Einfluss der Position und Stimmhaftigkeit von verdeckten Paketverlusten auf die Sprachqualität

Gabriel Mittag[1], Louis Liedtke[1], Neslihan Iskender[1],
Babak Naderi[1], Tobias Hübschen[2], Gerhard Schmidt[2], Sebastian Möller[1,3]
[1] *Quality and Usability Lab, TU Berlin*
[2] *DSS, Christian-Albrechts-Universität zu Kiel*
[3] *Language Technology, DFKI Berlin*

Abstract

Durch die Einführung von Paketvermittlung in Sprachkommunikationsnetzwerken hat sich die wahrgenommene Sprachqualität dieser Netzwerke in den letzten Jahren erheblich verbessert. Aufgrund einer durchgängigen Übertragung vom Sender bis zum Empfänger kommt es innerhalb des Übertragungskanals selbst kaum zu Beeinträchtigungen des Audiosignals. Wenn jedoch Übertragungsfehler oder Verzögerungen im Kanal auftreten, können Pakete verloren gehen. Moderne Codecs benutzen komplexe Paketverlust-Verdeckungsalgorithmen, die versuchen fehlende Sprachpakete zu ersetzen. Dabei synthetisieren sie ein neues Sprachsignal, basierend auf Informationen vorangegangener Sprachpakete. In dieser Arbeit untersuchen wir, wie sich diese Verdeckungsalgorithmen, in Abhängigkeit der Position des verlorenen Sprachpakets innerhalb eines Satzes, auf die wahrgenommene Sprachqualität auswirken. Hierfür wird zwischen Paketverlusten am Anfang, in der Mitte, oder am Ende eines Lautes unterschieden. Neben der Position, in Relation zum betroffenen Laut, spielt auch die Stimmhaftigkeit des Lautes eine Rolle darin, wie sehr die wahrgenommene Qualität beeinträchtigt wird. Anhand des Verdeckungsalgorithmus des EVS Codecs konnte gezeigt werden, dass hier besonders Paketverluste, die am Anfang stimmloser Segmente und am Ende stimmhafter Segmente auftreten ins Gewicht fallen.

Einleitung

Die Sprachqualität von Kommunikationsnetzwerken hat sich in den letzten Jahren, unter anderem durch die Umstellung von Leitungsvermittlung zu Paketvermittlung, deutlich erhöht. Diese Verbesserung der Sprachqualität lässt sich zum einen auf die Erweiterung der verfügbaren Audiobandbreite und zum anderen auf die damit mögliche vollständige digitale Übertragung zurückführen. Die übermittelte Sprache klingt durch die erhöhte Bandbreite "volleründ klarer – somit kann der typisch dumpfe Klang des klassischen analogen Telefonnetzes durch die neue Technologie hinter sich gelassen werden. Durch die vollständig digitale Übertragung bei der Paketvermittlung kann es im Netzwerk selbst kaum zu Störungen des Sprachsignals kommen, da die Pakete vom Sender zum Empfänger unverändert übertragen werden. Netzwerke in denen Paketvermittlung angewandt werden sind zum Beispiel VoIP (Voice over IP), VoLTE (Voice over LTE) oder VoWiFi (Voice over WiFi). Auch die meisten Over-The-Top Anbieter, wie Whatsapp, Skype, oder Facebook, benutzen für die Sprachübertragung VoIP Netzwerke. Jedoch treten auch hier Qualitätsminderungen auf, diese werden durch Paketverluste verursacht. Ein Paket geht verloren, wenn die Zeit, um das Paket vom Sender zum Empfänger zu übertragen, zu lang ist. Wenn die vom Jitter-Buffer festgelegte Zeit überschritten ist, wird das Paket verworfen und ist somit verloren. In diesem Fall wendet der Codec ein sogenannten 'Packet-Loss Concealment' Algorithmus an. Dieser versucht das verlorene Paket im Sprachsignal zu verschleiern, um störende Unterbrechungen zu verhindern. Dabei nutzt der Codec Informationen, aus vorher erfolgreich übertragenen Paketen, um ein neues Sprachsignal zu synthetisieren. Moderne Codecs, wie AMR-WB und EVS klassifizieren das Signal zunächst und wenden dann eine geeignete Synthetisierungsmethode an. So wird zum Beispiel bei stimmhaften Vokalen der Laut periodisch weitergeführt und dabei die Amplitude langsam verringert, damit das Signal bei mehreren verlorenen Paketen langsam abklingt. Wird das Signal zu lange periodisch verlängert, kann dies zu künstlichen Tönen oder robotischen Stimmen führen.

