WAHRGENOMMENE SPRACHQUALITÄT IN TELEFONGESPRÄCHEN BEI ZEITLICH VARIIERENDEN ÜBERTRAGUNGSEIGENSCHAFTEN

Benjamin Weiss[1], Sebastian Möller[1] und Jens Berger[2]

[1]Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin
[2]SwissQual AG, Solothurn, Schweiz
sebastian.moeller@telekom.de

Abstract: Auf Basis auditiver und instrumenteller Bewertung von 5–6 Sekunden langen Sprachsamples wurde ein neuartiges Verfahren zur Modellierung und Vorhersage wahrgenommener Qualität von Telefongesprächen evaluiert und bestätigt. Zugeschriebene Qualität für 1- und 2-Minuten Gespräche kann mit diesem Verfahren deutlich besser als durch das arithmetische Mittel der Einzelbewertungen erfasst werden.

1 Einleitung

Instrumentelle Messverfahren zur Schätzung der Sprachqualität, die heutzutage zum Beispiel zum Monitoring eingesetzt werden, verwenden üblicherweise Sprachsamples von ca. 4–8 Sekunden Dauer. Auf Basis dieser Samples werden Sprachqualitätswerte geschätzt, die ein Benutzerurteil für eine solche Dauer mit recht hoher Validität und Reliabilität widerspiegeln, vgl. z.B. ITU-T Rec. P.862 (2001). Allerdings sind diese Dauern nicht typisch für Telefonverbindungen.

Längere Stimuli mit variierender Qualität (z. B. 30–120 Sekunden) können mit solchen Verfahren nicht zufriedenstellend erfasst werden. Benutzerbewertungen fallen bei solchen Dauern geringer aus als mittlere Schätzungen instrumenteller Verfahren [4]. Abschnitte, die zeitlich nahe der Gesamtbewertung liegen, gehen stärker in diese ein, und Qualitätsminderungen werden schneller wahrgenommen als Steigerungen [5].

Im Rahmen der ETSI-Standardisierungsgruppe STQ-Mobile soll ein Verfahren zur Bestimmung der Sprachqualität eines simulierten Gespräches entwickelt werden. Ziel ist eine nutzbare Modellierung, die nicht nur längere Stimuli erfasst, sondern auch tatsächliche Telefongesprächsituationen mit ihren interaktiven Anteilen. Dieses Verfahren bestimmt aus kurzzeitigen Messungen mit 5–6 Sekunden langen Proben einen Schätzwert für die Qualität am Ende eines Gespräches von der Dauer von etwa ein bis zwei Minuten. Die Methode zur Bestimmung eines solchen Verfahrens ist bereits erfolgreich für 12 Sekunden lange Samples in einem informellen Experiment mit Gesprächen von 2 Minuten Dauer getestet worden, vgl. draft ETSI TR 102 506 [1].

Gemäß dieses Verfahrens konnten Telefonsituationen zufriedenstellend simuliert werden. Kurze Samples wurden für diese simulierten „Gespräche“ in verschiedenen Profilen zusammengestellt, um Systematiken in der Wirkung zeitlich variierender Sprachqualität auf den Gesamteindruck eines Gesprächspartners erkennen zu können. Erste Ergebnisse bestätigen, dass sich die Dialogbewertungen von Probanden nicht ausreichend über das arithmetische Mittel der Einzelbewertungen oder des instrumentellen Maßes PESQ modellieren lassen. Insbesondere traten zwei Effekte auf, die bereits in psychologischer Literatur zum Erinnerungsvermögen beschrieben sind: Spätere Eindrücke in einer linearen Abfolge bleiben dominanter und besser im Kurzzeitgedächtnis und wirken sich damit stärker auf Beurteilungen aus („recency-effect“) [3]. Zusätzlich wird ein herausragendes Ereignis „Ausreißer“ nicht vergessen [2].

In zwei formellen Experimenten in Anlehnung an die Empfehlungen der ITU-T Rec. P.800 (1996) wurde diese Methode nun angewendet.

