Referenzfreie Schätzung der perzeptuellen Dimension Rauschhaftigkeit von übertragener Sprache

Friedemann Köster, Gabriel Mittag, Sebastian Möller
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Deutschland,
Email: friedemann.koester@telekom.de, gabriel.mittag@gmail.com, sebastian.moeller@telekom.de

Zusammenfassung

Dieser Beitrag befasst sich mit der referenzfreien Schätzung wahrgenommener Qualität übertragener Sprache: Wir gehen davon aus, dass sich die Gesamtqualität durch parametrische Schätzung von mehreren perzeptiven Dimensionen vorhersagen lässt. Wältermann et al. [1] konnten Direktheit, Kontinuität, Rauschhaftigkeit und Lautheit als Hauptdimensionen identifizieren. Modelle mit Schätzern von Qualitätsdimensionen haben im Vergleich zu integralen Verfahren, welche nur die Gesamtqualität bestimmen, die Vorteile, durch analytische Informationen auf die Ursache einer Störung schließen zu können und robuster gegenüber Veränderungen des Übertragungssystems zu sein. Bisher wurden Dimensionsschätzer entweder mit parametrischen oder intrusiven instrumentellen Modellen realisiert. In dieser Arbeit basiert die Schätzung der wahrgenommen Qualität auf nicht-intrusiven instrumentellen Schätzungen. Als erste qualitätsrelevante perzeptive Dimension wurde Rauschhaftigkeit ausgewählt. Es konnten erste vielversprechende Ergebnisse auf unterschiedlichen Datensätzen erzielt werden, die zeigen, dass eine robuste Schätzung qualitätsrelevanter perzeptiver Dimensionen ohne Referenzsignal möglich ist.

Einleitung und Stand der Forschung

Für die Anbieter von Telekommunikationssystemen ist die Qualität der übertragenen Sprache der wichtigste Indikator um ihre Systeme zu evaluieren. Deswegen ist die Bestimmung der Sprachqualität essentiell und um eine Verbesserung zu erreichen, ist es nicht nur wichtig diese zu kennen, sondern auch analysieren zu können. So ist es bspw. ein großer Unterschied, ob ein Sprachsignal als schlecht bewertet wird, weil es oft unterbrochen ist oder weil es starkem Rauschen ausgesetzt ist. Aus diesem Grund wird versucht die integrale Sprachübertragungsqualität, die als ein multidimensionaler "Event" beschrieben wird [2], auf der Basis von mehreren perzeptiven Dimensionen zu modellieren. Wältermann et al. [1][3] haben vier perzeptive Dimensionen für Schmal- und Breitbandige Sprachübertragungen identifizieren können; Direktheit/Frequenzgehalt, Kontinuität, Rauschhaftigkeit und Lautheit. Außerdem entwickelte Wältermann eine Methode, mit der die einzelnen perzeptiven Dimensionen direkt subjektiv bewertet werden können [1]. Diese subjektive Methode liefert zwar gültige Ergebnisse, ist aber kosten- und zeitintensiv weshalb ein objektives instrumentelles Modell zur Schätzung der perzeptiven Dimensionen gebraucht wird. Ein instrumentelles Verfahren, welches jeweils für die ersten drei Dimensionen einen Schätzwert ermittelt und aus diesen, mit Hilfe eines linearen Qualitätsmodells, die integrale Qualität bestimmt, wurde von Scholz [4] präsentiert. Das Verfahren wurde später durch das DIAL Modell [5] von Côte verbessert und mit einem Schätzer für die integrale Qualität erweitert. Beide Verfahren basieren allerdings auf einem signalvergleichenden (intrusiven) Modell, für welches das unbeeinträchtigte Eingangssignal gebraucht wird. Solche intrusiven Modelle sind meist nur im Labor hilfreich, da in der Praxis das Eingangssignal, z.B. im laufenden Betrieb (zum Monitoring), nicht verfügbar ist. Aus diesem Grund werden referenz-freie (nicht-intrusive) Modelle benötigt, welche die Bewertungen nur anhand der Probe schätzen. Für die identifizierten Qualitätsdimensionen gibt es so ein nicht-intrusives Modell bis jetzt nicht. Ziel dieses Beitrags ist es, zu zeigen, dass es mit nicht-intrusiven Modellen möglich ist bei der Schätzungen der perzeptiven Qualitätsdimensionen robuste Ergebnisse zu erzielen. Als erste qualitätsrelevante perzeptive Dimension wurde Rauschhaftigkeit ausgewählt. Insgesamt werden 18 Parameter zur Beschreibung der Rauschhaftigkeit gesammelt von denen dann 7 endgültig verwendet werden.

