AUDITIVE UND INSTRUMENTELLE BEURTEILUNG DER QUALITÄT SYNTHETISCHER SPRACHE IN UNTERSCHIEDLICHEN ANWENDUNGSZUSAMMENHÄNGEN

Sebastian Möller und Florian Hinterleitner

Quality and Usability Lab, Telekom Innovation Labs, TU Berlin
sebastian.moeller@telekom.de

Abstract: In diesem Aufsatz beleuchten wir Methoden zur Beurteilung der Qualität synthetischer Sprache. Zunächst werden psychophysikalische Merkmale analysiert, welche der menschlichen Wahrnehmung und Qualitätsbeurteilung zugrunde liegen. Darauf aufbauend werden Testmethoden und –ergebnisse beschrieben, die diese Merkmale in unterschiedlichen Anwendungsszenarien auditiv quantifizieren. Abschließend wird ein Überblick auf laufende Arbeiten gegeben, welche versuchen, auditive Testergebnisse zu einem gewissen Maße instrumentell zu schätzen. Die Nutzbarmachung solcher instrumenteller Verfahren für die Syntheseentwicklung und –optimierung wird diskutiert.

1 Motivation

"If you cannot measure it, you cannot improve it." Dieses Zitat, welches Lord Kelvin zugeschrieben wird, bezog sich eigentlich auf physikalische Sachverhalte, wie aus folgenden Erläuterungen in seinem Werk „Popular Lectures and Addresses" (1891-1894) hervorgeht:

> „In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."[1]

Wenn man den Begriff der Messung auch auf psychophysikalische Sachverhalte erweitert, so weist das Zitat auf die zentrale wissenschaftliche Bedeutung der Messung für die wissenschaftliche Forschung wie auch die praktische Nutzbarmachung der Forschungsergebnisse hin. Dies gilt in besonderem Maße für die maschinelle Sprachsynthese, bei der weiterhin grundlegende Mechanismen unbekannt und – teilweise als Folge davon – Anwendungen von nicht ausreichender Qualität sind, um den Massenmarkt zu bedienen. Rüdiger Hoffmann trug nicht unwesentlich dazu bei, diesem Umstand Abhilfe zu verschaffen.

Eine moderne Definition sieht Qualität als einen Abgleich von wahrgenommenen und erwarteten oder erwünschten Eigenschaften einer Einheit [2]. Das Wahrgenommene wird häufig durch ein physikalisches Ereignis (z.B. die Schallwelle, die von einer Sprachsynthese über einen Lautsprecher ausgegeben wird) getriggert; es „geschieht" also gewissermaßen, und ist daher durch den Ort, die Zeit und die Eigenschaften des Ereignisses gekennzeichnet. Wird ein solches Ereignis mit dem Erwarteten oder Erwünschten verglichen, entsteht daraus wiederum ein Ereignis – hier ein Qualitätsereignis, welches bislang nur durch Introspektion der Versuchsperson erfasst werden kann (zu zukünftigen alternativen Methoden vgl. den Ausblick dieses Abschnitts).

Zur qualitativen und quantitativen Erfassung der Qualität synthetisierter Sprache bedient man sich normalerweise psychoakustischer Experimente mit menschlichen Versuchspersonen. Dabei werden vorher synthetisierte Sprachproben bezüglich ihrer relevanten Qualitätsmerkmale „vermessen“ oder vorher bekannte Qualitätsmerkmale „gemessen“, d.h. quantifiziert. Die Messung geschieht durch menschliche Versuchspersonen, welche ihre Wahrnehmungsereignisse nach vorher definierten Regeln mit Erwartungen und Wünschen abgleichen.

Wichtig ist in diesem Zusammenhang der Begriff der Relevanz: Wenn man davon ausgeht dass die Wahrnehmungsmerkmale konstant sind, so sollten sie sich eigentlich vom Anwendungszusammenhang unabhängig für alle Wahrnehmungsereignisse eines bestimmten Typs (z.B. die Wahrnehmungen synthetischer Sprache) gleichermaßen quantifizieren lassen. Ergebnisse psychophysikalischer Untersuchungen, wie sie in Abschnitt 2 vorgestellt werden, zeigen jedoch, dass je nach Anwendung unterschiedliche Merkmale von den Versuchspersonen genannt und beurteilt werden. Diesem Umstand trägt die Referenz (d.h. das Erwartete oder Erwünschte) Rechnung, die einzelne Wahrnehmungsmerkmale verstärkt oder abschwächt – ähnlich einem Filter für Wahrnehmungsmerkmale.

