Hierarchisches Modell zur Detektion von Sprache im Störgeräusch

Jörg-Hendrik Bach[1], Jörn Anemüller[2]
[1,2]*Abteilung Medizinische Physik, Universität Oldenburg, 26133 Oldenburg*
Email: j.bach@uni-oldenburg.de, joern.anemueller@uni-oldenburg.de

Einleitung

In diesem Beitrag wird eine hierarchische Methode zur Detektion von Sprache in Störgeräusch vorgestellt. Die Sprachdetektion basiert auf Amplitudenmodulationsmerkmalen, von denen bis zu 30 Merkmale ausgewählt worden sind, die den größten Beitrag zur Erkennungsrate leisten. In den Simulationen wurden Sprachsignale mit zwei real aufgenommenen Hintergrundgeräuschen (Straßenverkehr oder Fußgängerzone) bei Signal-Rauschabständen von -20dB bis 20dB abgemischt, der Klassifizierer wurde dann darauf trainiert, diese Sprachsignale von der reinen Hintergrundaufnahme zu unterscheiden. Diese Klassifikation wurde in einer hiearchischen Organisation aus drei Klassifizierern durchgeführt. Aus Vergleichsgründen wurde die Klassifikation auch mit einem binären Klassifizierer durchgeführt, der die verrauschte Sprache von reinem Hintergrundgeräusch unterscheidet.

Daten

Die verwendeten Sprachsignale entstammen dem TIMIT-Datensatz. Es wurden nur die Daten der „dialect region 1“ verwendet. TIMIT ist bereits in unterteilt in Trainings- und Testdatensatz, diese Einteilung wurde übernommen. Diese Sprachsignale wurden mit realem (additivem) Rauschen versehen: Es wurden zwei Aufnahmen von Hintergrundgeräuschen gemacht, einmal während eines Ganges durch eine Fußgängerzone, einmal an einer stark befahrenen Straßenkreuzung. Die Sprachsignale wurden mit diesen Geräuschen bei Signal-Rauschabständen von -20dB bis +20dB in 5dB-Schritten abgemischt, um eine SNR-abhängige Analyse durchzuführen. Es wurden in allen Fällen ca. 5-6 Minuten Audiomaterial pro Klasse zum Training verwendet.

Merkmale

Merkmalsberechnung

Die ausgewählten Merkmale sollten insbesondere robust gegenüber Rauschen sein, da die Simulationen auf Sprachdaten bis zu einem SNR von -20dB gerechnet wurden. Weiterhin sollten Sprechereinflüsse nicht allzu groß sein, da die verwendete Datenbank sowohl männliche als auch weibliche Sprecher enthält und außerdem die Sprecher im Trainings- von denen im Testmaterial verschieden sind. Aus der menschlichen und automatischen Spracherkennung ist diesbezüglich die Wichtigkeit der Modulationscharakteristika von Sprache bekannt, besonders der Einfluß des Bereichs zwischen 2Hz und 16Hz Modulationsfrequenz [1-3]. Dies findet auch Anwendung in der akustischen Szenenanalyse für den Hörgeräteeinsatz (z.B. [4]). Vor diesem Hintergrund wurden die Daten in der Repräsentation der sog. Amplitudenmodulationsspektrogramme (AMS) verwendet. Diese werden folgendermaßen berechnet (vgl. Abb. 1): Zunächst wird mit einer Kurzzeit-Fourier-Transformation ein spektrotemporales Muster mit einer Auflösung von 32ms (bei 4ms Vorschub) berechnet. Davon wird das Betragsquadrat gebildet und logarithmiert. Die Frequenzbänder werden auf eine Bark-Skala transformiert, die eine in etwa logarithmische Abbildung enthält. Innerhalb jedes der 17 Bark-Bänder wird eine Fourier-Transformation mit 1s langen Hann-Fenstern (0,5s Vorschub) durchgeführt.

