EINFLUSS VON PAKETVERLUSTEN AUF DIE QUALITÄT VON SPRACHERKENNUNG UND SPRACHSYNTHESE

Sebastian Möller, Jan Krebber und Alexander Raake

Institut für Kommunikationsakustik (IKA), Ruhr-Universität Bochum

sebastian.moeller@ruhr-uni-bochum.de

Abstract: In diesem Beitrag wird der Einfluss eines zeitlich instationären Telefonkanals auf die Erkennungsleistung von Sprach- und Sprechererkennern sowie auf die Qualität maschinell erzeugter Sprache untersucht. Die betrachtete Situation ist von Bedeutung z.B. bei der Interaktion mit einem Sprachdialogsystem über eine gestörte paketvermittelte Telefonverbindung. Zur Untersuchung der Kanaleigenschaften wurde ein Simulationssystem erstellt, welches Paketverluste auf kontrollierte Weise generiert. Mit Hilfe dieses Simulationssystems wurden Sprachdaten verarbeitet und unterschiedlichen Sprach- und Sprechererkennern bzw. beurteilenden Versuchspersonen präsentiert. Die Versuchsergebnisse zeigen einen annähernd linearen Abfall der Erkennungsleistung in Abhängigkeit von der eingestellten Paketverlustrate bei den deutschen Erkennern, während in allen anderen Fällen eine deutliche Verschlechterung erst ab Verlustraten von ca. 3-5% zu beobachten ist. Gründe für diese Unterschiede werden diskutiert, und es wird ein Ausblick auf weitere notwendige Untersuchungen gegeben.

1 Einleitung

Moderne Telekommunikationsnetze benutzen in immer stärkerem Maße paketorientierte Übertragungstechniken zur Übermittlung von Sprache und Daten. Die dabei auftretenden Störungen unterscheiden sich wesentlich vom traditionellen Telefonnetz; insbesondere ist ein zeitvariantes Übertragungsverhalten zu beobachten, welches aus dem Verlust von Paketen resultiert. Zeitvariantes Verhalten stellt insbesondere dann ein Problem dar, wenn zeitkritische Daten übertragen werden müssen, wie dies z. B. bei der Sprachkommunikation über das Internet (*Voice-over-IP*, VoIP) der Fall ist. Die Sprachübertragung erfordert nämlich einen kontinuierlichen Datenstrom – eine Forderung, die im Gegensatz zum Prinzip der asynchronen Übertragung des Internets steht. Dennoch stellt das Internet auch für die Sprach- und Bildübertragung in Realzeit eine interessante Alternative zum herkömmlichen Telefonnetz dar.

Paketverluste entstehen hauptsächlich durch die fehlende Prioritäts-Zuweisung in paketvermittelten Netzen. Die Datenpakete nehmen i.A. unterschiedliche Übertragungswege vom Sender zum Empfänger und weisen daher unterschiedliche Laufzeiten auf, wenn sie beim Empfänger eintreffen (sog. *delay jitter*). Diese Laufzeitunterschiede werden zunächst in einem Jitter-Buffer abgefangen, welcher die eintreffenden Datenpakete zwischenspeichert und sortiert. Dies führt jedoch zu einer Gesamtverzögerung des Sprachsignals, die möglichst gering gehalten werden sollte um die Kommunikationsfähigkeit nicht zu beeinträchtigen. Überschreitet die Verzögerung eines Datenpakets die maximale Verzögerung, die durch den Jitter-Buffer ausgeglichen werden kann, so gilt dieses Paket als verloren. Korrekturalgorithmen (sog. *packet loss concealment*, PLC) ersetzen das fehlende Paket entweder durch Nullen oder generieren ein Ersatzpaket aus den Informationen vorangegangener Pakete.

Die durch Paketverluste hervorgerufenen wahrnehmbaren Beeinträchtigungen sind in der Vergangenheit bereits Gegenstand mehrerer Untersuchungen gewesen, vgl. z.B. [9] oder [1]. Dabei konnte gezeigt werden, dass eine Qualitätsbeeinträchtigung nicht nur von der Anzahl

verlorener Pakete pro Zeiteinheit (Verlustrate) abhängt, sondern auch von der zeitlichen Verteilung verlorener Pakete, dem verwendeten Kodierer und dem Korrekturalgorithmus. Für den menschlichen Zuhörer ist das Einfügen von Nullen (Pausen) oft stärker störend als eine Interpolation des fehlenden Sprachabschnitts, bspw. durch eine pitch-synchrone Wiederholung vorangegangener Pakete. Auch treten Verluste in der Realität meist gehäuft (in sog. *bursts*) und nicht rein zufällig auf; diese Häufung kann ebenfalls zu einer weiteren Verschlechterung der Sprachqualität gegenüber einer rein zufälligen Verteilung der Paketverluste führen.