2 Material

Aus hochqualitativen Aufnahmen von zwei weiblichen und zwei männlichen Sprechern (vgl. [1]) wurden jeweils 10 Samples von 5–6 Sekunden Länge erstellt, die inhaltlich zueinander in Bezug stehen. Daraus wurden „Gesprächspartner“ für 1-Minuten bzw. 2-Minuten lange Dialoge simuliert. Anschließend wurde die Abtast-Rate auf 8 kHz verringert, und die Äußerungen wurden IRS-gefiltert und prozessiert, um verschiedene Qualitätsstufen jedes Samples zu erhalten. Aus diesen wurden Folgen von Stimuli für die beiden Experimente zusammengestellt, die in ihrer zeitlich variierenden Qualität jeweils 10 verschiedenen Profilen entsprechen: Drei mit gleichbleibenden, zwei mit kontinuierlich verändernden Qualitäten; sowie fünf Profile hoher Qualität mit einzelnen „Ausreißern“ an verschiedenen Positionen, drei davon als „Burstfolgen“. Dies bezeichnet Samples, die sich nicht durch gleichmäßig schlechte Qualität auszeichnen, sondern Aussetzer und abrupten Störungen aufweisen, wie sie gerade bei Mobilverbindungen auftreten können. Abbildung 1 zeigt schematisch einen solchen Dialog:

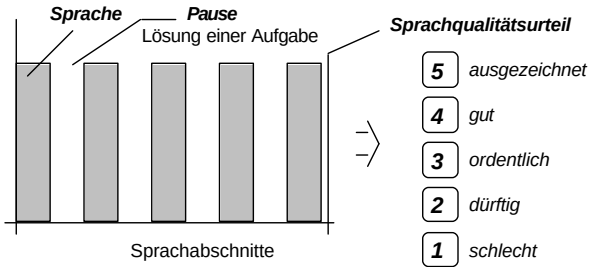

Abbildung 1: Schematische Darstellung eines Dialoges und seiner Bewertung

Jeweils fünf inhaltlich zusammenhängende Stimuli wurden den Probanden in einem Dialog präsentiert. Bei dem Test mit 1-Minuten langen Dialogen wurde die Hälfte dieser 10 Äußerungen für einen Dialog verwendet (Sample 1–5 oder 6–10). Um alle 10 Störprofile für jeden Sprecher zu erstellen (insgesamt 40 Dialoge), wurden die Profile variierend aus beiden Hälften der Äußerungen erstellt, sodass genau 5 Störprofile für die erste Hälfte, 5 für die zweite Hälfte der Äußerungen vorliegen. Bei dem zweiten Test bestand ein Stimulus aus zwei aufeinanderfolgenden Samples, sodass alle 10 Äußerungen eines Sprechers für einen Dialog verwendet wurden. In diesem Fall wurden jeweils 10 Störprofile pro Geschlecht erstellt, gleichmäßig für beide Personen. Insgesamt wurde also jeder Dialog in beiden Tests inhaltlich fünf mal wiederholt.

Nach jedem Stimulus erfolgte eine etwa gleich lange Pause von etwa der Länge eines Stimulus, also 6 bzw. 12 Sekunden.

In den Pausen zwischen den Sprachsamples beantworteten die Probanden mündlich eine Frage zum gerade gehörten Stimulus, um die Aufmerksamkeit von der Sprachqualität auf den Inhalt zu lenken und sich einer Gesprächssituation anzunähern. Damit handelt es sich genaugenommen nicht um einen Dialog, sehr wohl aber um eine Gesprächssituation, sodass sich die Bewertungen der Probanden auf Qualitätsurteile echter Telefongespräche übertragen lassen.

Nach dem letzten Stimulus wurde keine Frage mit fester Pausenzeit präsentiert, sondern eine 5-stufige „Mean Opinion Score“ Skala (MOS) [6], auf der die Versuchspersonen mit einer Maus ein abschließendes Urteil zur Gesamtqualität des Gespräches abgaben. Mit dieser manuellen Handlung sollte auch die Gesprächssituation demonstrativ beendet werden. Die Gesamtdauer eines solchen Dialoges betrug dabei etwa 60 Sekunden (1-Min. Test) bzw. 120 Sekunden (2-Min. Test).

