Sprachaktivitätserkennung mittels eines Mustererkenners für Atemschutzmasken

Michael Brodersen[1,2], Achim Volmer[1], Marcus Romba[1], Gerhard Schmidt[2]

[1] *Dräger Safety AG & Co. KGaA, 23560 Lübeck, E-Mail: michael.brodersen/achim.volmer/marcus.romba@draeger.com*

[2] *Christian-Albrechts-Universität zu Kiel, 24143 Kiel, E-Mail: mibr/gus@tf.uni-kiel.de*

Einleitung

Im Atemschutzeinsatz ist die Kommunikation unter Feuerwehrleuten aufgrund der starken Dämpfung der Atemschutzmaske und der lauten Umgebungsgeräusche sehr erschwert. Um die Kommunikation zu verbessern, werden Kommunikationssysteme für Atemschutzmasken eingesetzt, die das Sprachsignal beispielsweise an ein taktisches Funkgerät weiterleiten. Entscheidend ist, dass die Sprache mit besonders wenig Anstrengung verstanden werden kann. Dies kann durch den Einsatz einer Sprachaktivitätsdetektion unterstützt werden, die Störgeräusche wie beispielsweise Atemgeräusche aus dem Signal entfernt. In diesem Beitrag wird eine Sprachaktivitätsdetektion basierend auf einem Mustererkenner vorgestellt. Die Leistungsfähigkeit wird für unterschiedliche Ansätze wie für ein Neuronales Netz und ein Codebuch evaluiert und die Eignung für eine Echtzeitimplementierung auf einem eingebetteten, batteriebetriebenen System diskutiert.

Eigenschaften von Atemschutzmasken

Atemschutzmasken schützen das Gesicht sowie die Atemwege vor toxischen Gasen und Rauch (siehe [1]). Die Maske wird durch eine Dichtlinie um das Gesicht abgedichtet, Nase und Mund werden von der Innenmaske bedeckt, welche die Ausatemluft so lenkt, dass das Visier nicht beschlägt (siehe Abb. 1).

Abbildung 1: Atemschutzmaske auf einem Kunstkopf.

Der Raum vor Mund und Nase wird mit frischer Luft aus einer auf dem Rücken getragenen Flasche versorgt. Durch die Abdichtung der Atemschutzmaske am Gesicht, wird die Sprache stark gedämpft. Um diese Dämpfung zu minimieren, ist vor dem Mund eine Sprechmembran angeordnet.

Kommunikationseinheiten für Atemschutzmasken

Um die Kommunikation weiter zu verbessern, werden Kommunikationseinheiten verwendet, welche die Sprache mit einem Mikrofon vor der Sprechmembran aufzeichnen, verarbeiten und verstärkt auf die Lautsprecher oder ein taktisches Funkgerät ausgeben. Dieser Signalfluss ist in Abb. 2 dargestellt.

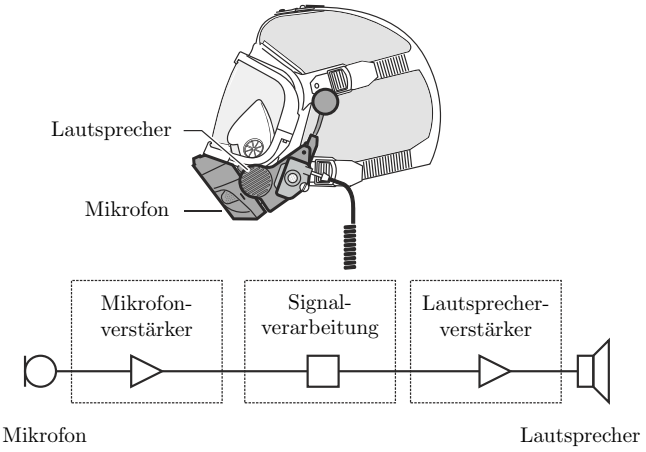

Abbildung 2: Struktureller Überblick der Kummunikationseinheit [1].