Im folgenden Abschnitt 2 stellen wir Ergebnisse psychoakustischer Studien vor, die die wahrgenommenen Merkmale synthetischer Sprache zu ergründen suchen. Abschnitt 3 beschreibt experimentelle Untersuchungen, in denen diese Merkmale in unterschiedlichen Anwendungszusammenhängen quantitativ erfasst wurden. Abschnitt 4 erläutert das Prinzip der Qualitätsschätzung und zeigt, dass auch mittels instrumenteller Messungen Aussagen zu einzelnen Qualitätsmerkmalen getroffen werden können. Abschnitt 5 beschließt den Aufsatz mit einem kritischen Rückblick und Ausblick auf offene Fragen.

2 Wahrnehmungsdimensionen synthetischer Sprache

Obwohl moderne Text-To-Speech (TTS) Systeme inzwischen ein Qualitätsniveau erreicht haben, das sie geeignet für eine Vielzahl unterschiedlicher Anwendungen macht, so z.B. Email-Vorlesedienste und Telefoninformationssysteme, werden einzelne Qualitätsaspekte synthetischer Sprache immer noch durch unterschiedliche systemabhängige Störungen beeinflusst: Diphone-Synthesizer hören sich wegen der zahlreichen Verkettungen von Spracheinheiten meistens sehr künstlich an, HMM-Synthesizer können zwar zu sehr natürlich klingender, aber gleichzeitig auch stark verrauschter Sprache führen, und die Qualität von Unit-Selection-Systemen hängt nicht nur davon ab, wie gut die einzelnen Einheiten zueinander passen, sondern auch davon, wie gut sie zu dem zu synthetisierenden Text passen.

Diese Störungen klingen allesamt unterschiedlich, d.h. sie stören die wahrgenommene Qualität somit entlang unterschiedlicher Wahrnehmungsdimensionen. Die Qualität synthetischer Sprache muss also als multidimensional angesehen werden.

Im Folgenden stellen wir die Ergebnisse mehrerer Studien vor, in denen die synthetischer Sprache zugrundeliegenden perzeptiven Qualitätsdimensionen untersucht wurden.

In einer ersten Versuchsreihe [4] wurden die Qualitätseindrücke der Versuchspersonen mittels eines semantischen Differentials (SD) erfasst. Dabei bewerten die Probanden vorgespielte Stimuli, wie sie für Telekommunikationsdienste typisch sind, auf vorgefertigten Attributskalen. Um zu garantieren, dass in diesem Versuch auch alle perzeptiven Qualitätseindrücke erfasst werden, müssen die Attributskalen durch aufwendige Vorversuche entwickelt werden. In einem ersten Schritt wurde eine Testdatenbank erstellt. Um eine möglichst große Anzahl unterschiedlicher Störungen zu erfassen wurden synthetische Sprachdaten von 14/15 unterschiedlichen weiblichen/männlichen TTS-Systemen erstellt. Einige von ihnen lagen mit bis zu 6 unterschiedlichen Stimmen vor.

In einem ersten Vorversuch wurden 296 unterschiedliche qualitätsbeschreibende Attribute gesammelt, die in 44 Attributskalen komprimiert wurden. Ein zweiter Vorversuch diente dazu, aus diesem Satz an Skalen die relevantesten herauszufiltern. Diese dienten im eigentlichen Hauptversuch dazu, die Qualität der präsentierten Stimuli zu erfassen.