Datenbanken

Zur Evaluation der Parameter stehen drei Datenbanken zur Verfügung. Sie beinhalten Sprachproben mit verschiedenen Sätzen und Sprechern sowie Ergebnisse von subjektiven Experimenten, wobei nur die in einen MOS Wert transformierten Bewertung der Rauschhaftigkeit verwendet werden. Die ersten beiden Datenbanken $DAT1$ und $DAT2$ wurden von Wältermann [1] erstellt und bestehen aus 66 bzw. 76 Konditionen, die dritte Datenbank, die $SWISS$ Datenbank [5], besteht aus 54 Konditionen. Alle Datenbanken (196 Konditionen) beinhalten Beeinträchtigungen in Form von verschieden NB und WB Kodierern, Rauschen (Hintergrund, MNRU), Paket Verlusten und Bandpassfiltern.

Vorverarbeitung

Für die Vorverarbeitung der Signale wird zunächst eine sog. Voice Activity Detection (VAD) verwendet. Diese bestimmt anhand der Kurzzeitleistung und der Nulldurchgänge (Zero-Crossing), ob ein Signalabschnitt aktiv oder still ist (Bestimmung von Hintergrund-Rauschen). Eine VAD ist besonders für die Bestimmung von nichtsignalkorrelierten Rauschen wichtig, da in Sprachpausen Störungen wie z.B. Hintergrundgeräusche sehr gut detektiert werden können.

Parameterauswahl

Mit den 18 gesammelten Parametern lässt sich die Rauschhaftigkeit nicht isoliert abbilden. Aus diesem Grund muss eine Kombination von diesen 18 Parametern gefunden werden, die über die drei Datenbanken robust die Rauschhaftigkeit modellieren kann. Hierfür wird eine wiederholte sequentielle Parameterauswahl in Verbindung mit einer Kreuzvalidierung verwendet. Dabei werden für einer Trainingsmenge zu einem linearen Regressionsmodell solange Parameter hinzugefügt, bis das Modell die Rauschhaftigkeit bestmöglich modelliert und anschließend auf einer Testmenge die Korrelation überprüft. Mit diesem Verfahren wurden 7 der 18 Parameter ausgewählt die im Folgenden genauer beschrieben werden sollen.

Dimensionsparameter

Zusammenfassend aus [6] sind die ausgewählten Parameter wie folgt definiert:
NL (Noise Level):Mittlerer Rauschpegel - gibt die Intensität des Hintergrundrauschens in sprachlosen Segmenten über die mittlere Energie an, abhängig von der VAD.
PKSL (PeaKS Lowpassfilter):Anzahl der Maxima in Zeitbereich bei einem tiefpassgefiltertem (300 Hz) Singnal - Indikator für das Hintergrundrauschen, robuster gegenüber der VAD.
PKSB (PeaKS Bandpassfilter):Anzahl der Peaks im Zeitbereich des zwischen bandpassgefilterten (3000 Hz - 4300 Hz) Signals - ähnlich zu PKSL, zeigt, dass auch höhere Frequenzen eine Einfluss haben.
PV (Power Variation):Standardabweichung der Kurzzeitleistung - Es wird davon ausgegangen, dass die Variation der Kurzzeitleistung bei nicht stationärem Rauschen höher ist als bei stationären.
FVSN (Frequency Variation Spline):Variation des Frequenzganges - RMSE zwischen dem Leistungsdichtespektrums (LDS) und der Spline-Glättung des LDS.
HFG (Frequency Gravity):Einfluss der Tonhöhe des Rauschens - Dieser Parameter berechnet den Schwerpunkt des Rauschleistungsdichtespektrums.
ROSC (Roughness):Rauhaftigkeit der Einhüllenden - Dieser Parameter bestimmt über die Einhüllende des Sprachsignals die Intensität von signalkorrelierten Rauschen an.

