Integriertes Konzept zur Evaluierung von Freisprecheinrichtungen im Kraftfahrzeug

Wilfried Gaedicke[1]

[1] *Volkswagen AG, 38444 Wolfsburg, Deutschland, Email: wilfried.gaedicke@volkswagen.de*

Einleitung

In den letzten zwei Jahrzehnten gab es viele Unternehmungen mit dem Ziel, die Sprachqualität von Telefonsystemen durch geeignete Algorithmen objektiv bewertbar zu machen. Das Resultat dieser kreativen Arbeit sind mehr als ein Dutzend von Formeln zur Quantifizierung von Störungen und einige etablierte Algorithmen zur Nachbildung der Resultate umfangreicher Probandentests. Der Schwerpunkt bei der Entwicklung dieser Verfahren lag bisher hauptsächlich in der Bewertung von Codecs und Signalverarbeitungsalgorithmen. Für diese Anwendungsfälle werden bereits hohe Korrelationen mit subjektiven Urteilen erreicht. Bei der umfassenden Bewertung von Freisprechsystemen im Kraftfahrzeug sind die Szenarien deutlich komplexer. Hierbei erweitert sich die Palette der qualitätsbeeinflussenden Faktoren zusätzlich um das im Kraftfahrzeug vorhandene Hintergrundgeräusch, sowie um die akustischen Eigenschaften der Fahrzeugkabine. Zur Entwicklung eines effizienten Algorithmus, der auch für die fahrzeugspezifischen Gegebenheiten geeignet ist, darf der Schwerpunkt nicht ausschließlich auf eine Abfolge prozeduraler Elemente gelegt werden, die in das Verfahren integriert werden. Stattdessen muss die komplette Kette - Gewinnung von Sprachmaterial, Probandenuntersuchungen und letztlich die Modellbildung - untersucht und optimiert werden. Gerade die Gewinnung von bewertetem Audiomaterial ist zeit- und kostenintensiv und auch die Arbeitsbedingungen bei der Aufzeichnung im fahrenden Fahrzeug erlauben oftmals keine präzisen und reproduzierbaren Resultate. Will man Besonderheiten, wie zum Beispiel *Doubletalk* untersuchen, erhöht sich dieser Aufwand nochmals erheblich. Ziel dieser Veröffentlichung ist es ein Gesamtkonzept zur effizienten Entwicklung eines Algorithmus zur Sprachqualitätsbewertung von Kfz-Freisprecheinrichtungen vorzustellen. Abschließend werden erste Ergebnisse der aktuellen Entwicklung vorgestellt.

Sprachdatengenerierung

Umfangreiche Sprachdaten bleiben die wichtigste Grundlage eines jeden Sprachbewertungsverfahrens. Zwar sind nicht für jedes erdenkliche Verfahren zur Sprachqualitätsbewertung umfangreiche Sprachdaten zur Modellbildung notwendig, spätestens jedoch bei der Überprüfung des Verfahrens werden bewertete Sprachdaten unverzichtbar. Schließlich soll es auch möglich sein, bei vorhandenem und verifiziertem Algorithmus, die zu bewertenden Daten auf einfache und reproduzierbare Weise zu erzeugen.

Bisher war es üblich, Sprachdaten direkt im Fahrzeug aufzuzeichnen. Hierzu wurde ein Kunstkopf notwendigerweise auf dem Beifahrersitz platziert und Sprachdaten wurden während der Fahrt wiedergegeben und am Freisprechmikrofon aufgezeichnet. Später wird mit diesem Gemisch aus Sprache und Hintergrundgeräusch im Labor ein Freisprechsystem am Mikrofoneingang gespeist und die zu bewertenden Daten werden am Sprachausgang eines Netzwerksimulators aufgenommen. Nachteilig hierbei ist, dass die Aufenthaltszeit im Fahrzeug stets der Länge der zu generierenden Sprachsequenz entspricht. Zur Herstellung von zusätzlichem Sprachmaterial (andere Sprecher, andere Inhalte) mit derselben Konfiguration ist es notwendig den kompletten Aufbau im selben Fahrzeug wiederherzustellen um zumindest im Stand erneut Sprachdaten am Freisprechmikrofon aufzuzeichnen. Sollen *Doubletalk*-Situationen auf diese Weise reproduzierbar nachgebildet werden, wird die Vorgehensweise noch wesentlich aufwändiger. So müssen die Sprachdaten direkt bei bestehender Verbindung mit dem fahrenden Fahrzeug aufgezeichnet werden, wobei am fernen Ende Sprache eingespeist wird. Soll vollständige Reproduzierbarkeit erreicht werden, ist zu gewährleisten, dass Sprache am fernen und am nahen Ende synchronisiert zueinander ablaufen.

