Generierung von Sprachmaterial zum realitätsnahen Test von Sprach-erkennungssystemen für Kfz-Freisprecheinrichtungen

Frank Kettler, Marc Röber
HEAD acoustics GmbH, Ebertstraße 30a, 52134 Herzogenrath

1. Einleitung

Ein Sprecher adaptiert seine Sprechweise intuitiv an die akustischen Bedingungen. Diese als Lombard-Effekt bezeichnete Veränderung der Sprechweise hat eine höhere Sprachverständlichkeit zum Ziel. Es erscheint daher naheliegend zu untersuchen, in wie weit dies in Spracherkennungssystemen wie sie z.B. bei Kfz-Freisprecheinrichtungen eingesetzt werden, ausgenutzt werden kann, indem die durch den Sprecher „bereitgestellte“ adaptierte Sprechweise mit den veränderten charakteristischen Parameter der Sprache berücksichtigt wird.

Typischerweise wird das Training eines Spracherkenners mit neutraler Sprache in Ruhe durchgeführt. Im realen Einsatz kommen aber durch die Fahrgeräusche einerseits und den damit verbundenen Lombard-Effekt andererseits zwei für die Erkennungsrate unabhängige Einflussfaktoren hinzu. Um den Einfluss beider Faktoren von einander unabhängig untersuchen zu können, muss entsprechendes Testmaterial zur Verfügung stehen.

Sprachmaterial wurde unter Berücksichtigung des Lombard-Effektes aufgezeichnet. Das Aufnahmeszenario ist im folgenden geschrieben. Erste Analysen des Sprachmaterials und Tests wurden durchgeführt. Weitere Tests insbesondere mit unterschiedlichen Implementierungen von Freisprecheinrichtungen sind geplant.

2. Sprachmaterial und Aufnahmeszenario

Spracherkenner in Freisprecheinrichtungen lassen sich mit neutraler Sprache, die z.B. in Ruhe über ein Kunstkopf-Messsystem an Fahrerposition wiedergegeben wird testen. Für Tests in störschallerfüllter Umgebung stellen kommerzielle Datenbanken ein- oder auch mehrkanalige Aufnahmen mit Sprache und Störgeräusch zusammen bereit [1]. Solche Aufnahmen müssen direkt in den Spracherkenner eingespeist werden, Tests in Fahrzeugkabinen und somit in definierten Zielfahrzeugen und Fahrzeugtypen sind nicht möglich, da immer nur Sprache und Störgeräusch zusammen vorliegen. Der Einfluss des Lombard-Effektes auf die Erkennungsrate lässt sich aber isoliert nur testen und konsequenterweise auch ausnutzen, wenn Lombard-Sprache ohne Störgeräusch als Testmaterial vorliegt.

Aufnahmen wurden mit Versuchspersonen in einer realen Fahrzeugkabine (Mittelklasse), ausgestattet mit einer akustischen Fahrsimulation, durchgeführt. Die Sprecher saßen auf dem Fahrersitz. Um möglichst realitätsnahe Bedingungen nachzubilden, sollte von den Versuchspersonen durch Bedienung des Gaspedals eine vorgegebene Geschwindigkeit, die auf dem Tachometer angezeigt wurde, gehalten werden. Die Versuchspersonen trugen einen geschlossenen Kopfhörer zur Wiedergabe der Fahrgeräusche im Simulator. Die veränderte Eigenwahrnehmung durch den geschlossenen Kopfhörer kann durch entsprechende Rückkopplung der eigenen Stimme vermindert werde [2], jedoch ist davon auszugehen, dass der Kopfhörereinfluss auf den Lombard-Effekt gegenüber dem Störgeräusch selbst vernachlässigbar ist [3].

Aufgezeichnet wurde bei drei simulierten Geschwindigkeiten von 50 km/h (Fahrzeuginnnengeräuschpegel 58,4 dB(A)), 130 km/h (71,8 dB(A)) und 200 km/h (80,5 dB(A)) sowie ohne Kopfhörer (neutrale Sprache, Ruhegeräuschpegel 29,4 dB(A)). Um den Einfluss des geschlossenen Kopfhörers selbst zu beurteilen, wurde ebenfalls eine Aufnahme mit Tragen des Kopfhörers, jedoch ohne Störgeräuschwiedergabe durchgeführt. Die Versuchspersonen sprachen deutsch.