Für eine gute Kommunikation ist es notwendig, nur das Nutzsignal und nicht das Störsignal zu übertragen, damit der Inhalt verständlicher wird. Bei der Masken-Kommunikationseinheit sind besonders die Atemgeräusche störend. Sprache und Atemgeräusche sind jedoch im Mikrofonsignal vorhanden. Um diese störenden Signale nicht über die Lautsprecher oder über ein taktisches Funkgerät auszugeben, ist eine Sprachaktivitätserkennung und eine anschließende Filterung notwendig, welche diese Störgeräusche herausfiltern kann. Ein möglicher Ansatz dafür sind Mustererkenner, welche auf das Atemgeräusch trainiert werden. Im Folgendem werden ein Neuronales Netz und ein Codebuch in Bezug auf die Leistungsfähigkeit einer Störgeräuschunterdrückung gegenübergestellt.

Merkmalsextraktion

Ein Mustererkenner nutzt Signaleigenschaften für die Klassifikation aus. Daher wird das Mikrofonsignal zuerst einer Merkmalsextraktion unterzogen. Diese verarbeitet das Eingangssignal so, dass die relevanten Merkmale herausgearbeitet werden. Diese Eigenschaften des Eingangssignal müssen durch möglichst wenige signifi-

kante Merkmale beschrieben werden, um den Rechenaufwand zu minimieren. Bei der Merkmalsextraktion wird das Eingangssignal in den Frequenzbereich transformiert und die Merkmale anhand des Betragssprektrums, einer Melfilterung und einer Logarithmierung ermittelt. Aus 65 komplexen Stützstellen werden nach der Transformation $N_\mathrm{M} = 12$ Merkmale extrahiert und in dem Vektor

$$\boldsymbol{X}_\mathrm{M}(k) = \Big[X_\mathrm{M}(0,k),\ ...,\ X_\mathrm{M}(N_\mathrm{M}-1,k)\Big]^\mathrm{T} \qquad (1)$$

zusammengefasst, wobei k der Rahmenindex ist. Daraufhin werden die Merkmale dem Mustererkenner präsentiert. In Abb. 3 ist der Signalflussgraph der Sprachaktivitätserkennung abgebildet.

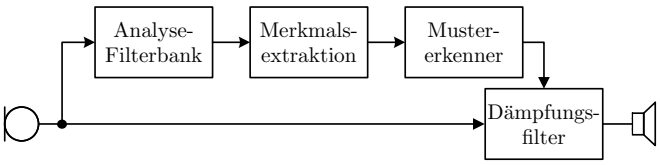

Abbildung 3: Signalflussgraph der Sprachaktivitätserkennung

Dieser Signalflussgraph ist stark vereinfacht und bezieht sich hierbei nur auf die Verarbeitung des Mustererkenners und nicht auf die Problematik des Gesamtsystems. Wird das Signal vom Mustererkenner als Sprache klassifiziert, wird dieses auf dem Lautsprecher ausgegeben.

Neuronales Netz

Das Neuronale Netz orientiert sich an der Funktion des Gehirns, welches aus Neuronen besteht [2]. Diese Neuronen, auch Knoten genannt, sind miteinander verknüpft und die Übergangspfade sind mit einer Gewichtung w_{ij} versehen, womit der Übergang von den Knoten i zu j beschrieben wird. Das hier verwendetet Neuronale Netz besteht aus einer Eingangsschicht, einer verdeckten Schicht und einer Ausgangsschicht. In der Eingangsschicht werden die einzelnen Elemente des Merkmalsvektors $\tilde{X}(m,k)$ mit

$$\tilde{X}(m,k) = 2 \cdot \frac{X_\mathrm{M}(m,k) - X_\mathrm{M,min}(m)}{X_\mathrm{M,max}(m) - X_\mathrm{M,min}(m)} - 1 \qquad (2)$$

auf den Bereich $-1 < \tilde{X}(m,k) < 1$ normiert, wobei m der Merkmalsindex ist. Dabei sind in $X_\mathrm{M,min}(m)$ die minimalen und in $X_\mathrm{M,min}(m)$ die maximalen Werte des zugehörigen Merkmals der Trainingsdaten, welche zum Training des Neuronalen Netzes benutzt werden, hinterlegt. Daraufhin werden die normierten Eingangsdaten von den zugehörigen Knoten zu den Knoten der verdeckten Schicht verteilt und mit dem zugehörigem Gewicht $w_{v,in}$ versehen. Zu den eingehenden Pfaden in dem Knoten wird zusätzlich ein Bias B_{Vi} für die verdeckte Schicht und B_{Aj} für die Ausgangsschicht an dem Knoten i hinzu addiert. Der Bias ist als Schwellenwert des jeweilige Knoten anzusehen. Die eingehenden Signale in jedem Pfad werden aufaddiert und daraufhin mit einer linearen begrenzten Übertragungsfunktion gewichtet, welche das Verhalten