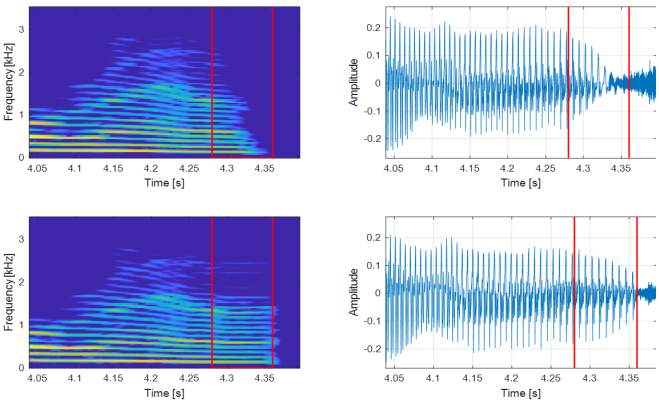

Abbildung 1: Beispiel eines Signals mit Paketverlusten. OBEN: Original Signal, UNTEN: Gestörtes Signal.

In Abbildung 1 ist exemplarisch ein Sprachsignal dargestellt, bei welchem Pakete während der Übertragung verloren gingen. Oben ist das originale Eingangssignal und im unteren Bereich das gestörte Ausgangssignal zu sehen. In dem Bereich zwischen den beiden roten Linien sind Pakete bei der Übertragung verloren gegangen. Im

Spektrogram kann man gut sehen, wie die Spektrallinien in dem verlorenen Bereich verlängert werden und der Ton so konstant gehalten wird. Auf der rechten Seite der Abbildung, im Zeitsignal, ist zu sehen, wie die Amplitude verringert wird während der Empfänger keine neuen Pakte erhält. Die perzeptiven Störungen, die durch solche Paketverluste verursacht werden können, sind neben robotischen Stimmen auch Unterbrechungen und rauschhafte Geräusche oder Klicks.

In dieser Arbeit wollen wir nun untersuchen, wie sich Paketverluste auf die wahrgenommene Sprachqualität auswirken, in Abhängigkeit der Position und Stimmhaftigkeit der betroffenen Stelle. Dazu haben wir zunächst eine große Datenbank mit Paktverlusten erstellt und die Sprachqualität objektiv mit P.OLQA [1] geschätzt. Anhand dieser Datenbank berechneten wir ein Modell, welches die Sprachqualität auf Basis der Position und Stimmhaftigkeit der Paketverluste schätzt. Anschließend führten wir einen Hörversuch durch, um den Einfluss der Paketverluste subjektiv zu untersuchen.

Trainingsdatenbank

Für die Trainingsdatenbank nahmen wir 16 Sprachdateien mit vier verschiedenen Sprachen aus der ITU-T P.501 [2] Datenbank als Referenzsignale. Diese bearbeiteten wir dann mit dem EVS Codec im Super-Wideband Modus und der höchsten Bitrate. Dadurch konnten wir sicherstellen, dass alle Qualitätsbeeinträchtigung nur durch die Paketverluste verursacht wurden und nicht durch die Kodierung selbst (EVS hat im höchsten Bitratenmodus keine Qualitätsbeeinträchtigungen, siehe auch [3]).

Insgesamt haben wir zwischen drei verschiedenen Positionen innerhalb eines Lautes unterschieden: Am Anfang, in der Mitte und am Ende. Für jede Position unterschieden wir zusätzlich, ob der Laut stimmhaft oder stimmlos ist. Für diese Unterscheidung verwendeten wir den RAPT Pitch-Tracker [4]. Insgesamt erhielten wir also sechs verschiedene Fehlerklassen, die wir gleichmäßig auf die Datenbank verteilten: stimmhaft am Anfang V_{FL}, in der Mitte V_{ML} und am Ende V_{EL}, und stimmlos am Anfang U_{FL}, in der Mitte U_{ML} und am Ende U_{EL}. In Abbildung 2 ist ein Sprachsignal aus der Datenbank beispielhaft gezeigt. Die blauen Stellen sind hier als stimmhaft und die roten Stellen als stimmlos markiert wurden. Abhängig von diesen Markierungen und der Fehlerklasse setzten wir dann Paketverluste in das Sprachsignal.

