Lombard-Sprache für Kfz-Anwendungen: eine Analyse verschiedener Aufnahmekonzepte

Markus Buck, Hans-Jörg Köpf, Tim Haulick

Harman/Becker Automotive Systems GmbH, Söflinger Str. 100, 89077 Ulm, Email: mbuck@harmanbecker.com

Einleitung

Durch Verwendung von Lombard-Sprache lassen sich Spracherkenner unter spezifischen Randbedingungen aufwandsgünstig testen bzw. trainieren. In diesem Beitrag werden verschiedene Konzepte zur Aufnahme von Lombard-Sprache anhand einer Messreihe untersucht.

Lombard-Sprache für Kfz-Anwendungen

Die akustischen Merkmale von Sprache hängen stark von der Umgebung ab, in der sich der Sprecher befindet. So spricht der Sprecher bei lautem Hintergrundgeräusch in der Regel lauter und deutlicher, um seinen eigenen Artikulationsvorgang besser kontrollieren zu können bzw. um für den Gesprächspartner verständlich zu sein. Dieser sogenannte Lombard-Effekt tritt auch unter anderen Randbedingungen auf, beispielsweise in halliger Umgebung oder unter Stress. Im Vergleich zur Sprache in ungestörter Umgebung ergeben sich akustische Unterschiede wie beispielsweise die Erhöhung der Sprachgrundfrequenz, eine Zunahme des Sprachpegels, eine längere Dauer von Vokalen sowie eine Verschiebung der Formantfrequenzen [1].

Für die Entwicklung von Spracherkennern sind umfangreiche Sprachdatensätze notwendig, die zum Training bzw. zur Evaluation von Erkennern in off-line Tests eingesetzt werden. Für robuste sprecherunabhängige Spracherkenner, wie sie im Kfz eingesetzt werden, sollte die Sprachdatenbank eine große Anzahl von Sprechern umfassen, sowie eine Vielfalt an Geräuschsituationen und Fahrzeugtypen abdecken. Für gezielte Untersuchungen von spezifischen Szenarien wie beispielsweise einem speziellen Fahrzeugtyp, einem speziellen Mikrofontyp bzw. einer speziellen Mikrofonposition sind diese Sprachkorpora jedoch nicht geeignet, da die Untermenge an passendem Sprachmaterial meist zu klein ist, um aussagekräftige Erkennertests durchführen zu können. Für solche Tests müsste jeweils geeignetes Sprachmaterial aufgezeichnet werden, was einen großen Zeitaufwand und hohe Kosten bedeutet.

Eine flexible und kostengünstige Möglichkeit, um spezifisches Sprachmaterial zu erhalten, ergibt sich durch die Trennung von sprecherspezifischen Daten und fahrzeugspezifischen Daten, wie in Bild 1 dargestellt. Die Mikrofonsignale werden mittels vorher aufgezeichneter fahrzeugunabhängiger Sprachäußerungen sowie in der Zielumgebung gemessener Impulsantworten und Geräuschaufnahmen simuliert. Die Sprachdaten werden dazu in hallarmer und ungestörter Umgebung aufgezeichnet, wobei dem Sprecher über Kopfhörer Fahrgeräusch eingespielt wird, um den Lombard-Effekt hervorzurufen. Um zu gewährleisten, dass die so aufgezeichneten Sprachdaten keine Raum- bzw. Mikrofoneinflüsse beinhalten, sollte ein Headset-Mikrofon mit ebenem Frequenzgang verwendet werden. Zur Aufnahme einer Lombard-Sprachdatenbank können verschiedene Geräusche eingesetzt werden. Bei der späteren Anwendung werden dann nur solche Äußerungen aus der Datenbank gewählt, für die die Aufnahmebedingungen zum zugemischten Geräusch passen. Neben der in Bild 1 skizzierten Simulation von Mikrofonsignalen lässt sich Lombard-Sprache auch direkt über einen Kunstkopf akustisch ausgeben, was für Erkenner- und Systemtests im Kfz genutzt wird [2].

Die Impulsantworten werden im Zielfahrzeug gemessen. Sie hängen von mehreren Randbedingungen ab, u. a. vom Fahrzeugtyp und der Art der Innenausstattung, von der Position und der Orientierung des Sprechers sowie von der Position, der Orientierung und dem Typ des Mikrofons. Ebenso hängt das Fahrgeräusch von zahlreichen Randbedingungen ab.