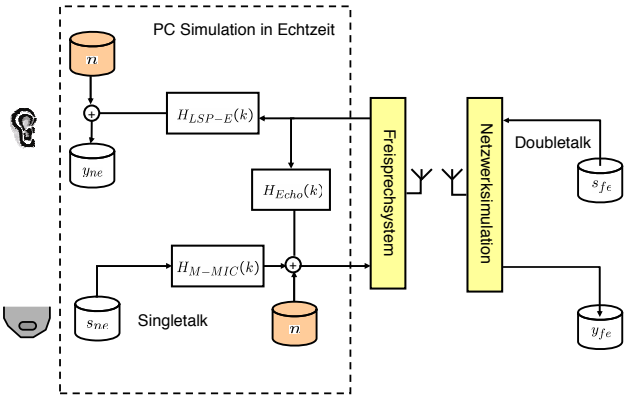

Abbildung 1: Echtzeitfähige PC-Simulation des Fahrzeuginnenraums mit angeschlossenem Freisprechsystem und verbundenem Netzwerksimulator für Doubletalk und Singletalk. Das Freisprechsystem wird in einem Prüfstand betrieben.

Das in Abbildung 11 dargestellte und bereits realisierte Verfahren vereinfacht die Erzeugung von Sprachmaterial erheblich. Das Konzept sieht vor, die Impulsantworten der Signalpfade im Fahrzeuginnenraum zu messen. Zusätzlich werden Datenbanken mit Fahrgeräuschen angelegt (bzw. das Fahrgeräusch wird vollständig synthetisiert). Mittels einer PC-Simulation werden die Signalpfade im Fahrzeuginnenraum in Echtzeit durch schnelle Faltungsalgorithmen und den aufgezeichneten Impulsantworten nachgebildet. Nur durch Simulation in Echt-

zeit und mit geringer Latenz der Simulation ist zu gewährleisten, das auch das Echo-Verhalten der Freisprechsysteme realistisch nachgebildet wird (Echokompensation adaptiert in Echtzeit). Erst hierdurch wird die Simulation von Doubletalksituationen möglich. Will man mehrere Freisprechsysteme miteinander vergleichen, bzw. Sprachmaterial für diese Systeme generieren, so müssen lediglich die an die Simulation angeschlossenen Freisprechsysteme ausgetauscht werden.

Probandenuntersuchungen

Wurde ausreichend viel Sprachmaterial generiert, können damit Probandenuntersuchungen durchgeführt werden. Hierbei ist zu beachten dass, das verwendete Sprachmaterial über einer geeigneten Skala gleichermaßen verteilt sein sollte. Will man Fahrgeräusche mit einbeziehen, so muss in den aufgezeichneten Sprachdaten das Komplette Spektrum von einem Mittelklassefahrzeug bei ausgeschaltetem Motor bis hin zu einem schlecht gedämpften Kleintransporter bei 180 km/h auf der Autobahn enthalten sein. Die Anzahl der zu bewertenden Sprachproben lässt sich wie folgt berechnen: Anzahl der Fahrzeuge x Anzahl der Geschwindigkeitsprofile x Anzahl der Freisprechsysteme x Anzahl der Softwareversionen x Anzahl der Sprechersituationen (*Singletalk/Doubletalk*). Der Vorteil einer Simulation zur Sprachdatengenerierung wird hier klar ersichtlich, da die Generierung der einzelnen Kombinationen fast vollständig automatisiert werden kann. Für die Probandenuntersuchungen wird ein geeignetes Werkzeug eingesetzt, das die Probandendaten entgegen nimmt und aufbereitet, um diese nahtlos zur Modellbildung bzw. -verifizierung zu verwenden [2].

