Automatische Schätzung wichtiger Nachhallparameter

Heiko Gölzer und Michael Kleinschmidt
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
heiko.goelzer@mail.uni-oldenburg.de

I. Einleitung

Die Erkennungsleistung aktueller automatischer Spracherkennungssysteme leidet erheblich unter dem Einfluss von Nachhall. Dies gilt bereits für moderate Nachhallsituationen in alltäglicher Umgebung. Für viele Anwendungen steht als Information über die Raumcharakteristik nur die implizit im verhallten Sprachsignal enthaltene Metainformation zur Verfügung. Da eine Enthallung der Signale selbst bei guter Schätzung der zugrunde liegenden Raumimpulsantwort schwer möglich ist, bietet sich stattdessen eine Adaptation der Merkmalsextraktion oder des Klassifikators an. Dazu ist es notwendig, die für die Spracherkennung wichtigen Parameter der Übertragungsfunktion aus dem verhallten Sprachsignal extrahieren zu können. Mit Hilfe des hier vorgestellten Verfahrens kann die Nachhallzeit und das Verhältnis von Direktschall zu Nachhallanteilen auch bei unbekanntem Sprachmaterial abgeschätzt werden. Als Merkmale werden AmplitudenModulationsSpektrogramme (AMS) verwendet, die aus dem kontinuierlichen Sprachsignal extrahiert werden. Zur Schätzung dient ein mehrschichtiges neuronales Netz, welches unter Benutzung künstlich verhallter Sprachsignale trainiert wird.

II. Automatische Schätzung

Zur Charakterisierung räumlicher Nachhallsituationen wurde hier neben der breitbandig berechneten Nachhallzeit T_{60} der in Anlehnung an die DIN EN ISO 3382 [din00] wie folgt definierte Früh-zu-spät-Index C_{30} verwendet.

$$C_{30} = 10\log\left(\int_0^{30ms} p^2(t)\,dt \;/\; \int_{30ms}^{\infty} p^2(t)\,dt\right) dB$$

Dabei ist $p^2(t)$ die quadrierte Raumimpulsantwort. C_{30} dient als Maß für das Verhältnis von frühen zu späten Reflexionen. Im Gegensatz zum gewöhnlich für Sprache verwendeten C_{50} [din00] wird hier eine Grenze von 30 ms Verzögerung zur Unterscheidung zwischen frühen und späten Reflexionen verwendet da dies für automatische Spracherkennungssysteme zu- und abträgliche Nachhallanteile besser voneinander abgrenzt. Die beiden Parameter T_{60} und C_{30} sollen für eine Raumsituation geschätzt werden, über die als einzige Information ein verhalltes Signal kontinuierlicher Sprache vorliegt. Dazu werden aus dem Sprachsignal AmplitudenModulationsSpektrogramme (AMS) berechnet, welche als Eingangsmuster für ein mehrschichtiges neuronales Netz dienen. Der generelle Aufbau des Algorithmus ist in Abbildung 1 wiedergegeben.

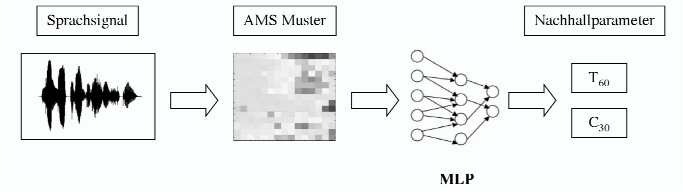

Abb. 1: *Verarbeitungsschritte*

A. Verwendete Merkmale

Als Merkmale für die Schätzung von Nachhallparametern werden AmplitudenModulationsSpektrogramme (AMS) verwendet. Diese von Kollmeier und Koch [kol94] vorgeschlagene Signalverarbeitung wurde durch psychoakustische und neurophysiologische Erkenntnisse motiviert. AMS sind 2-dimensionale Muster, welche die Modulationstiefe bei bestimmten Frequenzen und Modulationsfrequenzen abbilden. Tchorz und Kollmeier [tch00, tch02] verwendeten derartige Muster erfolgreich zur Kurzzeitschätzung des Signal-Rausch-Verhältnisses. Es erscheint sinnvoll den Einfluß von Nachhall auf Sprache im Modultionsspektrum in dem Bereich zu suchen und zu modellieren, dem Wichtigkeit für die Sprachverständlichkeit zugesprochen wird. Daher werden hier anders als bei [kol94] und [tch00, tch02] niedrige Modulationsfrequenzen zwischen 1 Hz und 25 Hz betrachtet. Die Bedeutung der Modulationsübertragungsfunktion (MTF) für die Evaluierung der Übertragungseigenschaften von Räumen im Bezug auf Sprache ist von Houtgast und Steeneken [hou85] gezeigt worden. Der charakteristische Einfluss von Nachhall auf die MTF ist danach eine Modulations-Tiefpassfilterung. Im Gegensatz dazu hat additives weisses Rauschen eine breitbandige Verringerung der Modulationstiefe zur Folge. Die hier verwendeten AMS Muster bilden eben solche charakteristischen Veränderungen der Modulationstiefe frequenz- und modulationsspezifisch ab.