Im Gegensatz zur wahrgenommenen Sprachqualität ist der Einfluss zeitvarianter Störungen auf die Leistung sprachtechnologischer Komponenten (Spracherkennung, Sprechererkennung, Sprachsynthese) bislang nur wenig erforscht[1]. Dabei kommen diese Komponenten in Sprachdialogsystemen zum Einsatz, welche gerade in paketvermittelten Netzen Anwendung finden. Ziel der hier vorgestellten Untersuchungen ist es, den Einfluss von Paketverlusten

- auf die Erkennungsleistung von Spracherkennern und Sprechererkennern sowie
- auf die wahrgenommene Qualität synthetischer Sprache

zu quantifizieren. Die betrachteten sprachtechnologischen Komponenten finden in einem sprachgesteuerten System zur Bedienung verschiedener Hausgeräte Anwendung, welches im Rahmen des europäischen IST-Projektes INSPIRE (INfortainment management with SPeech Interaction via REmote microphones and telephone interfaces) aufgebaut wurde.

Zur detaillierten Untersuchung des Einflusses von Paketverlusten wurde zunächst ein Simulationssystem für IP-Kanäle implementiert, mit dem Sprachdaten kontrolliert gestört werden können. Dieses System ist in Abschnitt 2 beschrieben. Mit Hilfe des Simulationssystems wurden dann aufgezeichnete Sprachdaten gezielt gestört und einem Sprach- bzw. einem Sprechererkenner zugeführt. Daneben wurden natürliche und synthetisierte Sprachprompts, wie sie im INSPIRE-System Verwendung finden, durch Paketverluste gestört und in einem anwendungsnahen Hörversuch von Versuchspersonen beurteilt. Der Aufbau dieser Versuche ist in Abschnitt 3 beschrieben. Die Ergebnisse werden bezüglich der unterschiedlichen Auswirkungen von Paketverlusten gleicher Verlustrate diskutiert und mit Modellvorhersagen verglichen (Abschnitt 4). Eine abschließende Diskussion in Abschnitt 5 beleuchtet die bislang gemachten Vereinfachungen und zeigt zukünftige Untersuchungsgegenstände auf.

2 VoIP-Simulation

In realen Kommunikationsnetzen ist eine genaue Kontrolle der Verbindung und der auftretenden Störungen i.A. nicht möglich. Deshalb wurde eine Simulationsumgebung erstellt, mit deren Hilfe Störungen generiert werden, wie sie für eine paketorientierte Übertragung typisch sind. Häufig wird die Verbindung nicht ausschließlich über paketvermittelte Netze aufgebaut, sondern ein Teil der Verbindung wird über das herkömmliche leitungsvermittelte Netz (z.B. ISDN) realisiert. Deshalb wurde die VoIP-Simulation mit einer Simulationsumgebung für leitungsgebundene Telefonie gekoppelt, welche in [6] beschrieben ist. Mit Hilfe der beiden Simulationen lassen sich alle wichtigen Störungen, die auf dem Weg zwischen dem Mund des Sprechers (oder einer an das Netz angeschlossenen Sprachsynthese) und dem Ohr des Zuhörers (oder einem Sprach- oder Sprechererkenner) auftreten, generieren.

Der schematische Aufbau der VoIP-Simulation ist in Abbildung 1 gezeigt. Das Sprachsignal wird zunächst digitalisiert und mittels einer logarithmischen PCM (nach ITU-T Rec. G.711) in ein 8-Bit-Format kodiert. Diese auch im ISDN verwendete Kodierung stellt die Eingangsschnittstelle zum Gateway dar, welches das digitalisierte Sprachsignal in IP-Pakete umsetzt.

[1] Eine Ausnahme bilden Untersuchungen zur Spracherkennung in IP-basierten Netzen, wie sie in [1] und [5] beschrieben sind.

