Heruntertasten der Signale auf eine niedrigere Abtastrate beinhalten. Auch ein räumliches Vorfiltern durch Beamformer ist hier denkbar, um bereits vor der Merkmalsextraktion Signale mit grober Richtungscharakteristik zu definieren. Achten Sie auf eine geeignete Skalierung beziehungsweise Normierung Ihrer Merkmale. In der Regel ist es sinnvoll, die (vorverarbeiteten) Mikrofonsignale in mehrere Signalfenster (Frames) einheitlicher Länge aufzuteilen und die Merkmalsextraktion sowie die Klassifikation fensterweise durchzuführen. Eine fensterweise Klassifikation erfordert eine finale Zusammenführung der Klassifikationsergebnisse aller Frames. Beispielsweise kann die Sprecherrichtung über alle Frames des aufgenommenen Sprachereignisses gemittelt werden oder es wird einfach die am häufigsten gewählte Richtungsschätzung bestimmt. Die Detektion und Selektion von Signalabschnitten mit Sprachaktivität kann für Ihr Klassifikationssystem von Vorteil sein, da Sprecherpausen vermutlich richtungsunabhängige Signalwerte liefern.

   Damit Sie Ihre Ideen und Herangehensweisen zu den genannten Punkten ausprobieren und vergleichen können, ist die Anwendung von Kreuzvalidierungsverfahren hilfreich.

Übersicht der Projektdaten

Es wird Ihnen im Ordner `hrir` ein Datensatz von HRIRs bereitgestellt, auf den bequem über die MATLAB-Funktion `getHRIR.m` zugegriffen werden kann. Dieser Datensatz[1] beinhaltet Aufnahmen von *behind-the-ear (BTE)* Hörgeräten vom Typ Siemens Acuris, gemessen an einem Kunstkopf im schalltoten Raum. Er stellt die Impulsantworten zu den 6 BTE Mikrofonen (jeweils 3 auf jeder Seite) für verschiedene Szenarien bereit. Dabei wurden die HRIRs in $-5°$-Schritten des Azimutwinkels gemessen, von $0°$ (vorne, Blickrichtung) über $-90°$ (links) bis $-180°$ (hinten), bzw. äquivalent für ein rechtsseitiges Schallereignis mit positivem Vorzeichen. Insgesamt ergeben sich also Richtungsdaten an 72 diskreten Richtungen des Azimutwinkels. Diese wurden für eine variierende Quellendistanz ($0{,}8\,\mathrm{m}$ und $3{,}0\,\mathrm{m}$) sowie einen variierenden Neigungswinkel (in $10°$-Schritten zwischen $-10°$ und $20°$) aufgenommen. Eine schematische Darstellung ist der Abbildung 1 zu entnehmen. Beachten Sie, dass für das Training nur jeweils 2 Mikrofone für jede Seite benötigt werden. Es steht Ihnen frei auch jeden anderen Datensatz gemessener oder simulierter HRIRs zu benutzen.

[1] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," EURASIP Journal on Advances in Signal Processing, Volume 2009, 10 pages, Article ID 298605, 2009



(a) (b) (c)

Abbildung 1: Konzept des bereitgestellten HRIR-Datensatzes: Belegungen (a) des Azimutwinkels, (b) des Neigungswinkels und (c) der Mikrofonkanäle. Der blaue Pfeil markiert die jeweilige Blickrichtung.

Entsprechende Trainingsdaten für den Mikrofonkanal c können Sie sich mittels

$$T_m(c, \phi, n) = h_m(c, \phi, n) * s_m(n) + \eta_m(c, n)$$

simulieren. Dabei kennzeichnet der Index m die Daten für verschiedene Messszenarien, $*$ ist der Faltungsoperator entlang der Zeitdimension, $h_m(c, \phi, n)$ bezeichnet die von der diskreten Zeitvariable n abhängige HRIR für den Azimutwinkel ϕ zum Mikrofon c, $s_m(n)$ ist ein Sprachsignal, und $\eta_m(c, n)$ bildet das Mikrofonrauschen nach. Generieren Sie sich für jede der 72 Richtungen Trainingssignale bezüglich unterschiedlicher Szenarien m. Variieren Sie dazu jeweils die Sprecherdistanz, den Neigungswinkel, das Sprachsignal sowie das Signal-Rausch-Verhältnis. Modellieren Sie das Messrauschen als weißes gaußsches Rauschen (MATLAB: `randn`). Sprachaufnahmen finden Sie in frei verfügbaren Sammlungen, bespielsweise in der CSTR VCTK Datenbank[2]. Achten Sie auf eine konsistente Abtastrate der Zeitsignale.

Es steht Ihnen frei, die Trainingsdaten noch wirklichkeitsgetreuer zu gestalten, z.B. können Sie jeweils leise (multidirektionale) Umgebungsgeräusche als Störquellen hinzufügen und auch verhallte Sprachsignale verwenden. Hilfreiche Datensätze und Tools dazu finden Sie im Internet[3].