Mittels einer Hauptachsen-Faktorenanalyse und einer daran anschließenden Promax-Rotation ließen sich 4 Faktoren extrahieren. Diese repräsentieren die *Natürlichkeit, Störungen, zeitliche Verzerrungen* sowie die *Geschwindigkeit.*

In einer zweiten Studie [5] sollten diese Ergebnisse durch einen Versuch mit multidimensionaler Skalierung (MDS) bestätigt werden. Während die Bewertungen der Probanden in einem semantischen Differential immer auf die vorgegebenen Skalen beschränkt sind, umgeht die multidimensionale Skalierung diese Einschränkung, indem die Versuchspersonen nicht die einzelnen Stimuli bewerten, sondern die Ähnlichkeit je zweier paarweise präsentierter Stimuli. Da die Anzahl solcher Paarvergleiche selbst bei wenig umfangreichen Datenbanken schnell Werte größer 100 annimmt und jeder Proband somit mehrere Stunden beschäftigt wäre, wurde in diesem Versuch statt auf Paarvergleiche auf eine Sortierungsaufgabe zurückgegriffen. Dabei sollten die Versuchspersonen die präsentierten Stimuli in mehrere Gruppen einteilen mit der Vorgabe, dass sich Stimuli einer Gruppe auditiv ähnlich sind während sie sich von Stimuli in anderen Gruppen unterscheiden. Aus den Ergebnissen aller Probanden wurde eine Unähnlichkeitsmatrix erstellt, die mittels eines MDS-Algorithmus ausgewertet wurde. Bei diesem Versuch ergaben sich drei Dimensionen, die sich der *Natürlichkeit der Stimme*, *zeitlichen Verzerrungen* und der *Gelassenheit des Sprechers* zuordnen lassen.

Damit ergeben sich in den Ergebnissen der beiden Versuche einige Übereinstimmungen, aber auch Unterschiede:

- Die sehr breit angelegte Dimension Natürlichkeit aus dem SD-Experiment findet sich in der Dimension Natürlichkeit der Stimme sowie in prosodischen Teil der Dimension zeitliche Verzerrungen des MDS-Versuchs wieder.
- Die Dimensionen zeitliche Störungen repräsentieren zeitliche Ansätze, wie sie beim Verknüpfen von Sprachbausteinen in einer Unit-Selection-Synthese vorkommen. Der Hörer hat dann, sofern die Spracheinheiten nicht genau zueinander passen, den Eindruck, als würden zwei Sprecher gleichzeitig reden.
- Die Dimension Geschwindigkeit und Gelassenheit des Sprechers repräsentieren beide Ruhe/Trägheit bzw. Unruhe/Hektik des Sprechers.
- Die Dimension Störungen aus dem ersten Versuch, die z.B. Rauschen, Zischen, Kratzen usw. wiedergibt, konnte im zweiten Versuch nicht bestätigt werden.

Die in den vorangegangenen Abschnitten untersuchten Stimuli wiesen alle eine Dauer von ca. 5-10s auf und bezogen sich somit auf Use-Cases wie z.B. Smart-Home-Umgebungen oder Telefon-Informationssysteme. Im nun folgenden Abschnitt werden die Ergebnisse zweier Studien zu Qualitätsdimensionen synthetisch erzeugter Hörbüchern vorgestellt. In [6] wurde ein Evaluationsverfahren zu Bestimmung der Qualität synthetisch erzeugter Hörbücher entwickelt. Neben bereits standardisierten Skalen [7] wurden einige Bewertungsskalen speziell für diese Aufgabe entwickelt. Um ein möglichst großes Spektrum unterschiedlicher Text abzudecken wurden Bücher aus mehreren Bereichen (Science Fiction, Komödie, Thriller, Kinderbuch usw.) und mit verschiedenen Schreibstilen (direkte/indirekte Sprache, gehobenes Vokabular, kurze/lange Sätze, komplexer Satzbau, poetische Sprache usw.) ausgewählt. Die von zwei Sprachsynthesen mit jeweils weiblicher und männlicher Stimme erstellten Stimuli hatten eine durchschnittliche Länge von ca. 50s. Die Stimuli wurden in einem Hörversuch auf den vorliegenden 8 Skalen bewertet und die Ergebnisse im Anschluss

mittels einer Hauptachsen-Faktorenanalyse ausgewertet. Dabei ergaben sich zwei Qualitätsdimensionen, die sich dem *Hörgenuss* und der *Prosodie* zuordnen lassen.