Dazu werden zunächst Sprachpausen entfernt und das Sprachsignal anschließend weiter komprimiert. Das von uns eingesetzte Gateway Cisco 2611 verwendet hierfür eine Kodierung nach ITU-T Rec. G.711 oder G.729. Die Pakete werden dann über das Netzwerk-Emulations-Programm NISTnet [8] "übertragen". Diese Emulation eliminiert ("verliert") einzelne Pakete mit einer vorher definierten Wahrscheinlichkeit. Die ankommenden Pakete werden im Jitter-Buffer zwischengespeichert und sortiert, in ein digitales Signal umgesetzt und dekodiert. Aus dem nach G.711 kodierten Ausgang wird abschließend das analoge Signal rekonstruiert.

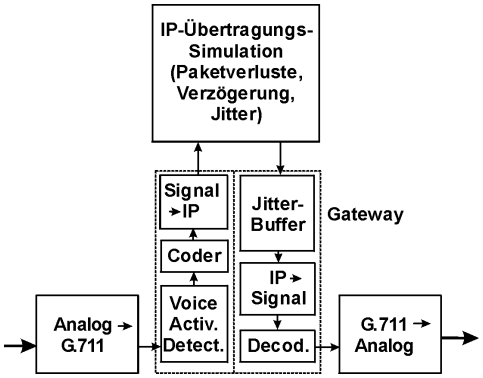

Abbildung 1 - Schematische Darstellung der VoIP-Simulation

Die Eingangssignale der VoIP-Simulation wurden zuvor über ein typisches Endgerät (sog. "7-er" Handapparat) eingespeist oder von einem Text-nach-Sprache-System synthetisiert. Bei natürlicher Sprache wurde die akustisch-elektrische Übertragungsfunktion des Handapparates mit Hilfe eines Korrekturfilters so modifiziert, dass sich eine IRSmod-Charakteristik nach ITU-T Rec. P.830 ergab. Anschließend wurde ein Standard-Rauschpegel von -70 dBm0p addiert, welcher typisch für das Leitungsrauschen eines Festnetzes ist. Das Signal wurde dann mittels eines Telefon-Bandpasses auf 300-3400 Hz bandbegrenzt und der VoIP-Simulation zugeführt. Am Ausgang der VoIP-Simulation wurde ein breitbandiges Rauschen von -64 dBmp addiert, welches empfängerseitige Rauschquellen simuliert. Das so verarbeitete Ausgangssignal wurde den menschlichen Versuchspersonen über einen Standard-Handapparat (Frequenzgang nach ITU-T Rec P.64) präsentiert, oder es wurde einem Sprach- oder Sprechererkenner (mit flachem Frequenzgang) zugeführt.

3 Versuchsaufbau

Telefonkanäle sind auf die Anforderungen eines menschlichen Zuhörers bzw. Dialogpartners abgestimmt. Bei der Interaktion mit einem Sprachdialogsystem müssen jedoch auch Sprach- und Sprechererkenner robust gegen die durch den Übertragungskanal eingebrachten Signalveränderungen sein. Es ist anzunehmen, dass sich die Anforderungen des menschlichen Zuhörers von denen eines Spracherkenners deutlich unterscheiden; daher werden auch die Auswirkungen von Störungen (hier von Paketverlusten) in beiden Fällen unterschiedlich sein. Hinzu kommt, dass die von einem Sprachdialogsystem generierte Sprache zunächst über den Telefonkanal übertragen wird, bevor sie vom Zuhörer wahrgenommen wird. Die Auswirkungen des Kanals werden von der Art der maschinell generierten Sprache (natürliche, vorher aufgezeichnete vs. synthetisierte Sprache) abhängen. Es werden also beide Schnittstellen eines Sprachdialogsystems durch den Übertragungskanal beeinflusst.

Im Folgenden werden Experimente beschrieben, die diese Unterschiede quantitativ erfassen sollen. Dabei stehen vier Situationen im Vordergrund:

(1) Ein Spracherkenner wird mit Sprache konfrontiert, die durch Paketverluste gestört ist.

(2) Ein Sprechererkenner wird mit Sprache konfrontiert, die durch Paketverluste gestört ist.

(3) Ein menschlicher Zuhörer wird mit natürlich erzeugter Sprache konfrontiert, die durch Paketverluste gestört ist.
(4) Ein menschlicher Zuhörer wird mit synthetisch generierter Sprache konfrontiert, die durch Paketverluste gestört ist.