Ihre trainierten Klassifikatoren sollen für die bereitgestellten Sprachereignisse des Entwicklungs- bzw. Bewertungsdatensatzes sinnvolle Richtungsschätzungen liefern. Diese Aufnahmen wurden in einem Computerlabor mit Nachhallzeit von etwa $0{,}5\,\mathrm{s}$ vorgenommen. Eintreffender Schall wurde durch die 4 Mikrofone der an einem Kunstkopf befestigten Hörgeräte des Typs Siemens Signia gemessen: An jedem Ohr sitzen 2 Mikrofone auf konstanter Höhe in einem (sagittalem) Abstand von $9\,\mathrm{mm}$. In Blickrichtung links sind die Mikrofonkanäle $c = 1$ (vorne) und $c = 2$ (hinten), rechtsseitig die Kanäle $c = 3$ (vorne) und $c = 4$ (hinten). Die Dimensionen des Kunstkopfes und der Hörgeräte unterscheiden sich leicht zum Setup des HRIR-Datensatzes. Vermeiden Sie es daher, Ihr Model an die Trainingsdaten überanzupassen (*overfitting*). Nutzen Sie die Ground-Truth-Winkel der 6 Szenarien des Entwicklungsdatensatzes, um die Performance Ihrer Klassifikatoren zu bewerten. Die tatsächlichen Azimutwinkel der Schallquelle sind hier in Radiant angegeben ($1\,\mathrm{rad} = 180°/\pi$ bzw. $1° = \pi/180$).

[2] https://datashare.ed.ac.uk/handle/10283/2950

[3] https://signalprocessingsociety.org/community-involvement/audio-and-acoustic-signal-processing/online-resources

Hintergrundwissen und Einordnung

Die Lokalisation akustischer Quellen in der Umgebung findet viele Anwendungen, insbesondere in den Bereichen Sicherheits- und Überwachungssysteme, Telekonferenzsysteme, Robotersysteme, Smart Home und Hörgeräte. Eine wesentliche Rolle spielt dabei die Richtungsschätzung (*direction-of-arrival estimation*). Sie bestimmt den Azimutwinkel, aus dem ein Schallereignis beim Empfänger eintrifft. Diese Information ist Grundlage etablierter Algorithmen zur Signalverbesserung und Szenenanalyse, darunter Beamforming, blinde Quellentrennung und automatische Spracherkennung. Zur Richtungsschätzung werden die Messsignale mehrerer Mikrofone (*array*) ausgewertet. Den einfachsten Fall stellt das bewegungslose Szenario mit nur einer Schallquelle dar. In realistischen akustischen Umgebungen wird dabei die Winkelgenauigkeit hauptsächlich durch die Faktoren Nachhall (*multipath arrival*), Hintergrundrauschen (*ambient noise*), Messrauschen und (temporäre) Quelleninaktivität negativ beeinflusst. Entsprechende Algorithmen zur Schätzung des Richtungswinkels lassen sich oft in die drei folgenden Kategorien einteilen:

- *Range-based* Methoden:
  Verzögerungen oder Signalstärkeunterschiede zwischen gemessenen Mikrofonsignalen werden bestimmt, um die (relativen) Entfernungen der einzelnen Mikrofone zur Schallquelle zu schätzen. Solche Entfernungsparameter werden dann genutzt, um die Schallquelle durch geometrische Überlegungen bzw. Triangulation zu lokalisieren.

  Relative Signalverzögerungen können durch das Maximum der Kreuzkorrelation ermittelt werden. Sprachsignale sind in kleinen Intervallen oft annähernd periodisch und bedürfen daher erweiterter Korrelationsverfahren, beispielsweise der *Generalized Cross-Correlation (GCC)-Phase Transform (PHAT)*.

  Ähnlich zu den Entfernungsparametern können für die Quellenlokalisation in Hörgeräten besondere binaurale Parameter benutzt werden. Die winkelabhängigen HRIRs an den Ohren des Hörgeräteträgers kodieren räumliche Anhaltspunkte auf eine relative Schallquellenposition in Form der Parameter *Interaural Level Difference (ILD)*, *Interaural Phase Difference (IPD)* und *Interaural Time Difference (ITD)*.
- *Beamforming* und *Spotforming* Methoden:
  Die akustische Umgebung wird virtuell gescannt, indem mögliche Positionen/Richtungen abgefahren werden und nach signifikannter Schallintensität gesucht wird. Beispielsweise werden diskrete Richtungen auf einem Gitter abgesucht und die Richtung des Maximums in der sogenannten *Steered Response Power (SRP)* bestimmt.
- *Supervised Learning* Methoden:
  Datenbasierte Verfahren des überwachten Lernens finden insbesondere zur binauralen Richtungsschätzung Anwendung. Detailierte Datensätze von winkelabhängigen HRIRs gemessen im schalltoten Raum, sowie eine Vielzahl an Audiodatensätzen mit Sprachsignalen und Hintergrundgeräuschen bieten viel Freiraum zur Generierung von umfangreichen Trainingsdaten, die Schallereignisse in realitätsnahen akustischen Umgebungen simulieren.

  Neben neueren Ansätzen des *Deep Learning*, die richtungsentscheidende Signalmerkmale eigenständig erlernen können, führen auch klassische Lernmethoden, beispielsweise im Rahmen der linearen Diskrimanzanalyse, zu robusten Richtungsschätzungen für unbekannte Testdaten. Die Merkmalsvektoren können dabei aus den bewährten Richtungsparametern zusammengesetzt werden, wie z.B. relative Signalverzögerungen, Kreuzkorrelationsfunktionen, GCCs, ITDs sowie (frequenzabhängige) ILDs und IPDs. Im Gegensatz zu den *Range-based* Methoden sind geometrische Überlegungen zur Eingrenzung der Quellenposition nicht notwendig, da die räumlichen Zusammenhänge der Parameter trainiert werden.