Für Tests der Spracherkennungssysteme im Kfz wurde spezifisches Testmaterial mit typischen Kommandowörter („Nummer wählen“, „Name wählen“, ...), Namen und Telefonnummern ausgewählt. Fortlaufend gesprochene Sätze waren so konzipiert, dass hieraus eine reale Konversation zwischen Festnetzteilnehmer und dem Benutzer einer mobilen Freisprecheinrichtung simuliert werden kann [4]. Diese Sätze können z.B. verwendet werden, um die Natürlichkeit der Sprache nach Übertragung über Störgeräuschreduktionsalgorithmen zu testen. Im Extremfall -bei einer zu hohen Störgeräuschreduktion- erscheint die Lombard-Sprache unnatürlich, da sie nicht mehr dem hörbaren Störgeräusch angepasst klingt.

Dieses so aufgezeichnete Sprachmaterial steht somit für Analysen und Testzwecke zur Verfügung.

3. Analysen des Sprachmaterials

Die ersten Aufnahmen wurden mit 3 männlichen Versuchpersonen durchgeführt. Der Einfluss des geschlossenen Kopfhörers selbst zeigt sich im Pegelunterschied zwischen der neutrale Sprache, aufgezeichnet ohne Kopfhörer und mit Tragen des Kopfhörers –aber noch ohne Störgeräuschdarbietung- von ca. 1,6 dB (gemittelt).

Im nachfolgenden sind für die 3 Sprecher für die 4 Aufnahmeszenarien die Sprachpegel gemittelt über eine Vielzahl der typischen Kommandowörter (Gesamtdauer ca. 20 s) analysiert. Der Sprachpegel wurde jeweils als Active Speech Level [5] bestimmt. Die Entfernung des Messmikrofons zum Mund-Referenzpunkt ist berücksichtigt. Zur besseren Übersicht sind die diskreten Messwerte durch Linien verbunden.

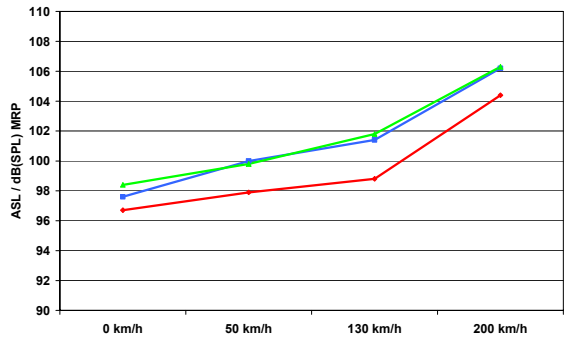

Abb. 1: Active Speech Level für 3 Versuchspersonen (Aufzeichnung bei 0 km/h ohne Kopfhörer)

Sprecherabhängige Unterschiede zwischen den Versuchspersonen betragen ca. 3 dB. Die Pegelvariation zwischen der neutralen Sprache und der Lombard-Sprache beträgt für jede Versuchsperson ca. 8 bis 9 dB. Zwischen den simulierten Fahrgeräuschen bei 200 km/h und 50 km/h beträgt der Pegelunterschied ca. 6 bis 7 dB wobei die Variation des Fahrgeräusches selbst mit einem Unterschied von 80,5 dB(A) und 58,4 dB(A) insgesamt ca. 22 dB beträgt. Durch den Lombard-

Effekt wird der Unterschied im Signal-Störabstand somit erwartungsgemäß abgesenkt jedoch nicht komplett ausgeglichen.