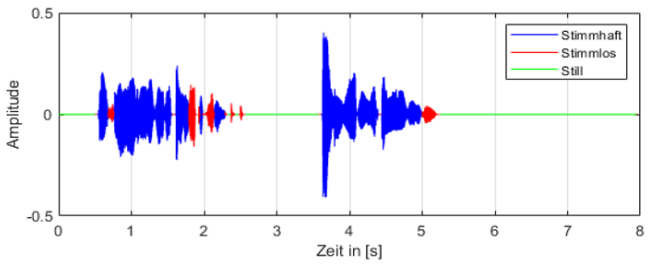

Abbildung 2: Unterscheidung zwischen stimmlosen und stimmhaften Lauten mit dem RAPT Pitch-Tracker

Weiterhin wird zwischen der Anzahl der hintereinander verloren gegangen Paketen unterschieden, was auch als Burstiness beschrieben werden kann. Wir untersuchten Bursts mit 2, 5 und 8 Paketen, wobei jedes Paket einen Frame von 20 ms Sprache beinhaltet – also Bursts mit 40 ms, 100 ms und 160 ms. Außerdem muss auch zwischen der Häufigkeit, in der eine Fehlerklasse auftritt, unterschieden werden. Wir verwendeten Häufigkeiten von 0-3 mal in der Trainingsdatenbank. Da wir auf eine sehr große Anzahl an Dateien kommen, wenn wir alle diese Kombinationen untersuchen wollen, haben wir uns auf 93 Bedingungen beschränkt, von welchen 18 nur die Einzeleinflüsse der sechs Hauptfehlerklassen beinhalten. Insgesamt erstellten wir somit 8835 Dateien, für welche wir mit P.OLQA die objektive Qualität schätzten.

Modellierung der Paketverluste

Um den Einfluss der Position und Stimmhaftigkeit der Paketverluste zu modellieren, greifen wir auf eine bereits existierende Formel aus dem E-Modell [5] zurück, welches aus technischen Parametern eines Sprachnetzwerkes die Qualität schätzen kann. Im E-Modell werden sogenannten Impairment-Faktoren von einer maximalen Qualität R_0 auf der R-Skala abgezogen, die Gesamtqualität kann dann folgendermaßen ausgerechnet werden:

$$R = R_0 - I_s - I_d - I_{e,eff} + A. \tag{1}$$

Hierbei beschreibt $I_{e,eff}$ die Qualitätsbeeinträchtigung durch Kodierung und Paketverluste. Die Formel für $I_{e,eff}$ ist gegeben als:

$$I_{e,eff} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{\frac{P_{pl}}{Burst} + B_{pl}}, \tag{2}$$

wobei I_e die Beeinträchtigung durch die reine Kodierung beschreibt und in unserem Fall dementsprechend $I_e = 0$ ist. P_{pl} gibt die Paketverlustrate im Sprachkanal an und $Burst$ ist die durchschnittliche Anzahl hintereinander verlorenere Pakete. B_{pl} gibt die Robustheit des Codecs gegenüber Paketverlusten an, diese wird basierend auf Hörversuchen für jeden Codec individuell festgelegt. Ausgehend von dieser Formel wollen wir nun Robustheiten für unsere sechs Fehlerklassen bestimmen. Dafür haben wir folgende neue Formel aufgestellt:

$$I_{e,eff} = C \cdot \sum_{i=1}^{6} \frac{P_{pl,i}}{P_{pl,i} + B_{pl,i} + \alpha \cdot Burst_i + \beta \cdot P_{pl,ges}}. \tag{3}$$

Hier beschreibt C eine allgemeine Skalierungsgröße, $P_{pl,i}$ die Paketverlustrate einer Fehlerklasse, $P_{pl,ges}$ die gesamte Paketverlustrate, $Burst_i$ die Burstiness einer Fehlerklasse und $B_{pl,i}$ die Robustheit der jeweiligen Fehlerklasse. Da die R-Skala mit den durchschnittlichen Qualitätsbewertungen aus Hörversuchen MOS (Mean Opinion Score) über die s-förmige Kurve in Abbildung 3 verbunden sind, können wir die mit P.OLQA geschätzten

MOS Werte nun nutzen, um die Koeffizienten aus Formel (3) über eine nicht-lineare Regression zu ermitteln.

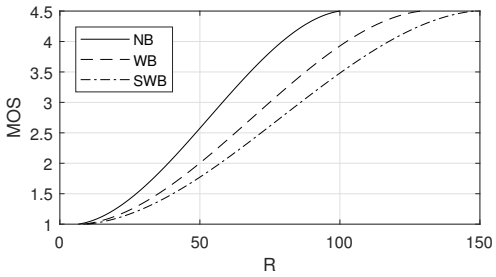

Abbildung 3: Transformationsregel des E-modell zwischen R-Skala and MOS

Ergebnisse Trainingsdatenbank

Die Ergebnisse des Fitting der nicht-linearen Regression sind in Abbildung 4 zu sehen. Der Fehler der Schätzung beträgt RMSE = 0.18 mit einer Pearsonkorrelation von $r = 0.87$. Wenn die Position und Stimmhaftigkeit nicht in der Formel berücksichtigt wird erhöht sich der Fehler auf RMSE = 0.28 mit einer Korrelation von $r = 0.68$.