Modellierung der Probenuntersuchungen

Die Gliederung von Algorithmen zur Sprachqualitätsschätzung die sowohl das originale als auch das zu bewertende Signal verarbeiten und die von uns implementiert wurden, sieht folgendermaßen aus: Zuerst erfolgt eine Vorverarbeitung der Sprachsignale, dann eine Parameterextraktion und schließlich die Abbildungsfunktion, die einen Zusammenhang zwischen den extrahierten Parametern und dem Qualitätsmaß herstellt. Die Vorverarbeitung enthält Elemente zur Pegelanpassung der Signale, sowie Elemente zur Kompensation von Effekten, die wenig oder keinen Einfluss auf das Sprachqualitätsmaß haben (z. B. mittlerer Frequenzgang, geringfügige Frequenzverschiebungen.) Die folgende Parameterextraktion kann sich auf einzelne Parametersätze wie z. B. ein Lautheitsmodell beschränken. Erfolgsversprechend erscheint auch die Einbindung von Kombinationen anderer Sprachqualitätsmerkmale. Die Parameterextraktion wurde an dieser Stelle als Plug-in Konzept realisiert, mit dem sich jederzeit neue Parameter und Sprachqualitätsmaße in den Algorithmus einbinden lassen. Um geeignete Abbildungsfunktionen zu generieren, kommen zum Beispiel neuronale Netze oder Regressionsfunktionen zum Einsatz. In unserer Implementierung wurde hierfür ein auf der Schätztheorie basierender *state-of-the-art* Ansatz gewählt. Hierfür werden für die Kombination u aus Parametervektoren x und dem Qualitätsmaß y mittels EM-Algorithmus Modellverteilungen berechnet und über einen MMSE-Ansatz wird eine Beziehung zwischen der Qualitätsdimension und den Parameterdimensionen hergestellt [3].

$$u = [y, x]^T \tag{1}$$

$$\epsilon_{MSE} = E[(y - \hat{f}(x))^2] \tag{2}$$

Erweiterte MMSE Schätzung

Eine mit diesem Paper eingeführte Neuerung bzw. Erweiterung des MMSE-Modells sieht eine iterative Verbesserung des EM-Modells vor. Aus Gründen praktischer Durchführbarkeit liegt pro Sprachsequenz nur eine subjektive Bewertung (bzw. Bewertungsverteilung) vor, jedoch werden für jede Sprachsequenz eine Vielzahl von Merkmalsvektoren extrahiert, und es kann davon ausgegangen werden, das die Sprachqualität innerhalb einer Sprachsequenz variiert. Jedem Merkmalsvektoren wird nun vor dem Training des Models der Mittelwert des entsprechenden Sprachsamples zugeordnet. Im folgenden werden alle Sprachseqenzen erneut gewertet und für jede Sprachsequenz wird der Mittelwert der subjektiven Wertung bestimmt. Im folgenden wird nun der Offset zwischen den subjektiven Bewertungen und den Mittelwerten für die Sprachsamples gebildet. Die Bewertungen eines jeden Merkmalsvektors eines Sprachsamples werden um diesen Offset korrigiert. Mit diesen neu bewerteten Vektoren wird erneut ein Modell aufgebaut und der beschriebene Vorgang wird so lange fortgesetzt bis die Korrelationen der berechneten Mittelwerte und der subjektive ermittelten Mittelwerte konvergieren. Für 65 bewertete Sprachsamples (incl. Fahrgeräusch und Doubletalk) wurde mittels leave-one out Verfahren das oben beschriebene Verfahren überprüft. Ohne die erweiterte MMSE-Schätzung wurden Korrelationen von 0.73 mit einer Standardabweichung von 1.425 erreicht. Nimmt man das erweiterte Verfahren so erreicht man unter den selben Bedingungen eine Korrelation von 0.93 mit einer Standardabweichung von 0.4.

Literatur

[1] Gädicke, W.; Fingscheidt, T.: Reproduzierbare Vermessung von Freisprecheinrichtungen mit Hintergrundgeräuschsimulation, IMA 2006

[2] Gädicke, W.; Lieb, M.: Ein multimediales Verfahren zur subjektiven Evaluierung der Sprachqualität von Kfz-Freisprecheinrichtungen, DAGA 2006

[3] Falk, T. H.; Chan, W.; Kabal, P.: Speech Quality Estimation Using Gaussian Mixture Models, Proc. Interspeech 2004, pp. 2013-2016