Spracherkennung in stark gestörten Unterwasserumgebungen

Tim Owe Wisch[1], Thorben Kaak[1], Alexej Namenas[1], Gerhard Schmidt[1]

Digitale Signalverarbeitung und Systemtheorie, CAU zu Kiel, E-Mail: {timw,thka,aln,gus}@tf.uni-kiel.de

Kurzfassung

Sprachkommunikation unter Wasser ist ein kompliziertes Tätigkeitsfeld. Aktuell gibt es nur wenige technische Ansätze, die es Schwimmern oder Tauchern erlauben mit Menschen außerhalb des Wassers Kontakt zu halten oder Kommandos zu empfangen. Der menschliche Sprach- und Hörapparat ist nicht an die Artikulation unter Wasser angepasst, weshalb das Sprechen in dieser Umgebung für den Menschen eine Herausforderung ist. In einem vorangegangenen Projekt an der Universität Kiel [1] wurde bereits eine Tauchermaske entworfen, die mit wasserfesten Mikrofonen und einem WLAN-Modul ausgerüstet ist. Da es sich jedoch um eine Vollgesichtsschwimmaske handelt, wird in den luftgefüllten Raum zwischen Maske und Gesicht artikuliert. In diesem Beitrag wird die Idee aufgegriffen und erweitert, sodass mithilfe von an einer Schwimmbrille befestigten Mikrofonen direkt ins Wasser gesprochene Kommandos detektiert werden. Als erster Schritt wurde eine Spracherkennung für einen MP3-Player aufgebaut. In Testreihen in Schwimmbädern und Pools wurden Sprachdaten unterschiedlicher Personen aufgezeichnet und mit verschiedenen Machine-Learning-Algorithmen prozessiert. Für die Spracherkennung kommen Gauß'sche Mischmodelle mit Erweiterungen zum Einsatz, bei denen zusätzlich zu den Wahrscheinlichkeiten der Modelle die Abfolge der Gauß-Verteilungen mit maximaler Wahrscheinlichkeit in Betracht gezogen wird. Mit dem Gaussian-Mixture-Model Mean-Value-Tracking (GMM-MVT) wird eine Erkennungsrate von 81,7% erreicht.

Grundlagen

Die Schallgeschwindigkeit c im Medium Wasser ist deutlich höher als im Medium Luft (1480 $\frac{m}{s}$ bzw. 330 $\frac{m}{s}$ [2]). Direkt verknüpft mit der Schallgeschwindigkeit ist auch der Wellenwiderstand eines Mediums (Tabelle 1).

Medium	Wellenwiderstand
Luft	$Z_{\text{Luft}} = 4,14 \cdot 10^2 \ \frac{kg}{m^2 \cdot s}$
Wasser	$Z_{\text{Wasser}} = 1,48 \cdot 10^6 \ \frac{kg}{m^2 \cdot s}$

Tabelle 1: Wellenwiderstände [3]

Aus den stark abweichenden Wellenwiderständen ergibt sich die erste Schwierigkeit für eine Spracherkennung direkt ins Wasser gesprochener Worte. Wellen werden an Medienübergängen zum Teil reflektiert, wobei sich das Verhältnis von reflektierter zu transmittierter Leistung aus dem Verhältnis der Wellenwiderstände der Medien am Übergang ergibt:

$$\frac{P_r}{P_e} = \frac{Z_{\text{Wasser}} - Z_{\text{Luft}}}{Z_{\text{Wasser}} + Z_{\text{Luft}}}. \tag{1}$$

Aus (1) ergibt sich zwar ein Reflexionskoeffizient von 99%, doch auch Körperschall spielt bei der Schallübertragung vom Sprechapparat ins Wasser eine signifikante Rolle, allerdings größtenteils im tieffrequenten Bereich. Neben den physikalischen Schwierigkeiten sind auch biologische Probleme anzutreffen. Durch den begrenzten Atem kann nicht dauerhaft bzw. nur schwer in ganzen Sätzen ins Wasser gesprochen werden. Einzelne Kommandos sind jedoch sehr wohl möglich.