B. Berechnung der Merkmale

Das verhallte Sprachsignal wird mittels Overlap-Add Verfahren mit 10 ms Vorschub in 20 ms lange Fenster zerlegt. Jedes Segment wird mit einem Hanning Fenster multipliziert und mit Nullen verlängert, um dann durch Fouriertransformation mit 512 Koeffizienten in ein komplexes Spektrum gewandelt zu werden. Die so gewonnenen komplexen Werte bilden für jeden Frequenzkanal ein Zeitsignal. Durch entsprechende Aufsummation der Energie benachbarter Kanäle wird die Frequenzachse in sieben Oktavbänder mit Mittenfrequenzen von 125 Hz - 8000 Hz unterteilt. Nach dem Ziehen der Wurzel liegt eine spektrogrammartige Repräsentation vor. Jeder Frequenzkanal wird mittels eines zweiten Overlap-Add-Verfahrens mit 5 s Vorschub in 10 s Fenster geteilt. Wiederum wird mit einem Hanning Fenster multipliziert und mit Nullen verlängert, um nach Fouriertransformation mit 1024 Koeffizienten ein komplexes Modulationsspektrum für jeden Frequenzkanal zu berechnen. Nach Bildung der Betragsquadrate werden durch Mittelwertbildung 14 Modulationsfrequenzkanäle erzeugt. Diese sind äquidistant auf einer logarithmischen Skala und decken den Bereich von 1 Hz - 25 Hz ab. Nach erneutem Wurzel ziehen werden in einem letzten Schritt die Amplituden des resultierenden zweidimensionalen Musters logarithmisch komprimiert. Die Berechnung der AMS ist vereinfacht in Abbildung 2 als Blockdiagramm dargestellt.

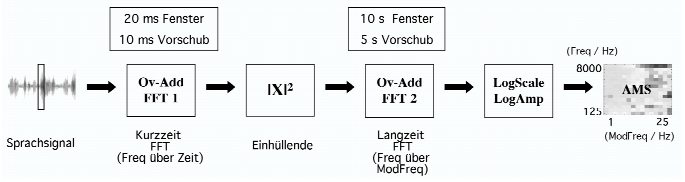

Abb. 2: *Berechnung der AmplitudenModulationsSpektrogramme (AMS)*

C. Simulation von Raumimpulsantworten

Für den Versuch wurden mit einem Spiegelschallquellenmodell Raumimpulsantworten simuliert. Ein Softwarepaket der Firma DSP Algorithms [1] wurde benutzt um für gegebene Abmessungen rechteckiger Räume, Absorbtionskoeffizienten der Wände und frei wählbare Quellen- und Empfängerpositionen Raumimpulsantworten zu berechnen. Die Ausdehnung der Räume wurde für die Querseiten, Längsseiten und die Höhe auf einen Bereich von 3 m – 10 m, 3 m – 15 m und 2.5 m – 4 m beschränkt. Die Absorbtionskoeffizienten wurden aus einem Bereich von 0.6 – 0.95 gewählt. Die Positionen von Quelle und Empfänger wurden in der Höhe fest mit 1.6 m und 1.5 m vorgegeben und sonst auf je eine Hälfte des Raumes beschränkt. Die Raumabmessungen wurden so gewählt, dass sich eine möglichst plausible Verteilung der beiden zu schätzenden Parameter ergibt. Insgesamt wurden für

[1] ROOM, www.dspalgorithms.com

zufällige Parameter 1000 Raumimpulsantworten generiert, die dann im Verhältnis 9/1 in disjunkte Gruppen für Training und Test geteilt wurden. Die zwei Parameter Nachhallzeit T_{60} sowie Früh-zu-spät-Index C_{30} wurden auf Basis der jeweiligen Raumimpulsantwort berechnet. Es ergeben sich Verteilungen wie sie in Abbildung 3 dargestellt sind.