Nach einer abschließenden Verbesserung am Messprotokoll kam das entwickelte Verfahren in der Blizzard Challenge [8] zum Einsatz. Diese ist ein jährlich stattfindender Wettbewerb für Entwickler von Sprachsynthesesystemen, in dem im Jahr 2012 zum ersten Mal auch synthetisch erzeugte Hörbücher evaluiert wurden. In dieser groß angelegten Studie wurden unter Anderem 230 TTS-Hörbuchstimuli auf 7 der im vorgehenden Abschnitt präsentierten Skalen bewertet. Aus einer Faktorenanalyse gingen auch hier zwei Faktoren hervor, die sich ebenfalls dem *Hörgenuss* und der *Prosodie* zuordnen lassen.

Im Vergleich zu den Dimensionen der anfangs beschriebenen Use-Cases zeigt sich, dass die Dimension Hörgenuss der Dimension Natürlichkeit der Stimme zugeordnet werden kann, während die Dimension Prosodie die Rhythmik aus den Dimensionen Natürlichkeit (SD) und zeitliche Störungen (MDS) wiedergibt.

3 Auditive Beurteilung bei unterschiedlichen Anwendungen

Exemplarisch sollen in diesem Abschnitt experimentelle Untersuchungen zur Quantifizierung der Qualität synthetischer Sprache in einem prototypischen System vorgestellt werden. Hierbei handelt es sich um das Infotainment-System INSPIRE, welches im Rahmen des gleichnamigen EU-geförderten IST-Projektes erstellt wurde und die Bedienung verschiedener Geräte in einem intelligenten Haus ermöglicht, wie den Fernseher, den Videorekorder, den Anrufbeantworter, die Leuchten und die Rollos. Einige Funktionen (elektronische Programminformationen, Videorekorder-Programmierung und Anrufbeantworter) sind auch über einen Telefonzugang erreichbar, und insbesondere in diesem Fall kommt der synthetischen Sprache besondere Bedeutung zu.

Ziel der Evaluierung war es deshalb, verschiedene Versionen des Systems mit verschiedenen Synthesesystemen auf ihre Gebrauchstauglichkeit für die geplante Anwendung zu untersuchen. Hierzu wurden 3 Experimente in unterschiedlichen Anwendungsszenarien durchgeführt:

- *Hausumgebung*: Der Fokus der Evaluation lag dabei auf dem Erscheinungsbild, welches der Nutzer vom System bekommt (sog. Metaphor), sowie auf dem Einfluss der Raumakustik bei der Sprachwiedergabe. Dabei konnte INSPIRE entweder als eine Ansammlung von „intelligenten Geräten“ agieren, die unabhängig voneinander eine sprachliche Interaktion mit dem Nutzer führen konnten, oder als „Assistent“, der auf einem Bildschirm in Form eines Avatars sichtbar war, oder als unsichtbarer Assistent (sog. „Geist“).
- *Telefonbedienung aus dem Büro*: Hier wurde der Effekt unterschiedlicher Übertragungskanäle, Endgeräte sowie möglicher Hintergrundgeräusche getestet.
- *Telefonbedienung aus dem fahrenden Auto*: Hier wurde die Robustheit der Sprachausgabe gegen Hintergrundgeräusche sowie bei kognitiver Belastung (Ablenkung durch die Fahraufgabe) des Nutzers gemessen.

Für das erste Szenario wurde das System in einer simulierten Wohnumgebung am IKA der Ruhr-Universität Bochum aufgebaut. Dabei hörten Testnutzer zum einen die Sprachausgabe des Systems vom Sofa aus und realisierten zum anderen die Reaktion der adressierten Hausgeräte. Nach jeder Sprachausgabe wurde zunächst der Inhalt des Gesagten abgefragt; hierdurch sollte erreicht werden, dass sich die Probanden auf den sprachlichen Inhalt konzentrierten (wie dies in einer realen Situation auch der Fall wäre), und nicht allein auf die Oberflächenform. Danach beurteilten sie 4 Merkmale der Sprachausgabe (Gesamtqualität,

Höranstrengung, Annehmlichkeit der Stimme, Passgenauigkeit der Stimme zum System) auf kontinuierlichen Skalen, welche durch Attribute beschriftet waren.