Schwimmbrille

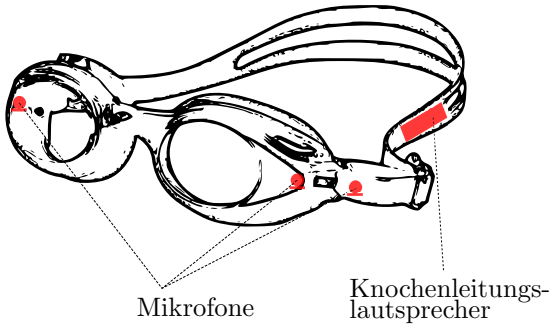

Abbildung 1: Schwimmbrille ausgestattet mit Mikrofonen und Knochleitungslautsprecher.

Als mögliches Anwendungsbeispiel für den Einsatz einer Spracherkennung unter Wasser wurde sich für einen sprachbedienbaren MP3-Player entschieden. Dieser benötigt nur eine begrenzten Anzahl von Befehlen und der Nutzer erhält sofort Feedback, ob das Kommando korrekt ausgeführt wurde. Zusätzlich gibt es durchaus einen Bedarf an derartigen Lösungen, wie die Verkäufe von MP3-Playern für Schwimmer zeigen.

Nummer	#1	#2	#3	#4	#5
Kommando	play	stop	pause	previous	next

Tabelle 2: Kommandoübersicht MP3-Player

Um Sprache unter Wasser aufzuzeichnen wurde eine mit wasserfesten Mikrofonen (IP67-Zertifizierung) ausgestattete handelsübliche Schwimmbrille verwendet. Die Mikrofonsignale werden durch Verstärker in einer wasserdichten Box vorverstärkt und von dort aus an das Aufzeichnungsgerät weitergeleitet. Zwei Mikrofone befinden sich innerhalb der Brille hinter den Gläsern, ein weiteres außerhalb am Kopfgurt (Abb. 1). Für die Anwendung der

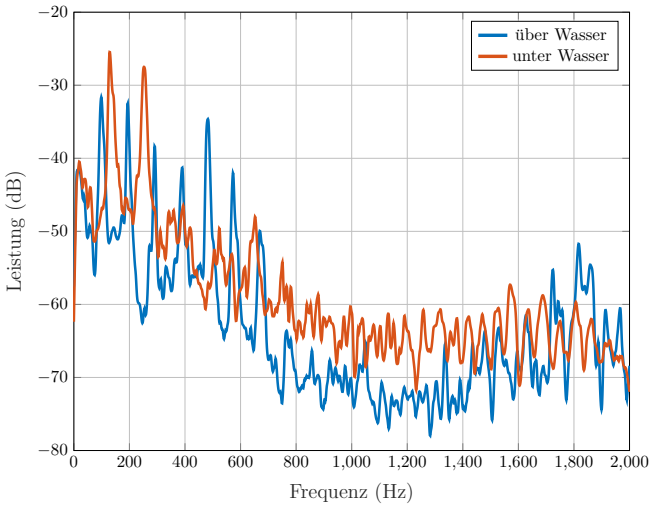

Abbildung 2: Geglättete Einhüllende des Spektrums des Kommandos „play“ über und unter Wasser.

Brille als MP3-Player unter Wasser wird auch ein Element zur Musikwiedergabe benötigt, weshalb ebenfalls ein Knochenleitungslautsprecher verbaut wurde. Dieser stellt eine einfache Möglichkeit dar auch unter Wasser Musik zu hören, indem der Schädelknochen in Schwingung versetzt wird. Die Qualität liegt zwar unter der von In-Ear-Kopfhörern über Wasser, erlaubt jedoch eine kostengünstige Alternative zu klassischen, wasserdichten Lautsprechern, welche unter Wasser häufig zusätzlich eine aufwendige Entzerrung benötigen. Weiterhin enstehen auch keine Schwierigkeiten mit eindringendem Wasser in den Gehörgang und Druckausgleich beim Tauchen.