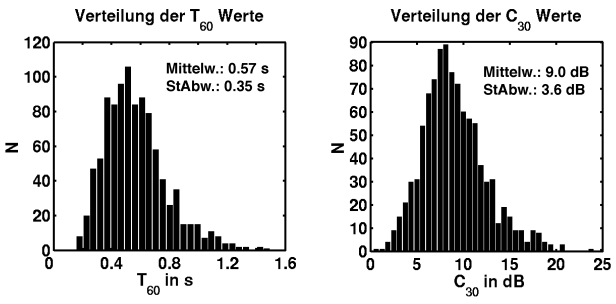

Abb. 3: *Verteilung der Parameter a) Nachhallzeit T_{60} und b) Früh-zu-spät-Index C_{30} für die verwendeten Impulsantworten*

D. Sprachmaterial

Als Basis für das Experiment dient ein Teil der Sprachdatenbank PhonDat1 [2]. Es handelt sich um zwei Geschichten, die von 90 verschiedenen deutschsprachigen SprecherInnen vorgelesen werden. Die Sprechdauer beträgt dabei zwischen 30 s und 140 s pro Person und liegt insgesamt bei etwa 2 h. Nach Aufteilung der Sprachsignale in disjunkte Gruppen für Training und Test im Verhältnis 9/1 wurden diese durch Faltung mit den simulierten Raumimpulsantworten künstlich verhallt.

E. Schätzer

Als Schätzer dient ein nichtlineares mehrschichtiges Neuronales Netz [3] mit einer verdeckten Schicht von 25 Neuronen und zwei Neuronen in der Ausgabeschicht. Es wird eine sigmoide Aktivierungsfunktion verwendet. Trainiert wird jeweils auf 9/10 aller AMS Merkmale als Eingangsvektoren (98 Elemente) und den vorher berechneten Parametern als den Zielwerten. Dazu wird ein Backpropagation-Trainingszyklus mit festgelegten Lernraten verwendet. Das austrainierte Netz dient zur Schätzung der Parameter mit dem restlichen 1/10 der Daten. Dieser gesamte Vorgang wird zehn mal mit jeweils disjunkten Trainings- und Testdaten wiederholt, um die Variabilität der Ergebnisse abschätzen zu können.

III. Ergebnisse

In Abbildung 4 ist die Korrelation zwischen Ziel- und Schätzwerten dargestellt. Die zugehörigen Korrelationskoeffizienten liegen bei 0.67 ± 0.04 und 0.73 ± 0.04 für die Nachhallzeit T_{60} und den Früh-zu-spät-Index C_{30}. Der RMS Klassifikationsfehler liegt bei $0.32s \pm 0.02s$ und $2.7dB \pm 0.2dB$ für die beiden zu schätzenden Parameter.