Im zweiten Szenario fehlte das direkte Feedback der Hausgeräte, dafür wurden Telefonverbindungen mit unterschiedlichen Übertragungseigenschaften (Leitungsrauschen, Kodierer mit und ohne Paketverluste) simuliert sowie Endgeräte mit unterschiedlichen akustischen Eigenschaften (Handapparat, realer und idealer Freisprecher), z.T. bei Hintergrundgeräusch verwendet. Im dritten Szenario wurden in einem Fahrsimulator ebenfalls unterschiedliche Übertragungseigenschaften simuliert, allerdings wurde hier nur der Freisprecher als realistisches Endgerät betrachtet; außerdem hatten die Probanden dort eine Fahraufgabe, und es wurde nur die (wahrgenommene) Verständlichkeit sowie die Gesamtqualität abgefragt, welche im Fahrsimulator auf einem Bildschirm bewertet werden sollten. Die Sprachprompts bestanden in allen Fällen aus Sätzen oder Doppelsätzen, welche sowohl statischen als auch dynamisch generierten Inhalte enthielten. Für die statischen Inhalte stand entweder ein natürlicher Sprecher oder verschiedene Sprachsynthesen zur Verfügung, für die dynamischen Inhalte nur die Sprachsynthesen; somit enthielten die Sprachprompts entweder nur synthetische oder auch gemischt natürlich-synthetische Stimmen.

Detaillierte Informationen zum Versuchsaufbau, den Sprachstimuli, den Bewertungsverfahren sowie den Ergebnissen finden sich in [3]. Im Folgenden sollen nur einige Kern-Ergebnisse aufgelistet werden, die veranschaulichen, dass die Qualitätsmesswerte – selbst bei einem einzigen getesteten System und einer einzigen synthetischen Stimme – in starkem Maße vom Anwendungsszenario bestimmt werden:

- In der *Hausumgebung* zeigten sich deutliche Unterschiede zwischen den Qualitätsurteilen für unterschiedliche natürliche und synthetische Stimmen. Allerdings kam die beste synthetische Stimme manchmal einer natürlichen bezüglich der Gesamtqualität gleich. Bei den Metaphoren „Assistent“ und „Geist“ war die Kombination von natürlicher und synthetischer Stimme am vorteilhaftesten. Je nach Metaphor schnitten die Stimmen unterschiedlich ab: Die natürlichen Stimme wurde beim „Assistenten“ besser beurteilt als bei den „intelligenten Geräte“, bei der synthetischen Stimme war es umgekehrt.
- In der *Büroumgebung* zeigte sich ein starker Einfluss des Übertragungskanals auf die Qualitätsurteile. Quantitativ war der Einfluss bei synthetischer Sprache ungefähr mit dem bei natürlicher Sprache vergleichbar, allerdings schien die synthetische Stimme robuster bei Leitungsrauschen. Hintergrundgeräusche am Empfangsort beeinträchtigten insbesondere die Höranstrengung.
- Im *Auto* wurde ebenfalls die natürliche bzw. die Kombination von natürlicher und synthetischer Stimme bevorzugt. Die Übertragungskanäle zeigten wiederum einen deutlichen Einfluss, allerdings war dieser für synthetische Sprache bei Sprachkodierern etwas geringer und bei Paketverlusten etwas stärker ausgeprägt als bei natürlicher Sprache.

Insgesamt kann das Testprotokoll als ein gutes Beispiel für die anwendungsorientierte Evaluierung synthetischer Sprache dienen. Es zeigt, dass das Anwendungsszenario einen signifikanten Einfluss auf die wahrgenommene Qualität hat, und dass vielfach einzelne Einflussfaktoren nur mit einem speziell gestalteten Testprotokoll adäquat berücksichtigt werden können. Dies ist für die Validität der Messergebnisse von großer Bedeutung.