Alle vier Situationen treten beim betrachteten INSPIRE-Sprachdialogsystem auf, vgl. Abschnitt 3.1. Der verwendete Versuchsaufbau ist in Abschnitt 3.2 und 3.3 beschreiben. Zur Vorhersage der Qualität in Situation (3) existieren bereits Modelle, die auf Grundlage der Eigenschaften des Übertragungskanals einen Schätzwert für die vom Benutzer im Mittel erfahrene Qualität liefern. Ein solches Modell wird in Abschnitt 3.4 kurz vorgestellt.

3.1 INSPIRE-Smart-Home-System

Im Rahmen des INSPIRE-Projektes wurde ein Sprachdialogsystem entwickelt, mit dessen Hilfe verschiedene Hausgeräte (Fernseher, Videorecorder, Lampen, Rollos, Ventilator, Anrufbeantworter, etc.) über Sprache gesteuert werden können. Die Steuerung kann entweder direkt im Haus erfolgen oder über einen Anruf beim System. Im letzteren Falle wird der Erfolg der Interaktion maßgeblich vom Telefonkanal beeinflusst; dieser Fall ist Gegenstand der hier geschilderten Untersuchungen.

Bislang wurden eine deutsche und eine griechische Variante des INSPIRE-Systems implementiert. Beide Varianten verwenden eine kommerziell verfügbare Spracherkennung: Das deutsche System arbeitet mit einem Einzelworterkenner, welcher Schlüsselwörter extrahiert, ohne grammatikalische Regeln anzuwenden; das griechische System arbeitet mit einer kontinuierlichen Spracherkennung, die eine Grammatik sowie ein vom Dialogzustand abhängiges Vokabular verwendet. Daneben verfügen beide Varianten über eine Sprecherverifikation, welche auf deutsche oder griechische Sprecher trainiert werden kann. Auf der Sprachausgabeseite verwenden beide Varianten entweder synthetisierte oder natürliche, vorher aufgezeichnete Sprache. Im hier getesteten deutschen System wurde die AT&T-Unit-Selection-Synthese benutzt, welche Sprachbausteine unterschiedlicher Länge verkettet.

3.2 Sprach- und Sprechererkennung bei Paketverlusten

Zur Untersuchung des Einflusses von Paketverlusten auf die Erkennungsleistung der verwendeten Spracherkenner wurden zunächst typische Benutzeräußerungen in einer ruhigen Umgebung aufgezeichnet. Dazu wurden für jede Sprache zunächst 10 "typische" Dialoge entworfen, welche dann von jeweils 10 Muttersprachlern (5 männlich, 5 weiblich) vorgelesen wurden. Diese Sprachdaten wurden über einen Kunstkopf Typ HeadAcoustics HMS II.3 abgespielt und über einen typischen Telefonhörer in einer ruhigen Büroumgebung aufgezeichnet.

Die so erzeugten Quelldaten (1370 deutsche und 500 griechische Äußerungen) wurden mittels der beschriebenen VoIP-Simulation verarbeitet. Zwei unterschiedliche Kodierer kamen (zusätzlich zur ISDN-Kodierung) hierbei zum Einsatz: Eine logarithmische PCM nach ITU-T Rec. G.711 (64 kbit/s) mit einem proprietären PLC-Algorithmus, sowie eine CS-ACELP-Kodierung nach Rec. G.729A (8 kbit/s) mit dem in dieser Empfehlung spezifizierten PLC-Algorithmus. Beide Kodierer werden typischerweise in paketvermittelten Netzen eingesetzt. Mit Hilfe der NISTnet-Netzwerkemulation wurden auf dem kodierten Datenstrom Paketverluste mit einer Verlustrate von 0, 3, 5, 10 oder 15% generiert. Die Sprach-Pausen-Detektion wurde ausgeschaltet, um einen Einfluss der Sprach-Pausen-Verteilung unserer Testdaten auf die eingestellte Verlustrate auszuschließen.