Fragen zum Inhalt der Stimuli wurden in ihrer Formulierung so kurz gehalten, dass sie in 6 Sekunden zu erfassen und zu beantworten sind. Jeweils drei optionale Antworten, von denen

eine inhaltlich richtig ist, wurden erstellt. Da die Stimuli zwei bis drei Informationen beinhalten, nach denen gefragt werden kann, jeder Dialog aber fünf mal auftrat, wurden als Varianten die falschen Antworten und die Position der richtigen variiert, damit sich keine Zusammenstellung wiederholte. Pro Dialog wurden mindestens 2 verschiedene Fragen erstellt. Als obligatorische optionale Antwort wurde zudem „habe ich nicht verstanden / mir nicht gemerkt“ präsentiert, auf die auch in der Instruktion hingewiesen wurde.

In einem zusätzlichen Experiment wurden die verwendeten Samples (nicht Stimuli) von den Versuchspersonen einzeln auf der MOS-Skala nach dem in [6] beschriebenen Verfahren bewertet.

3 Durchführung

Die Experimente wurden in einem akustisch gedämpften Raum durchgeführt. Nach der Begrüßung und Instruktion wurde zunächst ein Hörtest nach DIN EN 600645-2 durchgeführt. Keiner der Probanden zeigte ein eingeschränktes Hörvermögen.

Nach einem Trainingsdialog wurden pro Versuchsperson 40 (1-Min. Test) bzw. 20 (2-Min. Test) dieser „Dialoge“ automatisch über einen Handapparat präsentiert. Die Fragen zum Inhalt des Gehörten einschließlich der vorgeschlagenen Antworten erschienen als Text auf dem Bildschirm. Die Antworten der Versuchspersonen wurden über den Handapparat aufgezeichnet. Nach der Hälfte der Dialoge wurde jeweils eine etwa 5 bis 10-minütige Pause eingelegt. Die Abfolge der Dialoge wurde in fünf Varianten pseudo-randomisiert, sodass keine zwei Dialoge eines Sprechers direkt aufeinander folgten. Eine Durchführung inklusive Audiometer-Test dauerte zwischen 55 und 65 Minuten.

Nach einer weiteren Pause von 15 Minuten bewerteten die Versuchspersonen die verwendeten Samples direkt auf der MOS-Skala. Auch hier wurde nach der Hälfte der Stimuli eine Pause von 5–10 Minuten eingelegt. Eine Durchführung dauerte etwa 35–45 Minuten.

Die Versuchspersonen wurden für ihre Teilnahme an den Experimenten bezahlt. Von den zwei mal 24 Probanden waren 18 Frauen, 30 Männer; jeweils gleichmäßig in beiden Tests vertreten. Das Alter variierte von 17 bis 48 Jahre. Es wurden so jeweils 24 Beurteilungen für jeden Dialog und 45 bzw. 48 Einzelbewertungen für jedes verwendete Sample erhoben.

4 Diskussion der Ergebnisse

Beide Experimente zeigen vergleichbare Ergebnisse, obwohl die Dialoge in ihrer Dauer um das doppelte von einander abweichen. Besonders bei initialen niedrigen Qualitäten wäre ein Unterschied durch die unterschiedliche Dauer zur Bewertung möglich gewesen. Insgesamt sind die Einzelbewertungen der Probanden konsistent, auch die Dialogbewertungen zeigen keine großen Streuungen.

Die Dialogbewertungen sind mit zwei Ausnahmen gleich oder niedriger als der Mittelwert der Einzelurteile. Selbst das ansteigende Störmuster 4 wurde nicht durchweg besser bewertet (vgl. Abbildung 2). Bei konstanten und ansteigenden Mustern sind die Dialogbewertungen und mittleren Einzelurteile miteinander vergleichbar, ansonsten liegen die Dialogbewertungen deutlich niedriger. Dieser Unterschied zwischen den beiden Profilen mit kontinuierlichen Veränderungen zeigen eine Asymmetrie in der Bedeutung höherer und niedrigerer Qualitäten (vgl. Abbildung 2).

Anhand der „Burstfolgen“ lassen sich die Auswirkungen des „recency-effects“ und der „Ausreißer“ gut illustrieren (Abbildung 3).