Ergebnisse

Mit den 7 ausgewählten Parametern wird für jede Datenbank erneut ein lineares Regressionsmodell erstellt. Hierbei ist zu beachten, dass die gleiche Methode wie zur Auswahl der Parameter verwendet wird, mit dem Unterschied, dass jetzt nur die 7 vorher ausgewählten Parameter verwendet werden müssen. Insgesamt wurde dies 5 mal zur Verifizierung durchgeführt. Die gemittelten Ergebnisse sind in Tabelle 1 zu sehen.

Auswertung und Ausblick

Insgesamt zeigen die Ergebnisse, dass eine referenzfreie Schätzung der Rauschhaftigkeit gelingt. Für alle drei Datenbanken wurden Korrelationen von 0,88 und mehr auf den Testdaten erreicht. Es stellt sich heraus, dass die Bestimmung von Hintergrundrauschen bei einer geeigneten VAD gut gelingt und das eine Reihe von Parametern hierfür gefunden werden konnten (6 von 7, **NL**,**PKSL**,**PKSB**,**PV**,**FVSN**,**HFG**). Bei sprachinaktiven Segmenten werden mit Hilfe des Frequenzganges und der Maxima in dem Sprachsignal gute Ergebnisse erzielt. Für die Bestimmung des sprachkorrelierten Rauschens ist nur ein Parameter gefunden worden (**ROSC**) der über die Einhüllende in sprachaktiven Segmenten zufriedenstellende Ergebnisse zeigt. Die Ergebnisse der Korrelation und des Prädikationsfehler zu dem auditiven gemessenen Werte sind aber noch nicht hoch genug für eine zuverlässige Schätzung. Dennoch sind die Ergebnisse unter Berücksichtigung, dass es sich um einen neu untersuchten Ansatz handelt und der schwere einer referenzfreien Vorhersage, fürs Erste zufriedenstellend. So ist die Möglichkeit einer zuverlässigen Vorhersage der Dimension Rauschhaftigkeit durch ein referenzfreies Verfahren zukünftig nicht auszuschließen. In Zukunft sollten für die Verbesserung der Robustheit die vorgeschlagenen Parameter auf weiteren Daten verifiziert werden und auch eine Kreuzvalidierung unter den Datenbanken durchgeführt werden. Für einen endgültigen Schätzer sollten Klassifizierungen der Rauscharten die nicht-intrusive Schätzung vereinfachen und es wäre auch wünschenswert weitere Parameter, die das signalkorrelierte Rauschen beschreiben, zu finden.

Tabelle 1: Ergebisse der gemittelten Korrelationen für die drei Datenbanken

Ergebnisse	$\emptyset DAT1$	$\emptyset DAT2$	$\emptyset SWISS$
$\bar{\rho}$ Training	0,93	0,94	0,93
$\bar{\rho}$ Test	0,90	0,88	0,88
$\overline{RMSE}$	0,43	0,42	0,45

Literatur

[1] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*, Springer, Berlin, 2012.

[2] S Möller and A. Perkis, *Qualinet White Paper on Definitions of Quality of Experience*, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Novi Sad, 1.2 edition, March 2013.

[3] M. Wältermann, A. Raake, and S. Möller, "Quality dimensions of narrowband and wideband speech transmission," Acta Acustica united with Acustica, 2010, pp. 1090–1103.

[4] K. Scholz, *Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen*, Shaker Verlag, Kiel, 2008.

[5] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Springer, Berlin, 2011.

[6] G. Mittag, "Referenzfreie, instrumentelle bestimmung von bestimmten qualitätsrelevanten dimensionen übertragener sprache," Bachelor-Thesis, unpublished, November 2013.