$$\mathrm{f}(x) = \begin{cases} 1 & , \text{ wenn } x > 1, \\ -1 & , \text{ wenn } x < -1, \\ x & , \text{ sonst,} \end{cases} \qquad (3)$$

besitzt. Dadurch ergibt sich der Zusammenhang der Übertragung von der Eingangsschicht zur verdeckten Schicht inklusive der linearen begrenzten Übertragungsfunktion gemäß Gl. (3) zu

$$X_\mathrm{V}(i,k) = \mathrm{f}\left(B_{\mathrm{V}i} + \sum_{n=0}^{N_\mathrm{M}-1} \tilde{X}(n,k) \cdot w_{\mathrm{V}in}\right), \quad \text{für } 0 \le i < N_\mathrm{V}. \qquad (4)$$

wobei N_M die Anzahl der Merkmale, w_{Vin} die Gewichte von der Eingangsschicht zur verdeckten Schicht und $X_V(i,k)$ der anliegende Wert an dem Knoten i beschreibt. Der Ergebnisvektor der Ausgangsschicht wird mit $\boldsymbol{X}_\mathrm{A}(k) = [X_\mathrm{A}(0,k),\ ...,\ X_\mathrm{A}(N_\mathrm{A}-1,k)]^\mathrm{T}$ beschrieben und dessen Berechnung ergibt sich zu:

$$X_\mathrm{A}(j,k) = \mathrm{f}\left(B_{\mathrm{A}j} + \sum_{i=0}^{N_\mathrm{V}-1} X_V(i,k) \cdot w_{\mathrm{A}ji}\right), \quad \text{für } 0 \le j < N_\mathrm{A}. \qquad (5)$$

Dabei stellt N_V die Anzahl der Knoten der verdeckten Schicht, w_{Aji} das Gewicht von der verdeckten Schicht zur Ausgangsschicht und $X_A(j,k)$ das Ergebnis an dem Knoten j dar. Der Ergebnisvektor der Ausgangsschicht hat $N_\mathrm{A} = 5$ Elemente, was der Anzahl der zu erkennenden Klassen entspricht. Dabei wird zwischen Geräusch, Pause, Ausatmen, Einatmen und Sprache unterschieden. Diese werden in der Menge M zusammengefasst, so dass $M = \{\text{Geräusch, Pause, Ausatmen, Einatmen, Sprache}\}$ ist. Die Klasse Geräusch beinhaltet die auffälligen Geräusche wie beispielsweise Knackgeräusche vom Betätigen der Schalter an der Atemschutzmaske. Zur Detektion, welcher Pfad am Wahrscheinlichsten ist, wird der Index des maximalen Eintrags des Vektors $\boldsymbol{X}_\mathrm{A}(k)$ bestimmt:

$$d_\mathrm{max}(k) = \underset{p \in M}{\mathrm{argmax}}\Big\{\mathbf{X}_A(p,k)\Big\}. \qquad (6)$$

Mit dem Index des maximalen Elements $d_\mathrm{max}(k)$ kann nun durch das dabei angegebene p ermittelt werden, welches Element der Menge M aktiv ist. Solch ein verwendetes *feed forward*-Netz ist in Abb. 4 zu sehen. Damit die Berechnungen des Neuronales Netzes vollzogen werden können, muss das Neuronale Netz durch eine Trainingsphase initialisiert werden.