Interessanterweise liegt der Sprachpegel für die neutrale Sprache (in Ruhe, 0 km/h) mit ca. 97 dB_{SPL} am MRP (dies entspricht einen Wert von ca. +3 dB_{Pa}) und liegt somit deutlich höher als die in der Telefonometrie angenommenen -4,7 dB_{Pa} Nominalpegel. Die Versuchspersonen wurden zu Beginn des Tests darauf hingewiesen eine konstante Geschwindigkeit im Fahrsimulator zu halten und sich bewusst auf die Telefonsituation „zu konzentrieren“. Da sich hierbei um die Benutzung von Freisprecheinrichtungen handelt, kann man mit einer Sprachpegelanhebung von ca. 3 dB gegenüber dem Nominalpegel von –4,7 dB_{Pa} rechnen [6]. Die hier ermittelten Pegel liegen jedoch noch einmal um ca. 4 dB höher. Dies kann auch am analysierten Sprachmaterial liegen: unter Umständen werden Kommandowörter intuitiv lauter gesprochen. Hier müssen weitere Analysen mit heterogenem Sprachmaterial und weiteren Versuchspersonen ausgewertet werden.

Die typische, ausgeprägtere spektrale Struktur mit der Grundfrequenzanhebung bei der Lombard-Sprache ist in Abb. 3 und 4 gezeigt. Zugrunde gelegt wurde hier eine Sequenz von ca. 150 ms während eines gesprochenen „e“ im Wort „wegen“. Der Pegel beider Aufnahmen ist angepasst.

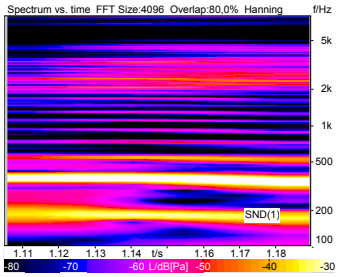

Abb. 3: Neutrale Sprache

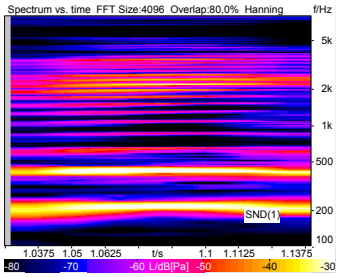

Abb. 4: Lombard-Sprache bei 200

Eine veränderte Zeitstruktur mit verlängerten Pausen fällt bei fortlaufende Sprache auf. Bei hohem Störgeräuschpegel wird aufgrund der Anstrengung nicht mehr zusammenhängend gesprochen. Für Störgeräuschreduktionsalgorithmen ergibt sich daraus die Forderung nach hoher Stabilität mit geringer eingefügter Dämpfung, um ein hörbares „pumpen“ des Störgeräusches in diesen Pausen zu vermeiden.

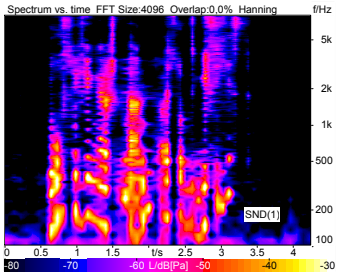

Abb. 5: Neutrale Sprache

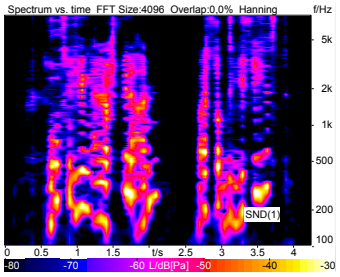

Abb. 6: Lombard-Sprache bei 200 km/h

4. Testmöglichkeiten

In den Tests wird ein Kunstkopfmesssystem in einem Fahrsimulator auf dem Fahrersitz positioniert. Über den künstlichen Mund wird die Lombard-Sprache wiedergegeben. Die entsprechenden Fahrgeräusche werden über die Fahrsimulation eingespielt. Getestet werden kann in diesem Szenario z.B. die Erkennungsrate bei der Wiedergabe von Namen aus dem zuvor mit neutraler Sprache trainierten Telefonbuch.

4.1 Statistische Erkennungsrate

Solche Tests, insbesondere Vergleichstests können statistisch z.B. mit der Angabe der Erkennungsraten, Vertrauensbereichen etc. ausgewertet werden. Motivation eines Tests ist es aber auch, die Unterschiede zwischen neutraler Sprache und Lombard-Sprache z.B. bei identischem Wiedergabepegel zu evaluieren. Hierbei wird der Pegel der Lombard-Sprache variiert.