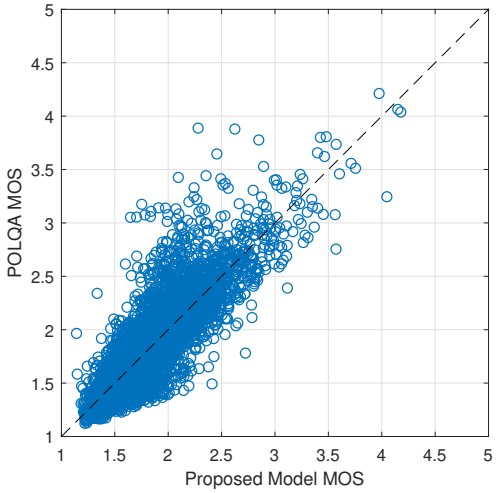

Abbildung 4: Schätzung des vorgeschlagenen Modells vs POLQA MOS für die Trainings-Datenbank

Die ermittelten Koeffizienten sind in Tabelle 1 dargestellt. Aus ihnen lässt sich nun ablesen, welche Fehlerklassen besonders zu einer Qualitätsbeeinträchtigung beitragen und welche weniger. Die Robustheit für die stimmlosen Paketverluste am Anfang eines Lautes und die stimmhaften Paketverluste am Ende eines Lautes sind deutlich niedriger als der Rest, das heißt, dass diese beide Klassen die Qualität am meisten beeinträchtigen. Hingegen sind die Paketverluste bei stimmlosen Lauten am Ende und stimmhaften Lauten am Anfang besonders hoch – hier scheint die Qualitätsbeeinträchtigung also merkbar geringer zu sein. Um die gefunden Ergebnisse zu validieren, führten wir anschließend noch einen Hörversuch durch.

Tabelle 1: Ergebnisse der nicht-lineare Regression für die Trainingsdatenbank

Skalierung Gesamt	C	222,023
Skalierung für $Burst_i$	α	-0,662
Skalierung für $P_{pl,ges}$	β	1,547
Robustheit U_{FL}	$B_{pl,1}$	2,885
Robustheit U_{ML}	$B_{pl,2}$	7,703
Robustheit U_{EL}	$B_{pl,3}$	9,484
Robustheit V_{FL}	$B_{pl,4}$	9,01
Robustheit V_{ML}	$B_{pl,5}$	6,925
Robustheit V_{EL}	$B_{pl,6}$	3,385

Hörversuch

An dem Hörversuch nahmen 37 deutsche Muttersprachler teil, welche die Sprachsignale auf einer 5-stufigen MOS-Skala nach ITU-T P.800 [6] bewerteten. Als Referenzsignale nutzten wir die deutschen Sätze der ITU-T P.501 Datenbank. Ingesamt beinhaltet die Datenbank 15 verschiedene Bedingungen die in Tabelle 2 dargestellt sind. Dabei entsprechen 6 der Bedingungen den Einzelfehlerklassen, mit einer jeweiligen Häufigkeit von 4 Paketverlusten mit einer Länge vom 100 ms. Die erste Bedingung ist das ungestörte Referenzsignal und die zweite Bedingung ein starkes Rauschen, welches als Ankerbedingung dient. Weiterhin gibt es noch 7 Bedingungen, in welchen Kombinationen verschiedener Fehlerklassen auftreten. Pro Bedingung wurden vier verschiedene Dateien prozessiert, mit jeweils zwei weiblichen und zwei männlichen Sprechern, so dass es ingesamt 148 Bewertungen pro Bedingung gibt. Die Ergebnisse des Hörtest sind in Abbildung 5 dargestellt und bestätigen den schon in der Trainingsdatenbank gefunden Trend: Paketverluste in stimmlosen Bereichen werden als nur wenig störend wahrgenommen, es sei denn sie befinden sich am Anfang eines Lautes, wohingegen Paketverluste in stimmhaften Bereichen im Allgemeinen störender wahrgenommen werden, vor allem wenn sie sich am Ende eines Lauter befinden. Die Kombinationen verschiedener Fehlerklassen ergibt mehr oder weniger den Mittelwert der derjenigen Klassen.