Die benötigten minimalen und maximalen Eingangswerte, die Übergangsgewichte und die Bias werden durch ein Training generiert. Dazu ist es notwendig, Merkmalsvektoren aus Audiosignalen der jeweiligen Klasse zu erzeugen. Das Training des Neuronales Netzes wird mittels des *back propagation*-Algorithmus durchgeführt. Das entworfene Neuronale Netz ist in Abb. 4 dargestellt,

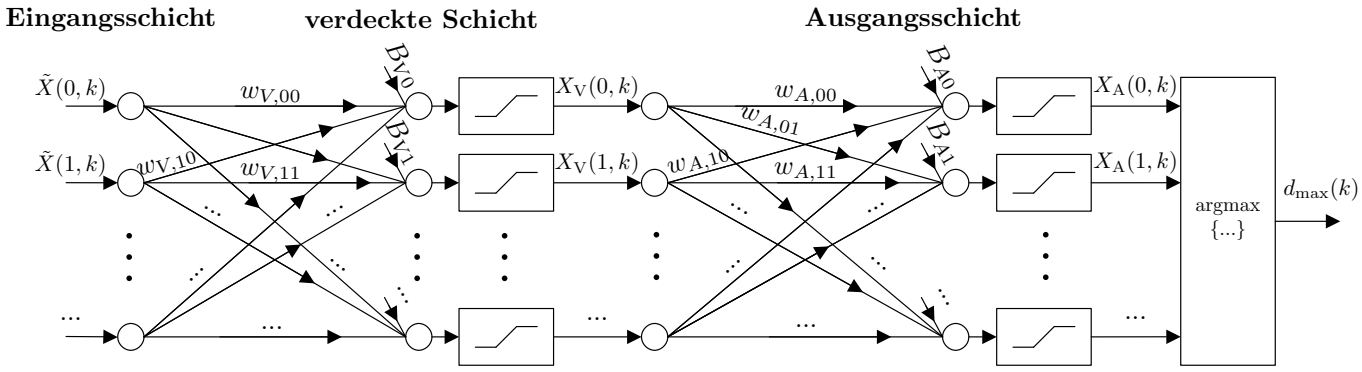

Abbildung 4: Neuronales Netz.

wobei die Übertragungsfunktion aus Gl. (3), die Anzahl der Merkmale und die Anzahl der Schichten der verdeckten Schicht und die Anzahl der Ausgangsmerkmale verwendet werden.

Codebuch

Das Codebuch ist ein Mustererkenner, der basierend auf einer trainierten Datenbank einen Vergleich zwischen den Datenbankeinträgen und dem aktuellem Merkmalsvektor vollzieht [3]. Der Datenbankeintrag mit minimaler, quadratischer, euklidischer Distanz wird als am wahrscheinlichsten interpretiert, so dass die Klasse dieses Datenbankeintrags als aktuelle Zustandsschätzung verwendet wird. Somit kann zwischen den Klassen Geräusch, Pause, Einatmen, Ausatmen und Sprache unterschieden werden.

Die Einträge der Datenbank werden durch ein Training erzeugt, für welches Merkmalsvektoren der jeweiligen Klasse erzeugt werden. Diese Merkmalsvektoren werden mittels des *K-means Clustering*-Algorithmus [4] so zusammengefasst, dass sich ein kompaktes Codebuch ergibt. Das Codebuch für Sprache besitzt mehr Einträge als die übrigen Klassen, da Sprache eine größere Varianz im Merkmalsraum aufweist. Die Klassen Geräusch, Pause, Ausatmen und Einatmen werden auf 16 Merkmalsvektoren und Sprache auf 64 Merkmalsvektoren für die entsprechenden Codebücher reduziert.

Mit dem trainiertem Codebuch kann die Aktivität der Klassen bestimmt werden. Zu dieser Bestimmung wird die minimale quadratische euklidische Distanz

$$d_{\mathrm{E},p}\big(\tilde{\boldsymbol{X}}(k), \boldsymbol{c}_{s,p}, k\big) = \min_{s=0\,\ldots\,N_{\mathrm{c}}-1} \left\{ \sqrt{\sum_{m=0}^{N_{\mathrm{M}}-1} \left| \tilde{X}(m,k) - c_{s,p,m} \right|^2} \right\}, \quad \text{für } p \in M \tag{7}$$

zwischen dem Merkmalsvektor $\tilde{\boldsymbol{X}}(k)$ und den Codebucheinträgen $\boldsymbol{c}_{s,p}$ berechnet [5]. Der Index s steht dabei für den Codebuchvektor s aus dem Set p, wobei es fünf Sets (Geräusch, Pause, Einatmen, Ausatmen und Sprache) gibt. Die fünf berechneten minimalen quadratischen euklidischen Distanzen von den Sets werden miteinander verglichen und das Argument des Minimums wird bestimmt:

$$d_{\min}(k) = \underset{p\in M}{\operatorname{argmin}} \left\{ d_{\mathrm{E},p}\big(\tilde{\boldsymbol{X}}(k), \mathbf{c}_{s,p}, k\big) \right\}. \tag{8}$$