4 Instrumentelle Schätzung

Auditive Bewertungen, wie sie in dem vorangegangenen Abschnitt vorgestellt wurden, stellen immer noch die einzige verlässliche Methode zur Qualitätsbestimmung von Sprachsynthesen

dar. Da TTS-Entwickler ihre Systeme innerhalb eines Entwicklungszyklus jedoch des Öfteren evaluieren müssen, um sowohl Algorithmus als auch Korpus zu optimieren, sind zeit- und kostenintensive Nutzerstudien ein ineffizientes Mittel. Um den Aufwand für Hörversuche auf ein Minimum zu reduzieren ist ein rein instrumentelles Verfahren zur Vorhersage der Qualität wünschenswert.

In den vergangenen Jahren wurden einige Verfahren vorgestellt, mit denen sich erste Ergebnisse im Bereich der Qualitätsvorhersage für synthetische Sprache erzielen ließen. Im Folgenden wird ein einführender Überblick über mögliche Ansätze gegeben.

Bereits in [9] wurden unterschiedliche Verfahren miteinander verglichen.

- *HMM-basierter Merkmalsvergleich*

  Im Vergleich zu natürlicher Sprache können in synthetischer Sprache durch die Verknüpfung einzelner Sprachbausteine zu einem Sprachsignal spontane spektrale Veränderungen auftreten, die durch die Trägheit des Vokaltrakts bei Menschen sonst nicht gegeben sind. Diese Störungen dienen als Hinweis auf die Qualität des erzeugten Sprachsignals. Im vorliegenden Ansatz werden deshalb Hidden-Markov-Modelle (HMM) auf natürlich erzeugte Sprache trainiert. Die spektralen vom HMM erfassten Informationen dienen dann dazu, Unterschiede z.B. zwischen natürlichen Wortendungen und durch Sprachsynthese erzeugten Signalunterbrechungen zu quantifizieren.

  Als Basis dienen melskalierte cepstrale Koeffizienten (MFCC) 12. Ordnung. Diese wurden zunächst aus einer Trainingsdatenbank mit natürlichen Sprechern extrahiert und anschließend zum Training des HMMs verwendet. In einem zweiten Schritt lassen sich dieselben Sprachmerkmale aus den zu evaluierenden TTS-Daten extrahieren.

  Über den Forward-Backward-Algorithmus kann nun ein Log-Likelihood-Wert bestimmt werden, der die perzeptive Ähnlichkeit zwischen einem TTS-Signal und dem auf natürliche Sprache trainierten HMM ausdrückt.

- *Linearkombination extrahierter Merkmale*

  Im zweiten Ansatz werden Parameter aus dem TTS-Signal extrahiert, die mit den durch den Syntheseprozess induzierten Störungen in Zusammenhang stehen können. Als Basis dienten zum einen die 44 internen Parameter des in der ITU-T-Empfehlung P.563 [10] beschriebenen Algorithmus. Dieser erlaubt eine zuverlässige Qualitätsschätzung von über einen Telefonkanal übertragener Sprache. Da diese Parameter unter Anderem Signalcharakteristiken wie z.B. Rauschen, zeitliches Clipping und Robotereffekte (metallisch klingende Stimmen) ausdrücken, erschienen sie auch für die durch Sprachsynthese erzeugten Artefakte von Bedeutung. Des Weiteren stand ein Set an Sprachmerkmalen zur Verfügung, das umfangreiche Informationen über den Vokaltrakt liefert und zur Klassifikation von menschlichen Emotionen eingesetzt wird. Enthalten sind unter anderem Informationen zu: Grundfrequenzverlauf, Lautheit, MFCCs, spektralen Komponenten, Formanten und der Intensität. Dadurch ergaben sich insgesamt 1495 Sprachmerkmale pro TTS-Signal.

  Um insbesondere aus dem zweiten sehr umfangreichen Satz an Merkmalen die für die Qualitätsschätzung bedeutsamsten Informationen zu ermitteln wurde eine korrelationsbasierte sequentielle Merkmalsauswahl durchgeführt. Aus der Untermenge an Merkmalen ließ sich mittels einer Hauptkomponentenanalyse ein kleiner Satz relevanter Faktoren ermitteln, die über eine lineare Regressionsanalyse zu einem Qualitätsschätzwert kombiniert wurden.