Auf diese Weise wurden die Quelldaten (1870 Dateien) mit je 10 unterschiedlichen Störprofilen (2 Kodierer x 5 Verlustraten) versehen. Diese Ausgangsdaten wurden dann den jeweiligen Sprach- bzw. Sprechererkennern zugeführt. Da es sich bei den Spracherkennern um

kommerzielle Module handelt war ein Training bzw. eine Adaption nicht möglich. Die Sprecherverifikation wurde zunächst mit Sprachdaten aus 5 Dialogen trainiert, um sprecherspezifische Modelle zu erzeugen. Die Rückweisungsschwelle wurde mittels Sprachdaten von 4 weiteren Dialogen eingestellt.

Da die beschriebenen Experimente noch vor der Systemoptimierung durchgeführt wurden waren die Erkennungsleistungen nicht in allen Fällen optimal. So erreicht der deutsche Spracherkenner nur eine maximale Erkennungsrate von 69,0%, der griechische Spracherkenner hingegen 98,7%. Dieser deutliche Unterschied ist durch das Fehlen einer Grammatik beim deutschen Erkenner zu erklären. Er ist im vorliegenden Fall tolerierbar, da nur die relative Verschlechterung der Erkennungsrate in Abhängigkeit von den Kanaleigenschaften (und nicht die absolute Erkennungsrate) betrachtet werden soll. Die Sprecherverifikation erreicht eine maximale Erkennungsrate von 93,3% (deutsch) bzw. 98,2% (griechisch).

3.3 Qualität natürlicher und synthetischer Sprache bei Paketverlusten

Zur Bestimmung der wahrgenommenen Qualität von durch Paketverluste gestörter Sprache wurde ein Hörversuch mit 24 Versuchspersonen durchgeführt. Die Versuchspersonen waren zwischen 20 und 70 Jahren alt (Mittelwert 31,6 Jahre). Etwa die Hälfte von ihnen hatte bereits Erfahrung mit Sprachdialogsystemen.

Für diesen Hörversuch wurden unterschiedliche deutsche Sprachprompts generiert, welche eine konstante Komplexität aufweisen und für die Versuchspersonen jeweils wechselnde Informationen beinhalten. Die enthaltene Information ist entweder statisch, d.h. sie wird nur einmal bei der Implementierung des Systems definiert, oder dynamisch, d.h. die Information kann sich während der Benutzung des Systems ändern (bspw. das aktuelle Fernsehprogramm). Statische Informationen wurden von einem männlichen Sprecher in einer Hörkabine vorgelesen und aufgezeichnet. Dynamische Informationen wurden synthetisiert.

Die so erzeugten Quelldaten wurden wiederum durch die VoIP-Simulation kontrolliert "gestört" (hier nur mit einem G.729A-Kodierer) und den Versuchspersonen über einen Standard-Handapparat in einer Büroumgebung vorgespielt. Die Versuchspersonen wurden zunächst nach dem Gerät befragt, auf welches sich die dargebotene Information bezieht; auf diese Weise sollte verhindert werden, dass sich die Versuchspersonen ausschließlich auf die Form und nicht auf den Inhalt der dargebotenen Sprache konzentrieren. Anschließend bewerteten sie vier Qualitätsaspekte der dargebotenen Sprachprobe auf kontinuierlichen Skalen mit 5 oder 2 Labeln, wie sie in ITU-T Rec. P.851 [4] empfohlen werden:

- Wie beurteilen Sie den Gesamteindruck des gerade Gehörten? (ausgezeichnet; gut; ordentlich; dürftig; schlecht)
- Welche Anstrengung war notwendig, um die Bedeutung des Gesagten zu verstehen? (keine Anstrengung erforderlich, entspanntes Zuhören möglich; keine wesentliche Anstrengung, aber Aufmerksamkeit erforderlich; mäßige Anstrengung erforderlich; erhebliche Anstrengung erforderlich; selbst größte Anstrengung genügt nicht zum Verstehen)
- Wie angenehm fanden Sie die gerade gehörte Stimme? (angenehm; unangenehm)
- Wie gut passt die gerade gehörte Stimme zum beschriebenen System? (ausgezeichnet; gut; ordentlich; dürftig; schlecht)

Das Urteil der Versuchspersonen wurde auf einen Zahlenbereich [0;6] abgebildet, wobei der Wert 5 dem positivsten Attribut der Skala und der Wert 1 dem negativsten Attribut entspricht, und die Werte 0 und 6 den Endpunkten der Skala. Die gemittelten Werte werden im Folgenden mit 'Gesamtqualität', 'Höranstrengung', 'Annehmlichkeit der Stimme' bzw. 'Angemessenheit der Stimme' bezeichnet.