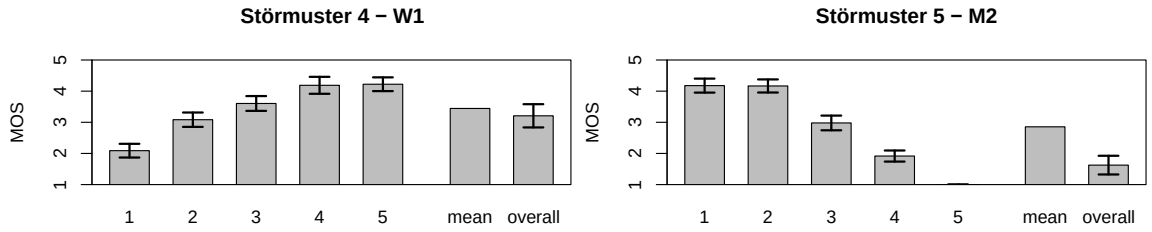

Abbildung 2: Ergebnisse des Dialoges mit Störmuster 4 und 5; mittlere MOS-Werte der 5 Samples; arithm. Mittel dieser 5 (mean), mittlere Dialogbewertung (overall)

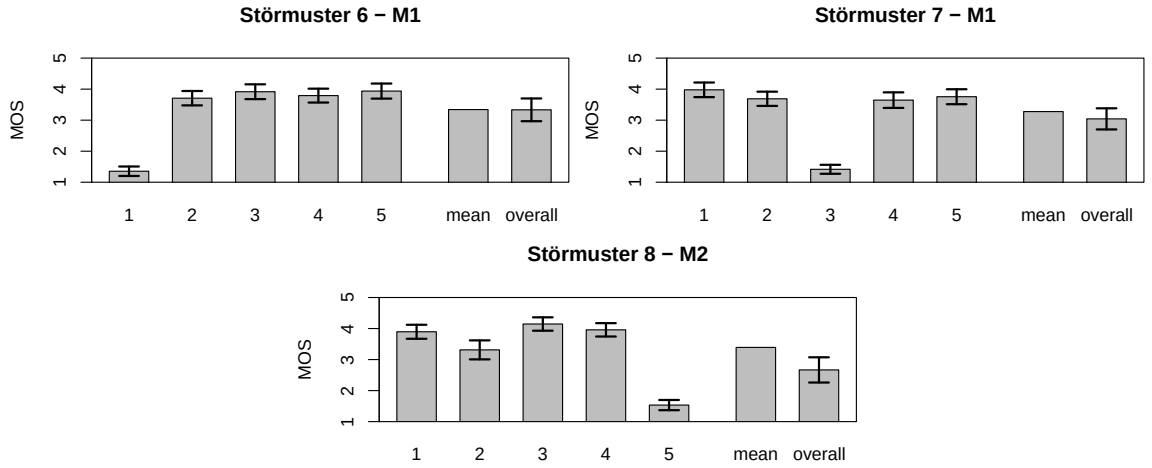

Abbildung 3: Vergleich dreier Profile mit Positionseffekt; mittlere MOS-Werte der 5 Samples; arithm. Mittel dieser 5 (mean), mittlere Dialogbewertung (overall)

Die Differenzen von gemittelten Einzelurteilen und Dialogbewertungen pro Störprofil zeigen eine deutliche Systematik. Sie sind getrennt für beide Tests in Abbildung 4 dargestellt. 4 weniger gut realisierte Störmuster wurden in dieser Darstellung ausgeschlossen. Es wird deutlich, dass das Muster 9 (mit einem mäßigen medialen Ausreißer) in seiner Gesamtbewertung nicht vom Mittelwert der Einzelurteile abweicht. Dagegen zeigen Muster 4 (kontinuierlich steigend), 6 (Bursts initial) und 7 (Bursts medial) deutliche Unterschiede, so dass nach dieser Sichtung der Ergebnisse der „recency-effect“ gegenüber dem Einfluss der Stärke einzelner Ausreißer weit geringer ausfällt.

Je niedriger die Qualität der Ausreißer bei ansonsten konstanten Qualitäten, desto stärker sowohl der absolute Einbruch der Dialogbewertung, als auch der relative gegenüber den mittleren Einzelurteilen. Je später der Qualitätseinbruch, desto stärker seine Wirkung auf das Gesamturteil. Im Vergleich von Mustern mit Ausreißern gegenüber kontinuierlich verändernden Qualitäten mit Qualitätsminima an gleichen Positionen (4 zu 6 bzw. 5 zu 8 und 10) zeigen sich keine deutlichen Unterschiede zwischen diesen Störmustern. Dieses Verhalten trat schon in [1] auf und führte dazu, auch in der Modellierung kontinuierlicher Störmuster nur den Mittelwert, „recency-effect“ und den Einzelstimulus mit minimaler Qualität zu berücksichtigen. Insofern haben sich durch die vorliegenden Ergebnisse beide Mechanismen bestätigt. Abweichungen von „overall“ gegenüber „mean“ erscheinen hier von der Qualität der jeweils schlechten Einzelstimuli abhängig. Da nur eine begrenzte Anzahl Störprofile getestet wurden, lassen sich keine Aussagen für positive Ausreißer treffen, wobei nach den hier dargestellten Ergebnissen und denen in [5] mit einem geringeren Effekt zu rechnen ist.