Neben diesen drei ermittelten Schätzwerten wurden alle Schätzer durch eine weitere lineare Regression miteinander kombiniert. Für die meisten der vorhandenen evaluierten TTS-Datenbanken konnten dadurch sehr gute Vorhersagegenauigkeiten erzielt werden, während die Ergebnisse auf einigen Datenbanken weiter optimiert werden müssen. Dies trifft insbesondere für die weiblichen TTS-Stimuli zu. Die Kombination mehrerer Schätzer führte durchweg zu deutlich stabileren Vorhersagen.

5 Ausblick

Wir haben in den vorangegangenen Abschnitten Methoden zur auditiven Messung und zur instrumentellen Schätzung der Qualität synthetisierter Sprache vorgestellt. Dabei diente die auditive Messung stets als Mittel der Wahl, lassen sich doch nur damit valide und zuverlässige Qualitätsmesswerte erzielen. Allerdings zeigte sich, dass die von Versuchspersonen gelieferten Urteile stark von der Anwendung abhängig sind; Merkmale – oder perzeptive Dimensionen – welche in einem Anwendungszusammenhang als relevant erachtet werden, sind dies nicht unbedingt in einem anderen Zusammenhang.

Die Abhängigkeit der auditiven Qualitätsmessung vom Anwendungszusammenhang erschwert auch eine instrumentelle Schätzung, denn hierbei müsste der Anwendungszusammenhang ja ebenfalls berücksichtigt werden. So lassen sich derzeit zwar für einzelne Datenbanken mit den o.a. Verfahren zufriedenstellende Korrelationen zwischen auditiven Urteilen und instrumentellen Schätzungen herstellen; es bleibt aber noch zu zeigen, in wie weit sich solche Schätzungen auch auf andere Datenbanken, die womöglich andere Anwendungszusammenhänge widerspiegeln, übertragen lassen.

Wir gehen davon aus, dass eine solche instrumentelle Schätzung sich einfacher für einzelne Qualitätsdimensionen als für die Gesamtqualität bewerkstelligen lässt. Dies gilt unter der Annahme, dass der perzeptive (Qualitäts-)Raum von synthetischer Sprache grundsätzlich feststeht, und dass einzelne Dimensionen dieses Raumes je nach Anwendungsszenario unterschiedlich gewichtet werden müssen. Dieser Annahme entspricht die Idee der in Abschnitt 1 genannten „Filterung“ der Wahrnehmungsmerkmale durch die Referenz. Im Ergebnis könnte eine Qualitätsschätzung also in Form eines multidimensionalen Profils erfolgen. Dieser Ansatz wird derzeit in einem gemeinsam mit der Christian-Albrechts-Universität zu Kiel bearbeiteten DFG-Projekt verfolgt (DFG MO 1038/11-2).

Ein Profil von Qualitätsschätzern wäre für die Entwickler von TTS-Systemen von großer Bedeutung. So könnten Entwickler einzelne Parameter ihres Systems (oder den verwendeten Korpus) verändern und direkt (ohne Zeitverlust für die Durchführung eines auditiven Tests) Feedback zur erzielten Qualitätsverbesserung oder –verschlechterung erhalten. Mehr noch, das Profil von Qualitätsschätzern, welche je eine perzeptive Dimension vorhersagen, würde Aufschluss über verbleibende Probleme – und somit Hinweise auf mögliche Lösungen – liefern. Dadurch könnte die Entwicklung besserer Sprachsynthesesysteme stark beschleunigt werden.