Anhand der minimalen Distanz kann nun ermittelt werden, welches Element der Menge M aktiv ist. Die Sprachaktivitätserkennung mittels Codebuch ist in Abb. 5 abgebildet.

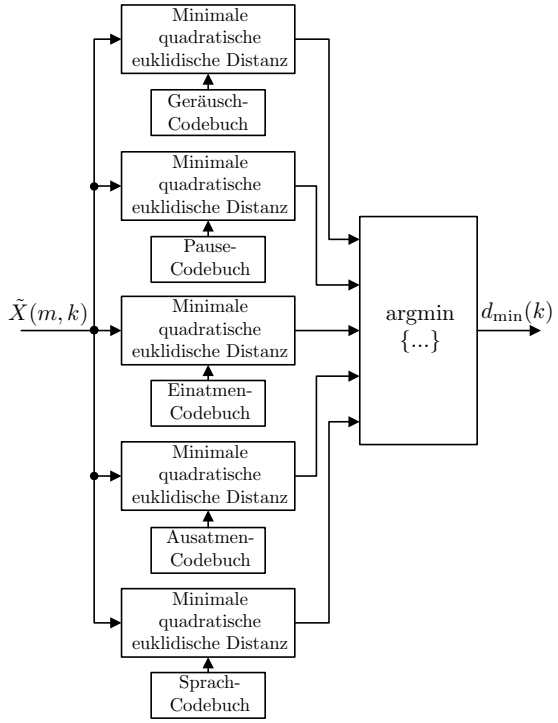

Abbildung 5: Blockschaltbild des Codebuchs

Performancevergleich der Mustererkenner

Die Kommunikationseinheit der Atemschutzmaske ist ein batteriebetriebenes Gerät, bei welchem die Batterielaufzeit sehr entscheidend ist. Um diese Zeit zu maximieren werden Prozessoren mit sehr geringer Leistung verwendet, wodurch die Anzahl der Rechenoperationen pro Zeit begrenzt sind. Aus diesem Grund werden die verschiede-

nen Mustererkennungsansätze hinsichtlich ihrer Erkennungsraten, benötigten Rechenoperationen und auf den benötigten Speicherbedarf verglichen.
Die Erkennungsraten des beschriebenen Neuronalen Netzes sind in Tab. 1 und die des beschriebenen Codebuches in Tab. 2 dargestellt, wobei die Klassen Geräusch, Pause, Ausatmen, Einatmen und Sprache durch G, P, A, E und S abgekürzt werden.

Eingang \ Ziel	G	P	A	E	S	Erkennung	
G	53%	23%	14%	2%	8%	53%	47%
P	14%	70%	8%	1%	7%	70%	30%
A	2%	2%	83%	1%	12%	83%	17%
E	<1%	<1%	<1%	99%	<1%	99%	1%
S	2%	<1%	2%	<1%	95%	95%	5%
Durchschnittliche Erkennungsrate						79%	21%

Tabelle 1: Erkennungsmatrix des Neuronales Netzes.

Eingang \ Ziel	G	P	A	E	S	Erkennung	
G	38%	40%	9%	3%	10%	38%	62%
P	4%	91%	3%	<1%	2%	91%	9%
A	6%	10%	68%	1%	14%	68%	32%
E	2%	<1%	<1%	96%	<1%	96%	4%
S	3%	3%	5%	<1%	88%	88%	12%
Durchschnittliche Erkennungsrate						76%	24%

Tabelle 2: Erkennungsmatrix des Codebuches.