Verarbeitung

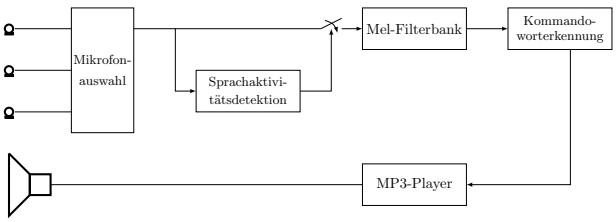

Abbildung 3: Systemübersicht

Für die Vorverarbeitung müssen die Signale der einzelnen Mikrofone entsprechend ihrer Position in bzw. an der Schwimmbrille unterschiedliche behandelt werden. Während das äußere Mikrofon wenig Störungen aufweist, zeigen die internen Mikrofone starke niederfrequente Störungen biologischen Ursprungs. Die Wimpern- und generelle Muskelbewegungen sind im aufgezeichneten Signal deutlich erkennbar. Diesen Störungen kann mit einem Hochpassfilter mit einer Grenzfrequenz von 100 Hz begegnet werden. Zusätzlich ist in einigen Umgebungen ein Kammfilter nötig, um 50-Hz-Störungen und deren Harmonische zu unterdrücken. Durch die begrenzte Bandbreite des Signals wird im gesamten System mit einer Abtastrate von 4 kHz gearbeitet. Als Eingang für die Erkennerstruktur kommen Mel-Koeffizenten [5] zum Einsatz, da diese eine kompakte Representation des Sprachspektrums darstellen und auch in weiteren Arbeiten zur Spracherkennung zum Einsatz kommen. Zusätzlich wird eine lineare Diskriminanzanalyse (LDA) zur Dimensionsreduktion eingesetzt [7].

Für die eigentliche Spracherkennung wurden verschiedene Erkennerarten in Betracht gezogen. Neben Codebüchern und Gauß'schen Mischmodellen (GMM) wurden auch Versuche mit neuronalen Netzen getätigt, um eine geeignete Erkennerstruktur für diesen Einsatzzweck zu ermitteln [4]. Die Nutzung von Hidden-Markov-Modellen ist in diesem Anwendungsfall nur schwer möglich, da für das Training der Modelle das Labeln der einzelnen Laute innerhalb einer Äußerung nötig ist, was durch die schlechte Qualität der aufgezeichneten Sprache unter Wasser und die dadurch bedingte schlechte Trennbarkeit der Laute erschwert wird. Der Vorteil von GMMs gegenüber Codebüchern liegt in der zusätzlichen Ausgabe der Wahrscheinlichkeiten für eine bestimmte Äußerung. $\boldsymbol{g}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ stehen dabei für die Gewichte, Mittelwerte und Kovarianzmatrizen der trainierten GMMs. Aus diesen und dem Eingangsvektor $\boldsymbol{x}$ des Blocks k wird die Zugehörigkeitswahrscheinlichkeit von $\boldsymbol{x_k}$ zum jeweiligen Modell m gebildet:

$$p(\boldsymbol{x_k}|\boldsymbol{g^m}, \boldsymbol{\mu^m}, \boldsymbol{\Sigma^m}) = \sum_{l=1}^{L} g_l \mathcal{N}(\boldsymbol{x_k}|\boldsymbol{g_l^m}, \boldsymbol{\mu_l^m}, \boldsymbol{\Sigma_l^m}). \quad (2)$$

Bei Sprach- bzw. Worterkennungen über Wasser wird oftmals die Blockgröße von Rahmen auf denen eine Mel-Extraktion und anschließende Erkennung ausgeführt wird so gewählt, dass die Sprache in diesem Zeitbereich als stationär angenommen werden kann. Für die Analyse der Unterwassersprache wurden verschiedene Blocklängen zwischen 16 ms und 512 ms mit und ohne einer vorgeschalteten LDA betrachtet.

	Pufferlänge	Samples	Überlappung
GMM I	64ms	256	50%
GMM II	256ms	1024	50%

Tabelle 3: Genutzte GMMs

Die für die Erkennung genutzten GMMs sind in Tabelle 3 dargestellt. Obgleich bei den höheren Pufferlängen nicht mehr Stationarität angenommen werden kann, werden im Sinne einer optimierten Detektionsperformanz die GMMs mit 64 ms und GMMs mit 256 ms trainiert. Eine Überlappung der einzelnen Blöcke wird zu 50 % gewählt. Die Dimension der GMMs für die einzelnen Wörter wird mit Hilfe des Akaike-Informationsmaßes bestimmt [6]:

$$AIC = -2l(\hat{\Theta}) + 2M, \quad (3)$$

mit der logarithmierten Likelihoodfunktion $l(\hat{\Theta})$ und M als Modelldimension. Dies gewährleistet, dass die Modelle einen Kompromiss aus nötiger Anpassungsgüte und Komplexizität unter Vermeidung von *overfitting* abbilden.