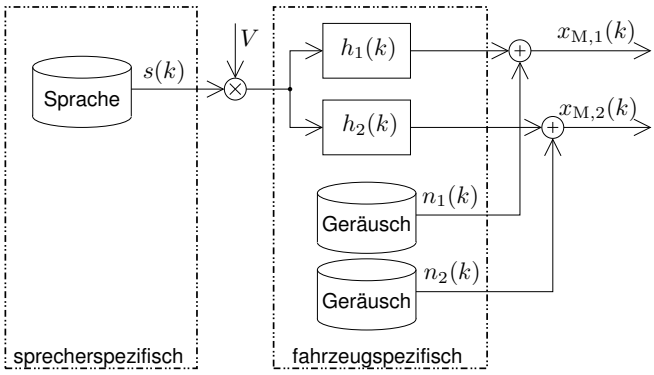

Bild 1: Generierung von Lombard-Daten für zwei Mikrofone.

Bild 1 zeigt die Simulation von zwei Mikrofonsignalen, wie sie z. B. für ein Mikrofon-Array notwendig sein kann. Das Sprachsignal $s(k)$ wird mit den gemessenen Impulsantworten $h_1(k)$ und $h_2(k)$ gefaltet und mit den Geräuschaufnahmen $n_1(k)$ und $n_2(k)$ addiert. Der Verstärkungsfaktor V ermöglicht die Anpassung des Sprachpegels, falls das zur Aufnahme der Lombard-Sprachdaten verwendete Geräusch nicht genau zur simulierten Geräuschsituation passt. Maßgebend ist dabei jeweils der Geräuschpegel an den Ohren des Sprechers.

Als Vorteile dieser künstlichen Generierung von Mikrofonsignalen sind die exakte Reproduzierbarkeit von Tests, sowie die Vergleichbarkeit zwischen Tests unter verschiedenen Randbedingungen zu nennen. Dabei erlaubt dieser

Ansatz, Tests sehr gezielt durchzuführen, beispielsweise zur Untersuchung einer bestimmten Geräuschsituation oder zur Ermittlung von optimalen Mikrofonpositionen im Fahrzeug. Zeitveränderliche Impulsantworten, wie sie in der Praxis häufig auftreten, können mit diesem Ansatz jedoch nur unzureichend nachgebildet werden.

Analyse verschiedener Aufnahmekonzepte

Um mit dem vorgestellten Ansatz realitätsnahe Mikrofonsignale generieren zu können, muss die Sprachdatenbank den Lombard-Effekt möglichst korrekt nachbilden. Zur Aufnahme von Lombard-Sprache wird üblicherweise den Sprechern über Kopfhörer Geräusch vorgespielt, während sie vorgegebene Äußerungen sprechen sollen. Um zu prüfen, ob dieses Aufnahmekonzept zu realistischer Lombard-Sprache führt, wurde eine Studie durchgeführt, die dieses Konzept mit Sprachaufnahmen bei realer Fahrt vergleicht. Für die Teilnehmer der Studie gliederten sich die Sprachaufnahmen jeweils in drei aufeinanderfolgende Phasen:

1. Der Sprecher befindet sich im stehenden Fahrzeug, Fahrgeräusch wird über Kopfhörer abgespielt, die Äußerungen sind durch eine Leseliste vorgegeben.
2. Reale Fahrsituation, die Äußerungen sind durch eine Leseliste vorgegeben.
3. Reale Fahrsituation, der Sprecher bedient ein Sprachdialogsystem lediglich mit der Vorgabe eine bestimmte Aufgabe zu lösen, beispielsweise eine Telefonnummer einzugeben.