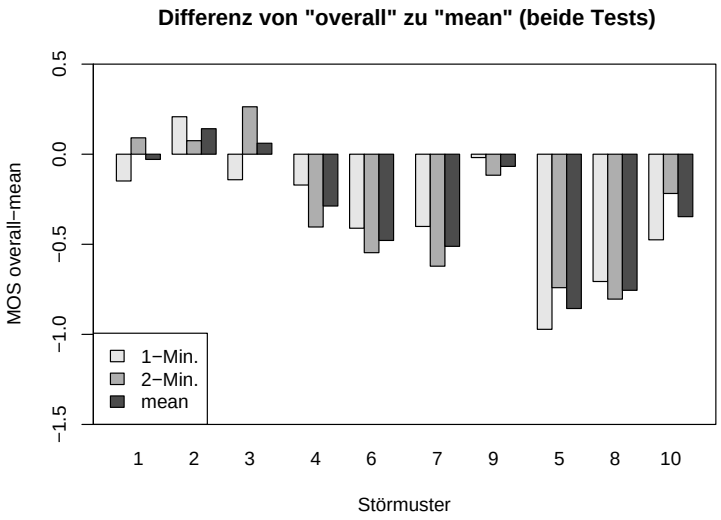

Abbildung 4: Differenz gemittelte Einzelurteile und Dialogbewertungen: Mittelwerte von Dialogen für den 1- und 2-Minuten Test sowie gesamt; für Störmuster konstanter Qualität (1–3) sowie niedriger Qualität initial (4,6), medial (7,9) und final (5,8,10)

5 Modellierung der Dialogbewertungen aus Einzelbewertungen

Eine Vorhersage der Dialogbewertungen aus reinen Mittelwerten ist nicht zufriedenstellend. Wie in Abbildung 5 deutlich zu sehen ist, ist eine solche Modellierung gerade für Gespräche mittlerer und niedrigerer Qualität zu optimistisch. Der Korrelationskoeffizient beträgt $R = 0.85$, bzw. $R = 0.83$.