Wie oben beschrieben beruhen Messungen von Qualität auf bewussten Urteilen von Probanden eines auditiven Tests. Dies ist notwendig, da das Qualitätsereignis (vgl. Abschnitt 1) innerhalb des Probanden liegt und damit unzugänglich erscheint. Neuere Untersuchungen belegen aber, dass sich durch physiologische Messungen Rückschlüsse auf die Wahrnehmung von Qualität ziehen lassen [11]. So wurde bspw. in [12] gezeigt, dass sich mittels Elektroenzephalographie (EEG) auch nicht bewusst wahrgenommene Störungen von Sprachsignalen messen lassen. Bei der Analyse ereigniskorrelierter Potenziale im EEG zeigte sich, dass vor allem die Amplitude der Reaktion des Gehirns ca. 300 ms nach dem Einsetzen eines Triggers bei gestörten Stimuli größer ist, und dass die Reaktion bei starken Störungen etwas früher eintritt. Die verringerte Latenz kann als geringerer „neuronaler Effort“ bei der

Störungsentdeckung verstanden werden. Am Quality and Usability Lab der TU Berlin wird derzeit in Zusammenarbeit mit dem INRS in Montreal untersucht, ob sich diese Methode auch für synthetisierte Sprache einsetzen lässt [13].

Danksagung

Dieser Beitrag wurde im Rahmen des Projektes „Instrumentelle Schätzung der Qualität synthetisierter Sprachsignale“ (MO 1038/11-2) von der DFG gefördert.

Literatur

[1] Thomson, W. Lecture on “Electrical Units of Measurement” (3 May 1883). Published in Popular Lectures Vol. I, p. 73. Quoted in Encyclopaedia of Occupational Health and Safety (1998) by Jeanne Mager Stellman, p. 1992.

[2] Jekosch, U. Voice and Speech Quality Perception — Assessment and Evaluation. Signals and Communication Technology Series, Springer, Berlin, 2005.

[3] Möller, S., Krebber, J., Smeele, P. Evaluating the Speech Output Component of a Smart-Home System, Speech Communication 48, 1-27, 2005.

[4] Hinterleitner, F., Möller, S., Norrenbrock, C., Heute, U. Perceptual Quality Dimensions of Text-To-Speech Systems, Proceedings Interspeech 2011, Florenz, S. 2177-2180, 2011.

[5] Hinterleitner, F., Norrenbrock, C., Möller, S., Heute, U. What makes this voice sound so bad? A multidimensional analysis of state of the art text-to-speech systems, Proceedings of the Spoken Language Technology Workshop (SLT) 2012, Miami, S. 240-245, 2012.

[6] Hinterleitner, F., Neitzel, G., Möller, S., Norrenbrock, C. An Evaluation Protocol for the Subjective Assessment of Text-To-Speech in Audiobook Reading Tasks, Proceedings of the Blizzard Challenge Workshop, Florenz, 2011.

[7] ITU-T Rec. P.85, A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union, Genf, 1994.

[8] Hinterleitner, F., Norrenbrock, C., Möller, S. Perceptual Quality Dimensions of Text-To-Speech Systems in Audiobook Reading Tasks, Proceedings of the ESSV 2012, Bielefeld, 2013.

[9] Möller, S., Hinterleitner, F., Falk, T., Polzehl, T. Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems, Proceedings Interspeech 2010, Kyoto, S. 1325-1328, 2010.

[10] ITU-T Rec. P.563, Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications, International Telecommunication Union, Genf, 2004.

[11] Möller, S., Antons, J.-N., Arndt, S., Porbadnigk, A., Schleicher, R. Zum Einsatz von Elektroenzephalographie bei der Messung der Wahrnehmung gestörter Sprache, in: Elektronische Sprachsignalverarbeitung 2012. Tagungsband der 23. Konferenz, Cottbus, 29.-31. Aug., M. Wolff (Hrsg.), TUDpress, Dresden, S. 81-88, 2012.

[12] Antons, J.N., Blankertz, B., Curio, G., Möller, S., Porbadnigk, A.K., Schleicher, R. Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise, in: Audio Engineering Society (AES) 129th Convention, 2010.

[13] Laghari, K.u.R., Gupta, R., Arndt, S., Antons, J.-N., Schleicher, R., Möller, S., Falk, T.H. Auditory BCIs for Visually Impaired Users: Should Developers Worry About the Quality of Text-to-Speech Readers?, angenommen zur Veröffentlichung: International BCI Meeting (Brain-Computer Interface 2013), Asilomar, Pacific Grove CA, 3-7 June 2013.