Tabelle 2: Bedingungen der Datenbank für den Hörtest

Nr	Condition
1	Clean Reference
2	MNRU Noise
3	4 x U_FL
4	4 x U_ML
5	4 x U_EL
6	4 x V_FL
7	4 x V_ML
8	4 x V_EL
9	2 x V_FL + 2 x U_FL
10	2 x V_ML + 2 x U_ML
11	2 x V_EL + 2 x U_EL
12	2 x V_FL + 2 x V_EL
13	2 x U_FL + 2 x U_EL
14	2 x V_FL + 2 x U_EL
15	2 x U_FL + 2 x V_EL

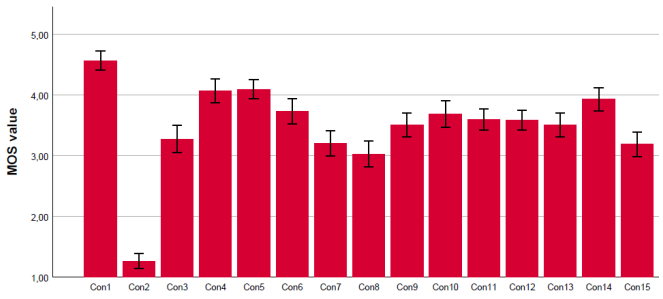

Abbildung 5: Ergebnisse des Hörtests: MOS pro Bedingung

Zuletzt prüfen wir unser Modell anhand der Daten aus dem Hörversuch. Dafür nahmen wir als Eingang zu dem Modell, die Position, Stimmhaftigkeit und Paketverlustrate, welche wir auch für die Erstellung der Datenbank nutzten, und berechneten mit der Formel (3) die geschätzte Qualität. Die Ergebnisse sind in Abbildung 6 zu sehen. Der Fehler bei POLQA beträgt RMSE = 0.61 mit einer Pearson Korrelation von $r = 0.61$, wohingegen der Fehler des vorgestellten Modells nur RMSE = 0.32 mit einer Pearson Korrelation von $r = 0.82$. Dies lässt darauf schließen, dass die Qualitätsschätzung verbessert werten kann, wenn bei Paketverlusten die Position und Stimmhaftigkeit berücksichtigt wird.

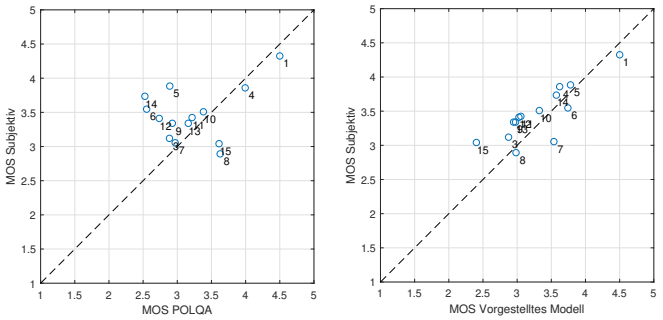

Abbildung 6: Korrelationsdiagramm zur Schätzung der Qualität aus dem Hörversuch. LINKS: POLQA, RECHTS: Vorgestelltes Modell (Formel (3)).

Fazit

In diesem Beitrag stellten wir Untersuchungen zur Qualitätswahrnehmung von verdeckten Paketverlusten, in Abhängigkeit der Position und Stimmhaftigkeit der betroffenen Stelle vor. Wir konnten zeigen, dass verlorene Pakete in stimmlosen Bereichen im Allgemeinen als weniger störend wahrgenommen werden, im Vergleich zu stimmhaften Bereichen. Stimmhafte Paketverluste, die in der Mitte oder am Ende stattfinden sowie stimmlose Paketverluste am Anfang eines Lautes, werden dabei als besonders störend wahrgenommen. Dies kann damit erklärt werden, dass sich langgezogenen Vokale oft künstlich bzw. robotisch anhören. Weiterhin kommt es am Anfang von stimmlosen Lauten eher zu hörbaren Unterbrechungen. Wir konnten Zeigen, dass sich die Qualitätsschätzung durchaus verbessern lässt, wenn die Position und Stimmhaftigkeit bei verdeckten Paketverlusten berücksichtigt wird.

Literatur

[1] ITU-T Rec. P.863: Perceptual objective listening quality assessment

[2] ITU-T Rec. P.863: Test signals for use in telephonometry

[3] G. Mittag, S. Möller, V. Barriac and S. Ragor, "Quantifying Quality Degradation of the EVS Super-Wideband Speech Codec,"2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, 2018, pp. 1-6.

[4] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", in "Speech Coding and Synthesis", W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995

[5] ITU-T Rec. G.107: The E-model: a computational model for use in transmission planning

[6] ITU-T Rec. P.800: Methods for subjective determination of transmission quality