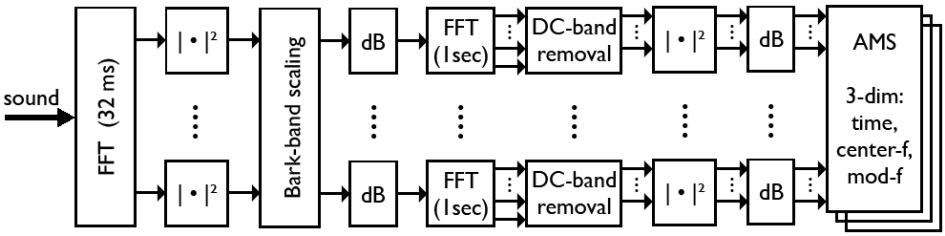
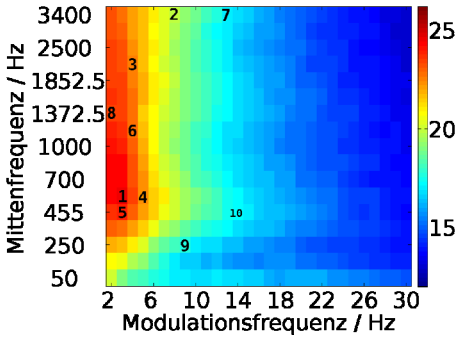


Abbildung 1: Berechnung der Modulationsmerkmale (AMS-Muster). Auf eine Kurzzeit-Einhüllenden-Extraktion folgt eine bandweise zweite Fouriertransformation, die in den Modulationsfrequenzraum transformiert. Für jedes 1s lange Fenster wird so ein Muster mit 17 Mittenfrequenzen und 29 Modulationsfrequenzen (d.h. 493 Merkmalen) erzeugt.

Die so erhaltenen Merkmale im Frequenz-Modulationsfrequenzraum werden oberhalb von 30Hz Modulationsfrequenz abgeschnitten, ebenso werden die unteren beiden Modulationsbänder entfernt, die den DC-Anteil enthalten. Dies resultiert in einer Repräsentation, die für jedes 1s-Fenster ein (17x29=)493-dimensionales AMS-Muster enthält.



Merkmalsauswahl

Abbildung 1: Beispiel für Merkmalsauswahl. Es sind die 10 wichtigsten Merkmale von Sprache in Fußgängerzonengeräusch (SNR=10dB) ausgezeichnet, die der iterative SFFS-Algorithmus findet. Das Hintergundbild ist das mittlere AMS-Muster dieser Signale gezeigt. Die Farbachse ist in willkürlichen Modelleinheiten.

Die Audiodaten wurden in eine 493-dimensionale Repräsentation transformiert (s.o.). Um die Dimensionalität zu reduzieren, wurde ein sog. „Sequential Floating Forward Search"-Algorithmus eingesetzt, um die 30 wichtigsten Merkmale zu identifizieren. Abb. 2 zeigt das Ergebnis der Auswahl für einen SNR von -15dB. Es ist klar zu erkennen, daß die Modulationsfrequenzen von 2-8Hz die wichtigste Information tragen, was die bekannten Ergebnisse bestätigt [1-3] Für die Mittenfreqeunzen ist der Trend weniger eindeutig, eine Häufung ist bei 500-600Hz - in der Nähe des ersten Formanten von Sprache - zu erkennen.

Methode

Die Spracherkennung wurde mit Support Vector Machines durchgeführt [5]. Dabei wurden zwei verschiedene Ansätze verglichen: Ein binäres Entscheidungsmodell, welches bei gegebenem SNR in den Hintergund eingebettete Sprache von reinem Hintergrundgeräusch unterscheidet, wobei beide aufgenommenen Hintergründe verwendet wurden, sowie ein hierarchisches Modell. Das hierarchische Modell besteht aus drei Knoten, von denen der erste darauf trainiert ist, die beiden Hintergründe voneinander zu unterscheiden, unabhängig davon, ob Sprache enthalten ist. Entsprechend dieser ersten Klassifikation wird ein zweiter Knoten nachgeschaltet, der auf die Unterscheidung von Sprache in diesem Hintergrund gegen den Hintergrund trainiert ist.
Dieser Vergleich zwischen binärem und hierarchischem Ansatz wurde für zwei verschiedene Merkmalssätze durchgeführt, erstens das volle AMS-Muster, zweitens die 30 besten ausgewählten Merkmale nach Merkmalsselektion (s.o).