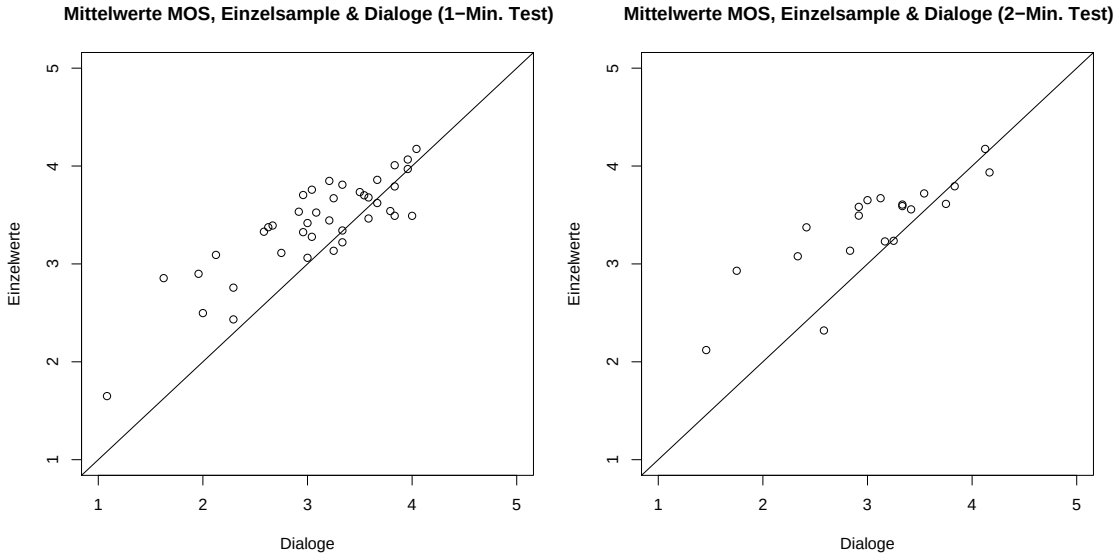

Abbildung 5: Einzelvergleiche der MOS-Werte „mean“ und „overall“

Aus den gerade dargestellten Ergebnissen beider auditiver Tests ergeben sich zwei Effekte, die Position (recency-effect) und Stärke schlechter Einzelstimuli (negativer „peak“) betreffen. Beide sollen nun so berücksichtigt werden, dass sich die Bewertung der Dialoge aus den Einzelbewertungen der 6 Sek. Stimuli zur Vorhersage modellieren lässt. Im Folgenden wird das Vorhersagemodell aus [1] kurz vorgestellt, da es bei einer Vergleichbarkeit der Resultate übernommen werden kann.

5.1 Modellierung der Ergebnisse nach Draft ETSI TR 102 506

Beide Effekte werden in separaten Schritten modelliert:

1. Modellierung des „recency-effect“

Die MOS-Werte der Einzelstimuli für die Mittelwertbildung werden je nach Position unterschiedlich gewichtet. Die Bedeutung der Sample für das Gesamturteil steigt mit zunehmender Dauer an.

$$\overline{MOS}_{\text{A\_mod1}} = \sum_{t=1}^{5}(a_t MOS_t) / \sum_{t=1}^{5} a_t$$

t: Sprachabschnitt; a: Gewichtungskoeffizient

Um das Modell auf die Daten des 2-Min. Tests anwenden zu können, wird jeweils für die beiden Samples, die zusammen einen Stimulus ergeben, der Mittelwert der Einzelurteile verwendet.

Bei Verwendung der vorgeschlagenen Gewichtungen von t_1 bis t_5 (0.4, 0.5, 0.6, 0.8, 1.0) beträgt die Korrelation $R = 0.91$ für den 1-Min. Test sowie $R = 0.85$ für den 2-Min. Test.

2. Modellierung des Einflusses besonders stark abweichender Sample

Der Einfluss des schlechtesten Samples eines Dialogs wird direkt mit dem modifizierten Mittelwert verrechnet:

$$\overline{MOS}_{\text{A\_mod2}} = \overline{MOS}_{\text{mod1a}} - 0.4(\overline{MOS} - min(MOS_t))$$

t: Sprachabschnitt; a: Gewichtungskoeffizient

Mit dieser Modifizierung wird eine Korrelation von $R = 0.93$ (1-Min. Test) und $R = 0.93$ (2-Min. Test) erreicht.

5.2 Anpassung der Modellparameter an die Ergebnisse des 1- und 2-Minuten-Tests

Das verwendete Modell ist bewusst einfach gehalten, um Überanpassung an die Daten zu vermeiden. Es beinhaltet nur zwei Mechanismen, die Systematiken der Ergebnisse umsetzen. Weitere Mechanismen zur Optimierung der Korrelation lassen sich nicht ohne Ergebnisse mit zusätzlichen Störmustern rechtfertigen.

Bei der Anpassung des Modells an die Ergebnisse der beiden hier vorgestellten Tests werden die Gewichtungsfaktoren nicht relativ zur Dialogdauer angewendet, sondern für ein konkretes Zeitfenster, da die Datenanalyse ergeben hat, dass sich der „recency-effect“ für beide Testlängen unterschiedlich auswirkt: Beim 1-Min. Test sind mindestens die letzten beiden Samples betroffen, beim 2-Min. Test nur der letzte. Der „recency-effect“ scheint sich im Bereich von etwa 20–30 Sekunden bis zur Bewertung abzuspielen, während die Gewichtung davor konstant gehalten werden kann. Wird das bestehende Modell entsprechend angepasst, ergibt sich eine generalisierte Form für variierende Gesprächsdauern und Sampleanzahl, bestehend aus 6 Sekunden langen Samples (Modell B).