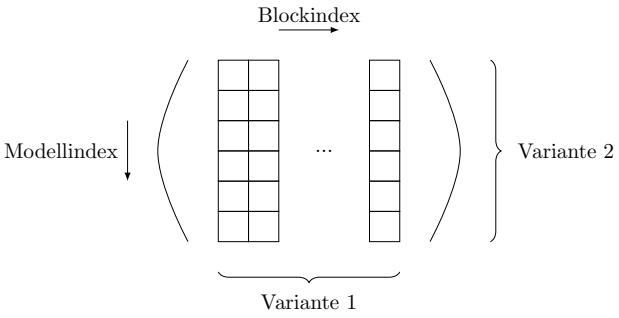

Abbildung 4: Varianten zur Auswertung der Wahrscheinlichkeitsmatrix einer Äußerung.

In diesem System gibt es für eine Äußerung zwei Arten der Klassifizierung mit GMMs (Abb. 4). Jede Äußerung hat eine Dauer von mehreren Blöcken, sodass sich die Möglichkeit ergibt einerseits jeden Block einzeln zu klassifizieren und nachdem alle Blöcke der Äußerung klassifiziert sind, diejenige auszuwählen, die in den meisten Blöcken die höchste Wahrscheinlichkeit aufweist (Variante 1). Alternativ bietet sich die Möglichkeit die Wahrscheinlichkeiten während einer Äußerung laufend in eine Matrix aufzunehmen und diese dann zeilen- statt spaltenweise auszuwerten (Variante 2). Für die zeilenweise Auswertung werden die Blockwahrscheinlichkeiten multipliziert, anschließend die Gesamtwahrscheinlichkeit je Modell gebildet und das Maximum ausgewählt. Zusätzlich zu den klassischen Varianten der Klassifikation mit GMMs wird ein neurartiges Pfadmodell eingeführt, welches eine gewisse Ähnlichkeit zu Codebüchern aufweist.

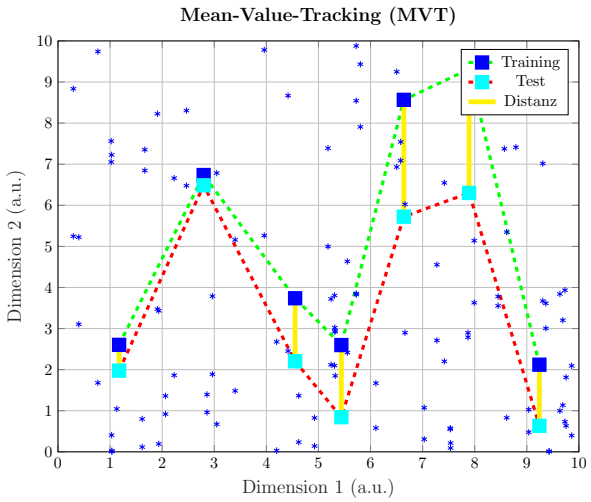

Abbildung 5: Beispielpfadmodell mit trainiertem und aufgezeichneten Pfad in zwei Dimensionen.

Das Pfadmodell betrachtet die Mittelwerte der Teil-Gauß-Glocken mit maximaler Wahrscheinlichkeit der GMM-Kommandomodelle während einer Äußerung (Abb. 5). Die Dimension dieser Mittelwerte entspricht der Dimension der GMMs. Jeder Teil-Gaußglocke eines GMMs wird ein Index i zugeordnet und für jeden Block k und für jedes Modell eines Kommandos diejenige Teil-

Wort / Sprecher	**#1**	**#2**	**#3**	**#4**	**#5**	$\sum$
Person 1	20	20	21	20	19	100
Person 2	35	34	31	32	37	169
Person 3	32	34	31	33	33	163
Person 4	57	54	51	58	63	293
$\sum$	144	142	144	143	152	725

Tabelle 4: Größe des Testdatensatzes von Probanden.