Ergebnisse

Die Ergebnisse werden in ROC-ähnlicher Form dargestellt, d.h. es wird die Erkennungsrate der korrekt positiven gegen die Erkennungsrate der falsch positiven, d.h. der als Sprache erkannten Hintergrundsegmente, aufgetragen. Der „Schwellwert", der entlang einer Kurve variiert wird, ist der Trainings-SNR, d.h. der Signal-Rauschabstand, auf dem das Modell trainiert wurde. Für jeden Test-SNR erhält man auf diese Weise eine charakteristische Kurve. Die Ergebnisse der beiden Klassifikationsansätze werden jeweils gegenübergestellt, außerdem werden die Ergebnisse unter Verwendung des vollen AMS-Musters mit denen unter Verwendung nur der 30 wichtigsten AMS-Merkmale verglichen. Im SNR-Bereich von -20dB bis +15dB liegt das hierarchische Modell über dem binären, bei +20dB befinden sich beide Modelle in idealer Sättigung bereits bei beliebig kleinen falsch positiven. Die Verwendung des gesamten Musters erhöht die Performance des hierarchischen Modells stärker als die des binären Klassifizierers.

Diskussion

Die Merkmalsselektion bestätigt psychophysikalische Ergebnisse. Der vorgeschlagene hierarchische Ansatz ist im gesamten SNR-Bereich besser als ein binärer Klassifizierer, der auf denselben Daten trainiert wurde.

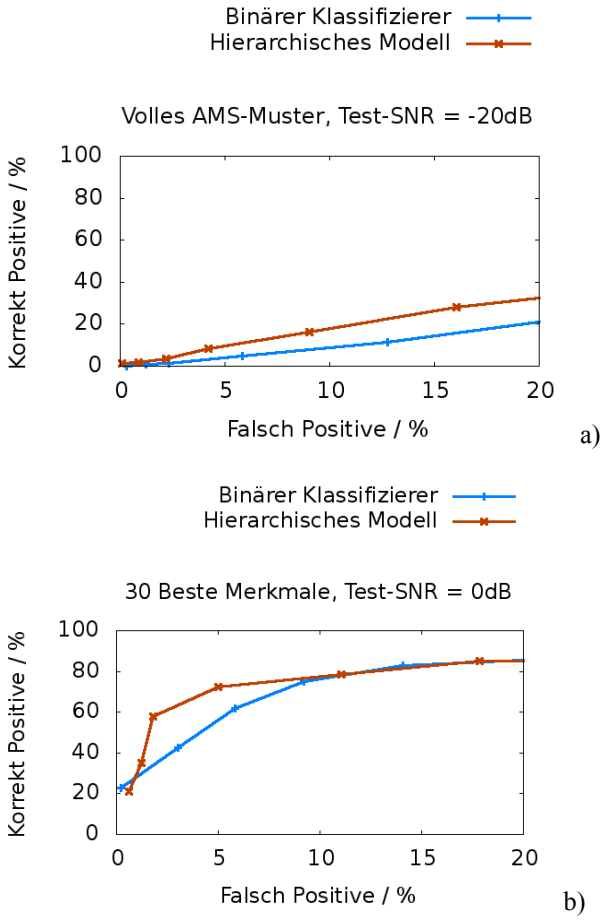


Abbildung 3: Ergebnisse der Klassifikation von Sprache in Fußgängerzonen- und Straßengeräusch. 3a) Das hierarchische Modell ist 1. besser als das binäre Modell, 2. auch bei -20dB SNR noch deutlich über der Diagonalen (Ratewahrscheinlichkeit). 3b) Das hierarchische Modell zeigt bessere Performance als das binäre; im Bereich hoher falsch positiver nähern sich die beiden Modelle einander an. Dies ist über alle SNR der Fall. Insgesamt liegen die Kurven des gesamten AMS-Musters etwas oberhalb derer der 30 besten Merkmale. Für das hierarchische Modell ist dies deutlicher ausgeprägt als für das binäre Modell.

Eine Erweiterung der Hierarchie auf Klassen von Hintergründen (z.B. Außen/Innenumgebungen) erscheint daher sinnvoll. Die Tatsache, daß der Trainings-SNR die Rolle des Schwellwert in den ROC-Plots spielt, läßt darauf schließen, daß ein bei niedrigem SNR trainiertes Modell auch die „leichtere" Aufgabe von Sprachdetektion bei höherem SNR mit guter Leistung löst.

Literatur

[1] Drullman et al, "Effect of temporal envelope smearing on speech reception", JASA 95(2), 1994
[2] Houtgast & Steeneken, "A review of the MTF concept in room acoustics and ist use for speech intelligibility in auditoria", JASA 77(3), 1985
[3] Kanedera et al, "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Comm. 28, 1999
[4] Büchler et al, "Sound Classification in Hearing Aids inspired by auditory scene analysis", EURASIP J. APP. Sig. Proc. 18 (2005)