3.4 Qualitätsmodellierung

Zur Planung von Telefonnetzen wurden in der Vergangenheit Vorhersagemodelle entwickelt, die einen Schätzwert für die Sprachqualität in Abhängigkeit von den Parametern der Übertragungsstrecke liefern. Das bekannteste Modell ist das sog. E-Modell, welches in ITU-T Rec. G.107 [3] beschrieben ist. Dieses Modell verwendet gerade diejenigen Parameter als Eingangsgrößen, welche mit der von uns verwendeten Simulation eingestellt werden können; es lässt sich also direkt ein Schätzwert für die Sprachqualität ermitteln, welcher den Einstellungen des Telefonkanals entspricht. Der Schätzwert des Modells kann auf eine 5-stufige ACR-Skala transformiert werden, die mit den selben Attributen wie die hier zur Bestimmung der Gesamtqualität verwendete Skala beschriftet ist. Auf diese Weise lassen sich die auditiv ermittelten Testergebnisse mit den Schätzungen des E-Modells vergleichen. Es ist zu beachten, dass sich die Schätzungen des Modells auf die Mensch-zu-Mensch-Kommunikation beziehen; sie sind also nicht zur Vorhersage der Qualität übertragener *synthetischer* Sprache gedacht.

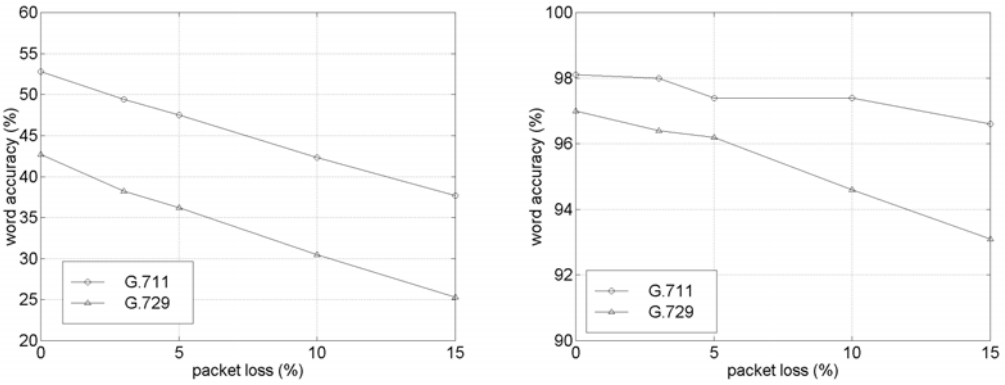

Abbildung 2 - Einfluss von Paketverlusten auf die Erkennungsleistung des deutschen (links) und des griechischen Spracherkenners (rechts)

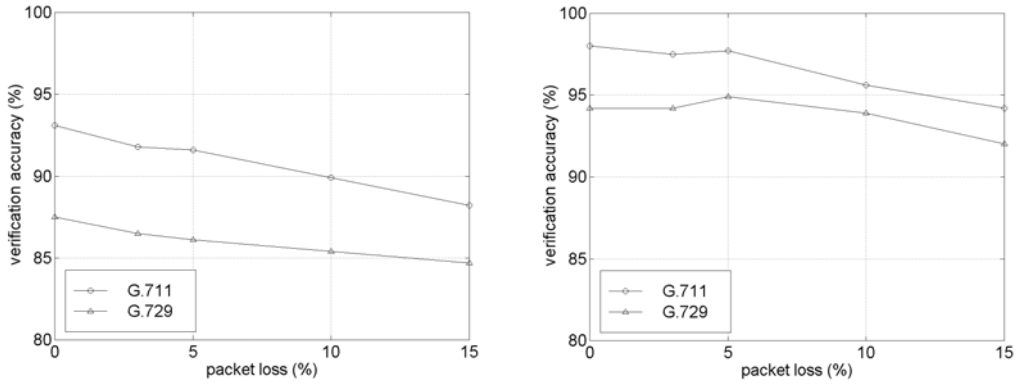

Abbildung 3 - Einfluss von Paketverlusten auf die Erkennungsleistung der deutschen (links) und der griechischen Sprecherverifikation (rechts)