In den Tabellen sind links die Eingangsklassen und oben die Zielklassen abgebildet. Für die Zielklassen sind die Erkennungsraten in % für die jeweilige Eingangsklasse abgebildet; rechts daneben ist die Gesamterkennungsrate für die Eingangsklasse zusammengefasst. Die Erkennungsraten der fälschlich zugewiesenen Merkmale sind in rot und die richtigen in grün dargestellt. In der untersten Zeile ist die durchschnittliche Erkennungsrate für alle Klassen dargestellt. Bei den Erkennungsraten sollte die Verwechslung zwischen Ausatmen und Sprache möglichst klein sein, da beispielsweise Zischlaute ansonsten als Ausatmen klassifiziert werden.
Der Vergleich der Erkennungsraten beider Mustererkenner zeigt, dass die Erkennungsraten mit dem Neuronalem Netz für die Klasse Sprache um 7 %, für Einatmen um 3 %, für Ausatmen um 15 % und für Geräusch um 15 % höher sind. Lediglich die Erkennungsrate für die Klasse Pause ist beim Codebuch um 21 % besser, wobei die Fehlerkennung von Pause beim Neuronalem Netz größtenteils in der Klasse Geräusch wieder zu finden ist. Somit werden diese Fehlerkennungen von Pause nicht fälschlicherweise als Sprache detektiert. Die Verwechslung von Ausatmen und Sprache ist beim Neuronalem Netz geringer und die durchschnittliche Erkennungsrate um 3 % höher. Somit ist das Neuronale Netz in den entscheidenden Fällen dem Codebuch vorzuziehen.
Die Erkennungsraten müssen dabei in Bezug auf die benötigten Rechenoperationen und den Speicherbedarf gesehen werden, welche für das Neuronale Netz (NN) und für das Codebuch (CB) in Tab. 3 dargestellt sind, wobei der Speicherbedarf in Bezug auf die Anzahl der benötigten Parameter im 16 Bit-Format für einen 4 ms Rahmen angegeben ist.

	Additionen	Multiplikationen	Speicherbedarf
NN	188	172	209
CB	3328	1792	1536

Tabelle 3: Benötigte Ressourcen der Mustererkenner.

Das Neuronale Netz ist gegenüber dem Codebuch sowohl hinsichtlich der benötigten Rechenleistung als auch des Speicherbedarfs sehr kompakt. Das Codebuch benötigt 17-mal so viele Additionen wie das Neuronale Netz, 10-mal so viele Multiplikationen und den 7-fachen Speicherplatz.

Im Gesamtvergleich der Leistungsfähigkeit ist das Neuronale Netz gegenüber dem Codebuch merklich im Vorteil, da die Erkennungsraten höher sind, allgemein sowie in den entscheidenden Fällen, und da es deutlich weniger Ressourcen benötigt.

Zusammenfassung

Mit Mustererkennern kann eine robuste Sprachaktivitätserkennung umgesetzt werden, die zwischen Geräuschen beim Bedienen der Atemschutzmaske, Pausen, Ausatmen, Einatmen und Sprache unterscheiden kann. Als Ansätze zur Mustererkennung wurden ein Neuronales Netz und ein Codebuch untersucht und hinsichtlich ihrer Leistungsfähigkeit verglichen. Die Erkennungsraten des Neuronalen Netzes und des Codebuches unterscheiden sich nur geringfügig, wobei das Codebuch für eine ähnlich gute Erkennung deutlich mehr Rechenoperation und Speicherplatz benötigt. Daher ist als Mustererkenner das Neuronale Netz im Vergleich zum Codebuch für die batteriebetriebene Anwendung in der Atemschutzmaske zu bevorzugen.

Literatur

[1] Dr.-Ing. A. Volmer, Dr.-Ing. M. Romba, C. Schmidt und M. Houssem Harbi: Optimization of Speech Intelligibility for Fire Fighters' Full Face Masks, DAGA, 2013

[2] C. M. Bishop: Pattern Recognition and Machine Learning, Springer, 2006

[3] G. A. Fink: Markov Models for Pattern Recognition: From Theory to Applications, Springer, London, 2014

[4] B. Pfister und T. Kaufmann: Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung, Springer, 2008

[5] M Bossert: Kanalcodierung, Oldenbourg Wissenschaftsverlag, 2013