Die Studie wurde mit 19 Teilnehmern (12 Männer, 7 Frauen) in einer Mercedes C-Klasse durchgeführt. Die Sprecher saßen jeweils auf der Beifahrerseite, mussten das Fahrzeug also nicht lenken. Zur Aufnahme der Sprache wurde ein Headset vom Typ AKG C477 verwendet. Ein Messmikrofon war an der Kopfstütze in der Nähe des linken Sprecherohrs befestigt, um den Hintergrundgeräuschpegel zu ermitteln. Für alle Phasen wurden Sprachaufnahmen für die Situationen 0 km/h, 90 km/h sowie 160 km/h durchgeführt, was Geräuschpegeln von etwa 43 dB(A), 65 dB(A) bzw. 72 dB(A) entsprach. In Phase 1 wurde den Sprechern das Fahrgeräusch über offene Kopfhörer mit kalibriertem Pegel präsentiert, um den realen Höreindruck möglichst gut zu simulieren. Insgesamt hatte jeder Sprecher etwa 300 Äußerungen zu sprechen, die sich etwa zu gleichen Teilen auf die drei Phasen und die verschiedenen Geschwindigkeiten verteilten. In Phase 1 und 2 hatten die Versuchsteilnehmer jeweils dieselben Äußerungen zu sprechen. Das Ziel in Phase 3 war, die reale Situation eines Anwenders wiederzuspiegeln.

In einer off-line Analyse wurden für jede Äußerung der A-bewertete Störsignalpegel am Fahrerohr sowie der linear bewertete Sprachsignalpegel ermittelt, wobei der Sprachpegel von der Headset-Position auf den Mundreferenzpunkt (MRP) umgerechnet wurde. Für die Wertepaare wurden Regressionsgeraden berechnet. Bild 2 zeigt die Ergebnisse für die drei Phasen, getrennt für männliche und weibliche Sprecher.

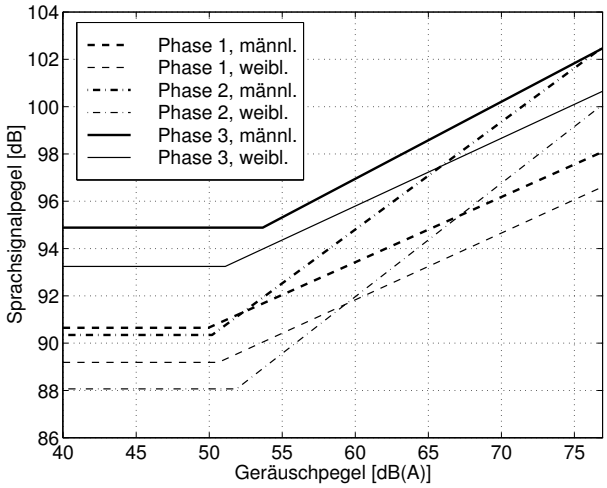

Bild 2: Sprachsignalpegel im MRP in Abhängigkeit vom Geräuschpegel am linken Sprecherohr.

Im Vergleich zu Phase 1 ergibt sich für Phase 2 ein stärkerer Anstieg des Sprachpegels über dem Geräuschpegel, was am realen Geräusch oder dem erhöhten Stress bei Fahrt liegen kann. In Phase 3 wurde bei geringem Geräusch lauter gesprochen als in Phase 2. Anstatt nur von einer Liste abzulesen, wird hier mit dem System kommuniziert, d. h. der Sprecher versucht sich stärker verständlich zu machen. Gemäß der ermittelten Regressionsgeraden steigt der mittlere Sprachpegel bei der realen Bedienung des Sprachdialogsystems (Phase 3) oberhalb von 55 dB(A) um etwa 0,3 dB/dB(A) über dem Geräuschpegel an. Bei geringem Hintergrundgeräusch ergibt sich hier ein mittlerer Sprachpegel von etwa 94 dB. Bei 65 dB(A) Geräuschpegel, was einer Fahrgeschwindigkeit von 90 km/h entspricht, beträgt der Sprachpegel etwa 98 dB. In Phase 1 liegt der Verlauf des mittleren Sprachpegels verglichen zu Phase 3 um etwa 4 dB tiefer. In allen drei Phasen sprachen die Männer im Mittel um etwa 2 dB lauter als die Frauen.

Zusammenfassung

Die drei untersuchten Aufnahmekonzepte führen zu abweichenden Sprachpegeln im Bereich von bis zu 4 dB. Für die reale Anwendungssituation und dem vorgestellten Aufnahmekonzept für Lombard-Sprache ergibt sich aber ein nahezu identischer Anstieg des Sprachpegels über dem Hintergrundgeräusch. Eine Korrektur kann daher mit einem konstanten Faktor erfolgen.

Literatur

[1] J.-C. Juncqua: The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. Speech Communication, Bd. 20, 1996, S. 13-22

[2] M. Lieb: Evaluating speech recognition performance in the car – a pragmatic approach, CFA/DAGA '04, Strasbourg, Frankreich, März 2004, S. 581-582.