4 Ergebnisse

4.1 Einfluss auf die Spracherkennungsleistung

Abbildung 2 zeigt die Beeinträchtigung der Worterkennungsrate in Abhängigkeit von der Paketverlustrate, für beide verwendeten Kodierer und Sprachen. Die Erkennungsleistung des deutschen Spracherkenners nimmt annähernd linear mit der Verlustrate ab, und zwar um etwa 1% pro Zunahme der Verlustrate um 1%. Ein ähnliches Verhalten wurde schon in [1] für den Kodierer nach ITU-T Rec. G.723.1 beobachtet. Demgegenüber scheint der griechische Spracherkenner etwas robuster gegen Paketverluste zu sein: Hier nimmt die Erkennungsrate

bei G.711 nur um etwa 1,5% bei 15% Paketverlusten ab. Der Einfluss auf G.729-kodierte Sprache ist größer, aber ebenfalls nicht so groß wie beim deutschen Erkenner.

4.2 Einfluss auf die Sprechererkennungsleistung

Wie schon bei der deutschen Spracherkennung fällt auch bei der Sprecherverifikation die Erkennungsrate fast linear mit der Verlustrate ab. Allerdings ist der Abfall deutlich schwächer, als dies bei der Spracherkennung zu beobachten war, vgl. Abbildung 3. Bei der griechischen Variante ist der Abfall erst bei höheren Verlustraten zu beobachten; es zeigt sich ein leichter "Schwelleneffekt", d.h. die Erkennungsleistung bleibt zunächst relativ konstant und fällt erst ab einer Mindest-Verlustrate, die hier etwa 5% (G.711) bzw. 10% (G.729) beträgt.

4.3 Einfluss auf die Qualität der Sprachausgabe

Den Einfluss von Paketverlusten auf die von den Versuchspersonen beurteilte Gesamtqualität natürlicher und synthetisierter Sprache zeigt Abbildung 4. Offensichtlich erfolgt eine merkliche Beeinträchtigung wiederum erst ab einer bestimmten Mindest-Verlustrate. Diese liegt für natürliche Sprache bei etwa 5%, für synthetisierte bei etwa 3%. Oberhalb dieser Mindest-Verlustrate fällt die Gesamtqualität deutlich ab, um etwa eine Skalenkategorie bei 15% Verlustrate. Ein ähnlich starker Abfall ist für die anderen Urteile zu beobachten; als Beispiel ist der Verlauf für die Höranstrengung im rechten Diagramm dargestellt.

Die Vorhersagen des E-Modells zeigen keinen Schwelleneffekt, sagen jedoch den relativen Verlauf bei höheren Verlustraten recht gut vorher. Dies gilt insbesondere für natürliche, eingeschränkt jedoch auch für synthetisierte Sprache. Der Gesamteindruck der synthetisierten Sprache ist gegenüber dem von natürlicher durch einen (negativen) Offset verschoben, der nicht vom Modell berücksichtigt wird, da es sich hierbei nicht um einen Einfluss des Kanals, sondern des Quellsignals handelt. Insgesamt startet die Beeinträchtigung bei synthetisierter Sprache schon bei niedrigeren Verlustraten als dies bei natürlicher Sprache zu beobachten ist. Ein Grund hierfür könnte eine im Vergleich zur natürlichen Sprache verringerte Redundanz der synthetisierten Sprache sein, die sie anfälliger für Paketverluste werden lässt. Dieses Verhalten steht im Gegensatz zu den in [7] dargestellten Untersuchungen, bei denen sich synthetisierte Sprache robuster gegenüber additivem Rauschen als natürlich erzeugte Sprache zeigte.