4.2 Einfluss der Lombard-Sprachcharakteristik?

In einem Vorversuch wurde der Pegel der Lombard-Sprache bei der entsprechenden Geschwindigkeit abgesenkt, bis von 10 Versuchen nur noch 8-mal der korrekte Namen erkannt wurde. Die Angaben auf der y-Achse entsprechen diesen Pegelabsenkungen (80 %-ige Erkennungsrate). Die Tests wurden mit drei Namen durchgeführt (rot: „Karin Brings“, blau: “Katrin Frings“, grün: „Peter Schmidt“). Jedem Einzelergebnis liegen somit 10 Testdurchläufe zugrunde.

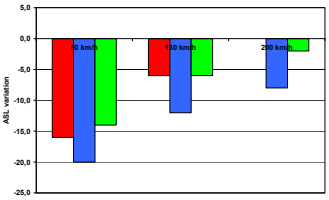

Abb. 7: Versuchsperson 1

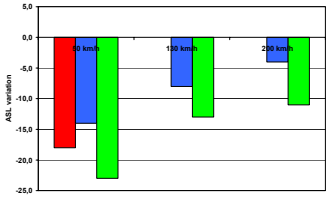

Abb. 8: Versuchsperson 2

Die linken drei Balken entsprechen der Fahrsituation bei 50 km/h, die mittleren 130 km/h und die rechten 200 km/h. Bei 50 km/h lassen sich die Sprachpegel für die drei Telefonbucheinträge bei dieser getesteten Freisprechereinrichtung um ca. 15 dB absenken, bei den höheren Geschwindigkeiten von 130 km/h bzw. 200 km/h sind die möglichen Pegelabsenkungen deutlich geringer. Tendenzen lassen sich schon aus diesen ersten Ergebnisse für diese Freisprecheinrichtung, die ohne adaptive Störgeräuschreduktion arbeitet, ableiten:

- Die Abhängigkeit der Erkennungsrate von der Fahrsituation ist für die verschiedenen Sprachsignale stark unterschiedlich (rot: starke Abhängigkeit, grün, blau: gering).
- die überproportionale Störgeräuschzunahme gegenüber dem Lombard-Effekt setzt die Erkennungsrate deutlich herab

Weitere Test mit unterschiedlichen Implementierungen der Spracherkennung werden sich anschließen. Neben Vergleichstests stellt sich die Frage, ob sich -bei gleichem Wiedergabepegel- mit Lombard-Sprache im Vergleich zu neutraler Sprache eine höhere Erkennungsrate erzielen lässt. Die konsequente Weiterführung dieses Gedankens besteht natürlich auch darin, diesen Effekt explizit auszunutzen.

5. Zusammenfassung

Durch die Trennung von Lombard-Sprache und Fahrgeräuschsimulation lassen sich interessante Punkte zu Testzwecken realisieren

- die realitätsnahe Aufzeichnung von Hörbeispielen zur Optimierung der Störgeräuschreduktion. Mit Lombard-Sprache lässt sich in einem beliebigen Zielfahrzeug die Natürlichkeit des Verhältnisses von übertragener Sprache und Hintergrundgeräusch beurteilen.
- Realistische Tests von Spracherkennern (insbesondere Vergleichstests) bei verschiedenen Geschwindigkeiten
- Evaluierung des Verbesserungspotential bei Spracherkennern

6. Literatur

[1] The Center of spoken Language Research CSLR, University of Colorado, http://cslr.colorado.edu/beginweb/speechcorpora/corpus.html

[2] Eigenwahrnehmung der Stimme in virtuellen akustischen Umgebungen, Christoph Pörschmann, DAGA 1998

[3] Eine Datenbank für deutsche Sprache mit Lombard-Effekt, Stefanie Köster, Christoph Pörschmann, Jürgen Walter, DAGA 2000

[4] Untersuchungen zur Erfassung der Konversationsqualität von Mobiltelefonen, Marc Röber, Diplomarbeit, Institut für Technische Akustik, RWTH Aachen, 2002

[5] ITU-T Empfehlung P.56, Objective Measurements of Active Speech Levels.

[6] ITU-T Empfehlung P.340, Transmission Characteristics of Handsfree Telephones.