Gaußglocke i_k des Modells m mit der maximalen Wahrscheinlichkeit gespeichert.

$$i_k^m = \arg\max_l g_l \mathcal{N}(\boldsymbol{x_k}|\boldsymbol{g_l^m}, \boldsymbol{\mu_l^m}, \boldsymbol{\Sigma_l^m}) \tag{4}$$

Am Ende einer Äußerung werden die zu den einzelnen Indizes gehörigen Mittelwertabfolgen μ_i der einzelnen Modelle mit denen von im Training zu den einzelnen Kommandos erstellen Mittelwertabfolgen verglichen.

$$\boldsymbol{d}^m = \sum_{k=1}^{K} (\boldsymbol{\mu}^{\boldsymbol{m}}_{\boldsymbol{i_k},\text{test}} - \boldsymbol{\mu}^{\boldsymbol{m}}_{\boldsymbol{i_k},\text{train}})^2 \tag{5}$$

Eine quadratische Abstandsfunktion gibt eine Distanz zwischen einer trainierten und aufgezeichneten Äußerung an. Die Distanzen werden über die Gesamtlänge der Äußerung in Blöcken K aufsummiert und die Äußerung mit der minimalen Distanz wird als die wahrscheinlichste angenommen.

$$m = \arg\min_m \boldsymbol{d}^m \tag{6}$$

Folglich wird dies neben den einzelnen Wahrscheinlichkeiten der GMMs als weiteres Entscheidungskriterium herangezogen werden kann.

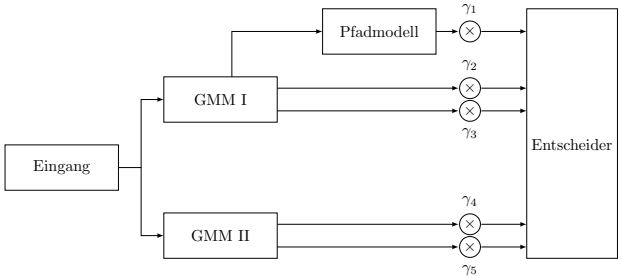

Abbildung 6: Erkennerstruktur

Test- und Trainingsdaten

Für das Training und die Tests der Daten wurden Trainingsdaten in Pools und Schwimmbädern aufgenommen. Insgesamt wurden Daten von vier Sprechern erhoben, zwei männlichen und zwei weiblichen Geschlechts.

Alle Kommandos wurden gesprochen, während der Kopf vollständig unter Wasser war. Die Probanden wurden gebeten, so normal wie unter diesen Umständen möglich

	Pufferlänge	γ
GMM I (Variante I)	64ms	0,1
GMM II (Variante I)	256ms	0,1
GMM I (Variante II)	64ms	0,1
GMM II (Variante II)	256ms	0,1
Pfad	-	0,6

Tabelle 5: Gewichte der einzelnen Pfade.

zu sprechen und mit kurzen Pausen so lange die Worte zu artikulieren, bis wieder Luft geholt werden musste. Hörbeispiele der Kommandos sind unter [8] zu finden. Durch die verhältnismäßig geringe Anzahl von Trainingsdaten für die Mustererkenner, wurde eine Aufteilung von 90 % Training und 10 % Test gewählt.

Auswertung

Das entscheidende Qualitätskriterium für einen Mustererkenner ist die Erkennungsrate. Die phonetische Ähnlichkeit der gewählten Äußerungen sorgt dafür, dass bei den beiden GMMs mit unterschiedlicher Blocklänge mit beiden Varianten der Auswertung keine optimale Erkennung und Trennung der einzelnen Kommandos möglich ist. Das GMM mit einer Blocklänge von 64 ms erreicht mit der Auswertevariante 2 (Abb. 4) eine Gesamtdetektionsrate von 55,9 %, das GMM mit 256 ms von 46,7 %. Wird jedoch das Pfadmodell, bzw. das Verfolgen der Mittelwerte (GMM-MVT) betrachtet, steigt die Erkennungsrate deutlich. Auch die Verwechslungsrate mit ähnlich klingenden Worten sinkt, sodass eine Erkennungsrate von 80,0 % realisiert wird.

Da die klassischen GMMs auch einen Beitrag zur Detektion leisten, können diese ebenfalls in das Gesamtergebnis mit einbezogen werden. Für das Kombinieren der Ergebnisse werden die Erkennerausgänge gewichtet und aus dem Ergebnis die Entscheidung getroffen. Die Gewichtungsfaktoren γ sind in Tabelle 5 abgebildet.

Durch die Kombination aller zur Verfügung stehenden Erkenner steigt die Erkennungsrate um weitere 1,7 %. Die Erkennerperformanz lässt sich aus der Verwechslungsmatrix [9] ablesen. Diese zeigt neben den korrekt klassifizierten Äußerungen auch die Verwechslungen zwischen den Äußerungen an. Während „stop“ sehr sicher klassifiziert wurde, zeichnen sich die phonetischen Ähnlichkeiten zwischen „play“, „previous“ und „pause“ sehr deutlich ab.