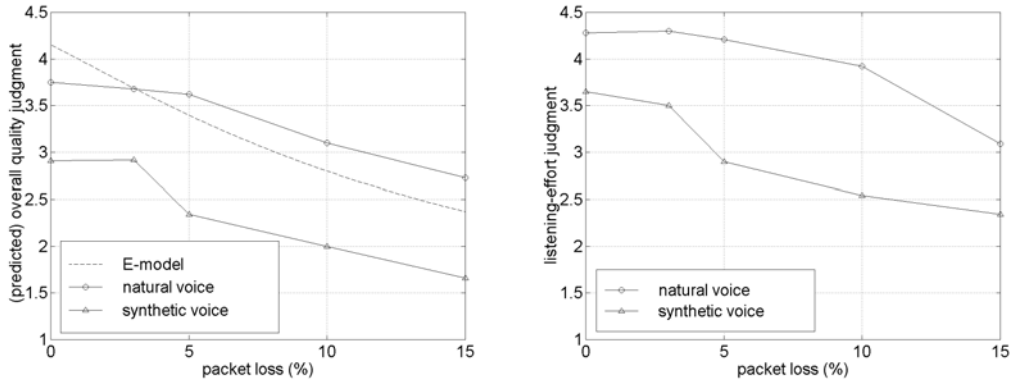

Abbildung 4 - Einfluss von Paketverlusten auf die wahrgenommene Qualität natürlich und synthetisch erzeugter Sprache (G.729A-Kodierer). Links: Gesamtqualität; rechts: Höranstrengung

5 Diskussion und Ausblick

Die experimentellen Daten zeigen, dass sich Paketverluste bei Sprach- und Sprechererkennung sowie bei der Beurteilung natürlicher oder synthetisierter Sprache unterschiedlich auswirken. Während bei der deutschen Sprach- und Sprechererkennung ein fast linearer Abfall

der Erkennungsleistung mit der Verlustrate zu beobachten ist, so zeigt sich bei den griechischen Erkennern ein merklicher Abfall erst ab einer bestimmten Mindest-Verlustrate. Ein ähnlicher "Schwelleneffekt" ist auch bei der Beurteilung natürlich erzeugter oder synthetisierter Sprache zu verzeichnen. Er wurde in früheren Untersuchungen (z.B. [9]) bislang noch nicht beobachtet.

Die beschriebenen Arbeiten stellen eine erste kontrollierte Untersuchung des Einflusses von Paketverlusten auf unterschiedliche sprachtechnologische Komponenten dar. Sie ist zugegebenermaßen noch recht begrenzt; insbesondere wurde der Einfluss der zeitlichen Verteilung der Verluste nicht näher betrachtet. Sie dürfte von grundsätzlicher Bedeutung sein und sollte in zukünftigen Versuchen realistischer simuliert werden, bspw. durch Modellierung von *bursts*. Auch sollte der Einfluss der Kodierer und PLC-Algorithmen näher betrachtet werden; beide werden hauptsächlich auf ihre perzeptiven Effekte bei natürlicher Sprache optimiert, aber nicht auf ihr Zusammenspiel mit sprachtechnologischen Komponenten.

Danksagung

Die vorliegende Arbeit wurde am IKA, Ruhr-Universität Bochum, im Rahmen des EU-geförderten Projektes INSPIRE (IST-2001-32746; `www.inspire-project.org`) durchgeführt. Die Autoren danken allen Partnern von INSPIRE, und insbesondere Alex Trutnev (EPFL), Anestis Vovos (Knowledge S.A.) und Todor Ganchev (WCL) für die Bereitstellung der Erkennungsergebnisse.

Literatur

[1] Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de-María, F.: A Robust Front-End for ASR over IP and GSM Networks: An Integrated Scenario. In: Proc. Eurospeech 2001 – Scandinavia, Vol. 2, 1103-1106, 2001.

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

[3] ITU-T Rec. G.107 : The E-Model, a Computational Model for Use in Transmission Planning. International Telecommunication Union, CH-Geneva, 2002.

[4] ITU-T Rec. P.851: Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, CH-Geneva, 2003.

[5] Metze, F., McDonough, J., Soltau, H.: Speech Recognition over NetMeeting Connections. In: Proc. Eurospeech 2001 – Scandinavia, Vol. 4, 2389-2392, 2001.

[6] Möller, S.: Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic Publ., USA-Boston, 2000.

[7] Möller, S.: Telephone Transmission Impact on Synthesized Speech: Quality Assessment and Prediction. Acta Acustica united with Acustica 90, 121-136, 2004.

[8] NIST Net, Network Emulation Software, v. 2.0. National Institute of Standards and Technology, `www.nist.gov`, 2001.

[9] Raake, A.: Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with Other Network Impairments. Acta Acustica united with Acustica, to appear, 2004.

[10] Rehmann, S., Raake, S., Möller, S.: Parametric Simulation of Impairments Caused by Telephone and Voice over IP Network Transmission. In: Proc. 3rd European Congress on Acoustics (Forum Acusticum Sevilla 2002), ES-Sevilla, Special Issue Revista de Acústica 33, 6 pages, 2002.