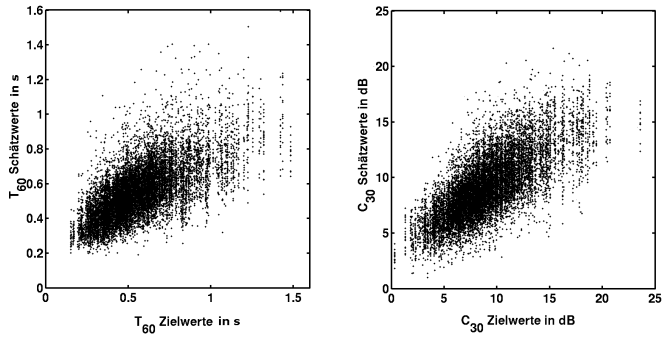

Abb. 4: *Korrelation zwischen Schätz- und Zielwerten*

[2] PhonDat 1, BAS, www.phonetik.uni-muenchen.de

[3] Quicknet, ICSI, www.icsi.berkeley.edu

A. Vergleich verschiedener Merkmale

In einem weiteren Versuch wurde evaluiert, welchen Einfluss die verschiedenen Anteile des AMS Musters an der Schätzung haben. Dazu wurde in Fall a) über alle Modulationskanäle im AMS gemittelt, wodurch ein Vektor mit sieben Elementen erzeugt wird, der rein spektrale Information trägt (spktr). Im Fall b) wurde über die Frequenzkanäle gemittelt, wodurch das Muster auf einen Vektor mit 14 Einträgen reduziert wird, eine Art breitbandiges Modulationsspektrum (mod). Im Fall c) schliesslich wurden die Information aus Fall a) und Fall b) durch aneinanderhängen der beiden Merkmalsvektoren kombiniert (komb). Fall d) ist das schon ausgewertete vollständige AMS Muster mit 98 Elementen (voll). Es zeigt sich, dass die rein spektralen Merkmale für die Schätzung wenig auswertbare Information tragen. Allein das breitbandige Modulationsspektrum in Fall b) führt schon auf ein weitaus besseres Ergebnis. Die Kombination von spektralen und temporalen Merkmalen wie in Fall c) führt zu einer weiteren Verbesserung, während das komplette AMS Muster nötig ist, um eine bestmögliche Korrelation von Ziel- und Schätzwerten zu erreichen. In Abbildung 5 ist ein Vergleich der Korrelationskoeffizienten zwischen Schätzwerten und berechneten Zielwerten für die vier verschiedenen Merkmalsklassen wiedergegeben.

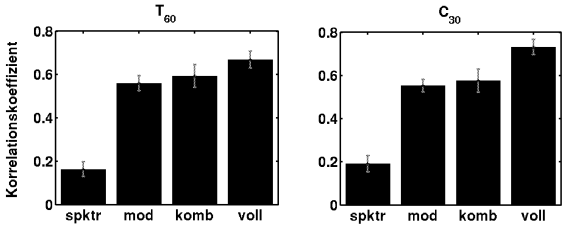

Abb. 5: *Korrelationskoeffizienten zwischen Schätz- und Zielwerten für verschiedene Merkmalsklassen: a) Spektrum allein, b) Modulationsspektrum allein, c) Kombination aus a) und b), d) volle AMS Muster. Gegeben sind Mittelwerte und Standardabweichungen über 10 verschiedene Testdatensätze.*

IV. Zusammenfassung und Ausblick

Es konnte gezeigt werden, dass mit dem hier vorgestellten Verfahren die Nachhallzeit T_{60} sowie der Früh-zu-spät-Index C_{30} einer zu Grunde liegenden Raumimpulsantwort mit moderatem Fehler geschätzt werden können. Als Basis für die Schätzung werden ausschließlich verhallte Sprachsignale verwendet.Es liegt also keine Kenntnis der Modulationsübertragungsfunktion vor, sondern lediglich die des Modulationsspektrums eines unbekannten Sprachsignals. Voraussetzung für die Schätzung ist die Extraktion vollständiger AMS Muster als Merkmale, die in charakteristischer Weise den Einfluss von Nachhall abbilden können.
Als weitere Arbeit ist die Nutzung der geschätzten Paramter für eine Modellauswahl bei automatischen Spracherkennungssystemen geplant. Dabei soll das vorliegende Verfahren genutzt werden, um bei unterschiedlichen Nachhallbedingungen durch Auswahl geeignet trainierter Modelle eine ausreichende Adaptation des Spracherkennungssystems an die jeweile Hallsituation zu gewährleisten.

Literatur

[kol94] Kollmeier, B. and Koch, R.: *Speech Enhancement Based on Physiologigal and Psychoacoustical Models of Modulation Perception and Binaural Interaction*, J. Acoust. Soc. Am., 95, p. 1593-1602, 1994.

[tch00] Tchorz, J. und Kollmeier, B.: *Schätzung des Signal-Rauschabstandes durch Analyse von Amplitudenmodulationen*, Fortschritte der Akustik - DAGA 2000, Oldenburg, DEGA Hrsg. A. Sill, p. 366-367, 2000.

[tch02] Tchorz, J. and Kollmeier, B.: *Estimation of signal-to-noise ratio with amplitude modulation spectrograms*, Speech communication 38, p. 1-17, 2002.

[hou85] Houtgast, T. and Steeneken, H.J.M.: *A rieview ot the MTF concept in room acoustics and its use for estimating speech intelligibility*, J. Acoust. Soc. Am., 77:1069-1077

[din00] DIN EN ISO 3382, Messung der Nachhallzeit von Räumen mit Hinweisen auf andere raumakustische Parameter, 2000.