Die Gewichtungsfaktoren für a betragen generell 0.5. Im Bereich der letzten 18 Sekunden vor der Bewertung, also der 3 letzten Abschnitte von 6 Sekunden bis Ende des letzten Samples, steigt die Gewichtung an (0.6,0.7,1.0). Dabei werden für den 1-Min. Test die Werte 1 und 3 verwendet (im Zeitbereich des 2. Wertes ist eine Interaktionspause); für den 2-Min. Test werden die Werte 2 und 3 verwendet (im Zeitbereich des 1. Wertes ist eine Interaktionspause). Diese Anpassung der Gewichtungsfaktoren entspricht eher dem reinen „End Affect“ [2] als dem „recency-effect“.

Unabhängig von der Gesprächsdauer oder der Sampleanzahl wird das gewichtete Minimum im zweiten Modellschritt subtrahiert. Bei Verwendung der alternativen Gewichtung im ersten Schritt und einem Vorfaktor von 0.3 anstatt von 0.4 im zweiten Schritt verbessert sich die lineare Korrelation nur leicht ($R = 0.95$ für 1-Min. Test; $R = 0.93$ für 2-Min. Test), ist aber damit flexibel für verschiedene Dialogdauern und Anzahlen von Interaktion.

Die Leistungsfähigkeit (R-Wert, Standardfehler) der verschiedenen Modelle ist in Tabelle 1 dargestellt. Man beachte, dass für den 2-Min. Test nur halb so viele Werte zur Verfügung stehen, so dass die Modellierung des 1-Min. Test stärker in die gemeinsame Korrelation eingeht.

Daten	reine Mittelwerte		Modell A_mod1		Modell A		Modell B	
	R	Ep	R	Ep	R	Ep	R	Ep
1-Min	0.85	0.26	0.91	0.23	0.93	0.22	0.95	0.19
2-Min	0.83	0.28	0.85	0.31	0.93	0.24	0.93	0.21
beide	0.84	0.26	0.89	0.25	0.93	0.22	0.94	0.19

Tabelle 1: R-Wert und Standardfehler der verschiedenen Modelle

Die modifizierte Version des Modells aus [1] (Modell B) ist im Vergleich zu den Dialogbewertungen in 6 abgebildet.

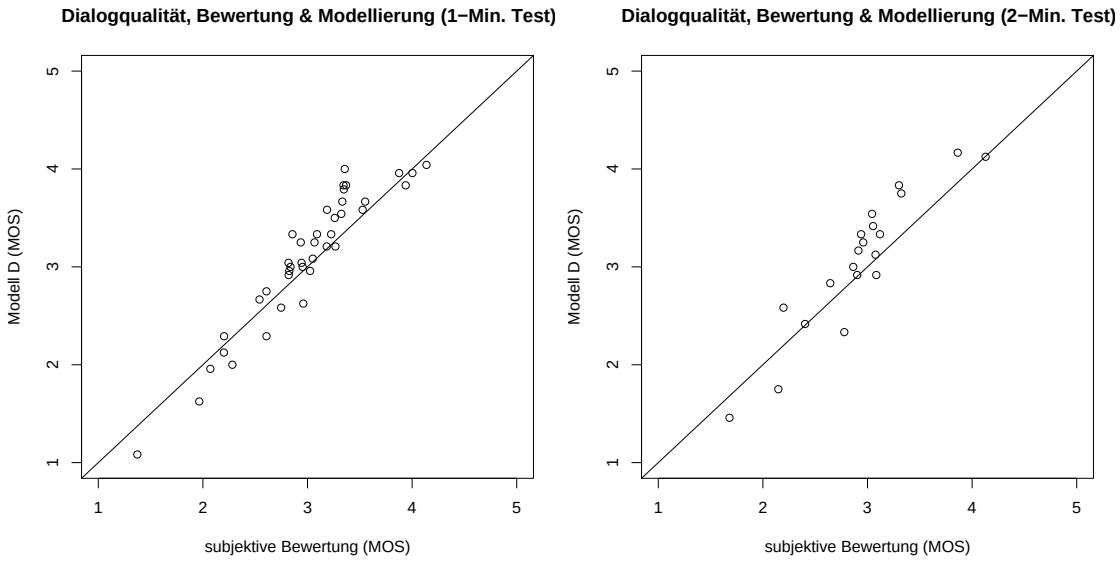

Abbildung 6: Einzelvergleiche von Modell B und Dialogbewertung

Zusammenfassend muss gesagt werden, dass sich getrennte Modelle für die beiden separaten Experimente nicht rechtfertigen lassen. Das in [1] vorgeschlagene Modell kann direkt übernommen werden. Parameteränderungen führen nur zu minimalen Verbesserungen der Vorhersage.