Fazit und Ausblick

Es wurde eine Möglichkeit vorgestellt, direkt ins Wasser gesprochene Äußerungen mittels einer Kombination verschiedener GMMs und deren Erweiterung zu erkennen. Der GMM-MVT-Ansatz bringt entscheidende Verbesserungen in der Detektionsrate. Die Kommandos wurden genutzt, um einen MP3-Player an einer Schwimmbrille zu bedienen. Während bisher nur eine kabelgebundene Versorgung in einer wasserdichten Box in Kombination mit einer MATLAB-basierten Verarbeitung verfügbar ist, könnte im nächsten Schritt das System weiter in-

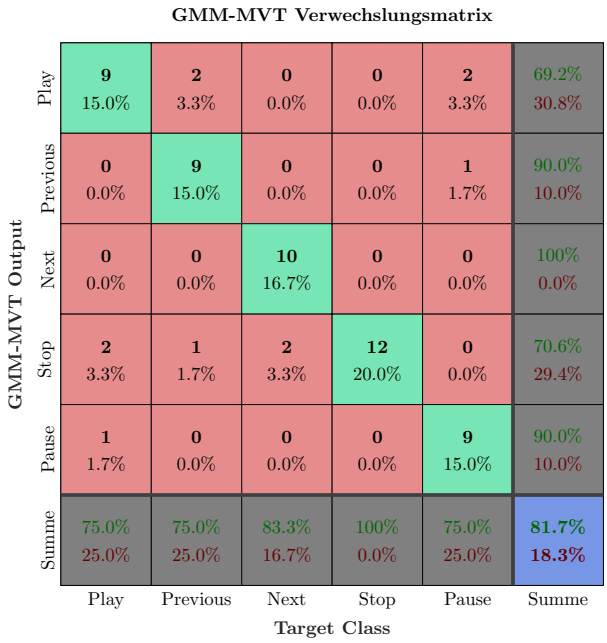

Abbildung 7: Gesamtergebnis als Verwechslungsmatrix.

tegriert werden, sodass es direkt an der Schwimmbrille Platz findet. Zusätzlich wurden bereits einige Versuche mit neuronalen Netzen unternommen, da diese durch ihre diskriminative Unterscheidungscharakteristik Vorteile gegenüber den auf Gemeinsamkeit zu bereits trainierten Modellen abzielenden GMMs haben könnten. Erste Versuche mit kleineren neuronalen Netzen zeigten bereits Ergebnisse, welche denen der GMMs teilweise überlegen sind. Eine weitere Optimierung der Detektionsrate würde das Einführen einer Zustandsbeobachtung des Systems erbringen. Beispielsweise würde der vorgeschlagene MP3-Player im „stop“-Zustand kein „pause“-Kommando sinnvoll annehmen können, sodass evtl. eher vom Benutzer das Kommando „play“gemeint gewesen sein könnte.

Literatur

[1] `https://dss.tf.uni-kiel.de/index.php/teaching/projects/comm-unit-diving-mask`, Stand: 07.08.2017

[2] Lurton, X. : An Introduction to Underwater Acoustics, Springer, 2011

[3] E.Hering, R.Martin, M.Stohrer: Physik für Ingenieure, Springer Verlag, 2007

[4] Wisch, T.O. : Spracherkennung in stark gestörten Unterwasserumgebungen unter Nutzung von wasserfesten Mikrofonen, Masterarbeit, DSS Uni Kiel, 2017

[5] Logan, B. : Mel Frequency Cepstral Coefficients for Music Modelling, McGraw-Hill, 1975

[6] Akaike, H.; Lovric, M. : Akaikes Information Criterion, International Encyclopedia of Statistical Science, 2011

[7] Geierhofer, S. : Feature Reduction with Linear Discriminant Analysis and its Performance on Phoneme Recognition, ECE272 - Individual Study in ECE Problems, 2004

[8] `https://dss.tf.uni-kiel.de/index.php/research/publications/publications-add-material/underwater-speech-examples`, Stand: 14.03.2018

[9] K.M. Ting: Confusion Matrix, Encyclopedia of Machine Learning, Springer, 2010