6 Modellierung der Dialogurteile aus instrumentellen Samplebewertungen

Subjektive Beurteilungen der Einzelsample und PESQ-Werte weisen eine lineare Korrelation von $R = 0.93$ auf, was im Bereich anderer Vergleiche von subjektiven Urteilen mit PESQ-Werten liegt. Werden nun die beiden Modelle mit diesen Werten anstatt der subjektiven MOS-Werte verwendet, ergeben sich folgende Ergebnisse:

Daten	reine Mittelwerte		Modell A_mod1		Modell A		Modell B	
	R	Ep	R	Ep	R	Ep	R	Ep
1-Min	0.75	0.31	0.82	0.31	0.89	0.27	0.89	0.25
2-Min	0.70	0.34	0.73	0.37	0.83	0.33	0.84	0.29
beide	0.73	0.32	0.79	0.32	0.87	0.29	0.87	0.26

Tabelle 2: R-Wert und Standardfehler der verschiedenen Modelle

Mit beiden Verfahren können die Dialogbewertungen auf Basis von PESQ-Werten nicht so gut vorhersagen wie mittels MOS-Werten. Hier addiert sich der Fehler der Modellierungen mit der Vorhersage der Samplebewertungen durch PESQ.

7 Abschließende Bemerkungen

Die in [1] entwickelte Methode konnte erfolgreich in den hier vorgestellten auditiven Tests verwendet werden. Die Systematiken wurden bestätigt, wie auch die dort entwickelte Modellierung von Dialogurteilen aus Bewertungen von Einzelsamples. Dieses Modell führt sowohl mit subjektiven MOS-Werten (10 Prozentpunkte) als auch mit Bewertungen nach PESQ (14 Prozentpunkte) zu deutlich besseren Resultaten als eine vergleichbare Modellierung mit jeweiligen Mittelwerten der Einzelsample. Mit einem gemeinsamen Verfahren für sowohl 1- wie auch 2-Minuten lange „Gespräche“ ist dieses robust für typische Schwankungen in der Gesprächsdauer.

Wird das modifizierte Modell auf die Daten in [1] angewendet, ergeben sich fast vergleichbare Ergebnisse: Statt eines Korrelationskoefizienten von $R = 0.85$ (Modell A) für eine Modellierung mit MOS-Werten erreicht Modell B einen Wert von $R = 0.83$. Werden die PESQ.1-Werte verwendet, betragen die Koeffizienten $R = 0.79$ (Modell B) anstatt $R = 0.84$ für Modell A. Dennoch benötigt es noch weiterer unabhängiger Daten, um die Validität des Modells abzusichern, z.B., was die zeitliche Domäne seiner Gültigkeit betrifft.

Danksagung

Den Mitgliedern der ETSI-Standardisierungsgruppe STQ-Mobile, insbesondere Joachim Riedel (T-Mobile) und Jürgen Krämer (E-Plus) möchten wir für ihre Unterstützung danken. Vielen Dank an Raphael Ullmann (SwissQual) für die tatkräftige Hilfe bei der Erstellung der Stimuli.

Literatur

[1] ETSI TR 102 506: Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call, 2006, draft.

[2] Kahneman, D.: *Objective Happiness*. In: D. Kahneman, E. Diener, N. Schwarz (eds.), Well-Being: The Foundations of Hedonic Psychology, 3–25. Russel Sage, 1999.

[3] Murdock, B. B.: The serial position effect of free recall. *Journal of Verbal Learning and Verbal Behaviour*, 64, 482–488, 1962.

[4] Gray, P., Massara, R. and Hollier, M.: An Experimental Investigation of the Accumulation of Perceived Error in Time-Varying Speech Distortions. *Audio Engineering Society 103 Convention*, New York, 1997.

[5] Gros, L. and Chateau, N.: Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships. *Acta Acustica united with Acustica*, 87, 367–377, 2001.

[6] ITU-T Recommendation P.800: *Methods for Subjective Determination of Transmission Quality.* International Telecommunication Union, Geneva, 1996.