SPRACHSYNTHESESYSTEME FÜR DEN EINSATZ IM KFZ – UNTERSUCHUNGEN ZUR SPRACHQUALITÄT

Nadya Stoyanova, Steffen Werner

Daimler AG – Mercedes-Benz Cars Development
{nadya_emilova.stoyanova | steffen.s.werner}@daimler.com

Abstract:

Bei der Integration von Sprachausgaben in Telematik-Geräten müssen Herausforderungen wie geringer Speicherplatz und geringere und/oder langsamere Rechenleistung gemeistert werden. Gleichzeitig erwartet der Nutzer eine höhere Sprachqualität von integrierten Lösungen, als von externen Geräten, die oft durch die Verwendung natürlicher Sprache realisiert wird.

Die bisher überwiegend statischen Informationen wie Navigations-Fahrempfehlungen werden in modernen Systemen durch dynamische Inhalte wie Straßennamen ergänzt. Da es speichertechnisch nicht sinnvoll ist, alle möglichen Straßennamen mit natürlicher Sprache aufzuzeichnen, werden immer häufiger Synthesesysteme auch in hochwertigen Telematik-Geräten integriert.

Der inhaltliche Fokus variiert von kurzen Äußerungen in der Navigation bis hin zu längeren Sätzen in SMS und/oder Email. Deshalb ist eine Systemauswahl über etablierte Challenges nur begrenzt möglich. Im vorliegenden Beitrag werden die Ergebnisse einer Untersuchung dargestellt, die aktuelle automotive-taugliche Systeme vergleicht.

1 Einleitung und Motivation

Der Fokus von Optimierungen bei Einsatz von Sprache in technischen Systemen liegt größtenteils im Bereich der Spracherkennung. Dies ist durchaus verständlich, da eine falsche Erkennung oftmals zu falschen Dialogabläufen führt, welche sogar als anekdotische Vorlagen für die Werbung dienen [1]. Dabei ließen sich solche Fehler mit einem gut durchdachten Dialogkonzept durchaus minimieren und/oder vermeiden.

Fast schon stiefmütterlich wird dagegen die Optimierung der Sprachausgaben gehandhabt. Dies liegt zum einen daran, dass häufig aufgenommene Sprache für die verschiedenen Dialogpromts verwendet wird. Zum anderen bekommt man als Entscheider bei der Auswahl eines Synthesesystems oft gut gestylte Hörbeispiele präsentiert, die kaum Handlungsbedarf erkennen lassen [2]. Die Qualitätsbeurteilung von Sprache ist jedoch nichts Absolutes, da sich die individuelle Hörererwartung an die Sprachqualität dynamisch ändern kann [3].

Bei der Auswahl von Sprachausgaben für Entwicklung von Telematik-Geräten müssen diverse Herausforderungen gemeistert werden. Sprachausgaben werden von Fahrern unmittelbar wahrgenommen und führen sofort zu einem Qualitätseindruck und damit auch zu einer Markendifferenzierung. Dabei ist die Nutzer-Erwartung an die Sprachqualität bei integrierten Lösungen wesentlich höher, als bei Aftermarket-Lösungen.

Der vorliegende Beitrag zeigt die Randbedingungen auf, die beim Einsatz von Sprachausgaben (synthetisch und/oder natürlich) beachtet werden müssen. Desweiteren werden die Ergebnisse einer Qualitätsuntersuchung aktueller automotive-tauglicher Sprachsynthesesysteme

vorgestellt, welche in Zusammenarbeit mit den führenden Anbietern von TTS-Systemen durchgeführt wurde.

Der Einsatz von Sprachsynthese bei hochwertigen Sprach-Dialog-Systemen und/oder Navigationssystemen im Automotive-Bereich erfordert den Aufbau einer geeigneten Teststrategie. Nur so lassen sich bestimmte realisierte Sprachqualitäten des Synthesesystems mit den Anforderungen und Erwartungen vergleichen und bewerten. Möglichkeiten dazu werden im Beitrag ebenfalls andiskutiert.

2 Sprachausgabesysteme für Telematik-Geräte im Kraftfahrzeug

In diesem Abschnitt werden einige der Randbedingungen für den Einsatz von Telematik-Geräten im KFZ kurz vorgestellt und auf die Bandbreite an Anforderungen und Qualitätsmaßstäbe eingegangen.

2.1 Sprachausgaben im KFZ

Moderne Telematik-Geräte haben mittlerweile einen wertbestimmenden und markenspezifischen Rang im Automobil erobert. Dabei werden immer mehr Aufgaben und Funktionen in diese Geräte integriert. Das Aufgabenspektrum reicht dabei vom Wiedergeben verschiedenster Inhalte (gespeichert, Broadcast, On-Demand, etc.) über die Bereitstellung von unterschiedlichen Informationen (zum Fahrzeug, zur Position, zum Umfeld, etc.) bis hin zur Kommunikation mit anderen Menschen, Fahrzeugen, und/oder Diensteanbietern. Auch immer mehr Fahrer-Assistenzfunktionen werden mit der Telematik verknüpft, sei es zur Zustandserkennung, über die Fahrerwarnung bis hin zur reinen Informationsbereitstellung [4].

Viele der einzelnen Aufgaben sind mit Sprachausgaben gekoppelt, wie z. B. die Navigation, Verkehrsinformation, Messaging, usw. Qualitativ hochwertige Sprachausgaben bei Telematik-Geräten im Fahrzeug werden meist immer noch mit natürlicher Sprache realisiert. Die Zunahme der Ausgaben mit variierenden und dynamischen Inhalten (wie z. B. Verkehrsmeldungen und SMS) verlangt eine Dynamisierung der Sprachausgaben, weg von Voraufgenommenen, so genannten Pre-Recordings, hin zu freier Sprachgestaltung mittels Text-To-Speech Systemen.

Durch Vergrößerung von Funktionsumfängen kommt es dabei zunehmend zur Vermischung von immer wiederkehrenden (statischen) Sprachausgaben (z. B. Fahranweisungen innerhalb der Navigation) und von wechselnden (dynamischen) Sprachausgaben (z. B. Fahrtrichtungsansagen innerhalb der Fahranweisungen). Dadurch gewinnt der Aspekt der Sprachqualitätsbewertung von synthetischen Äußerungen im Fahrzeug immer mehr an Bedeutung, da immer die natürlichsprachliche Referenz der Maßstab ist.

2.2 Randbedingungen und Sprachqualität

Nur wenige Hersteller integrierter Telematik-Lösungen wagten bisher den Weg, Sprachsynthese exklusiv für alle Sprachausgaben zu verwenden. Die Unterschiede zwischen den Klangbildern von natürlicher und synthetischer Sprache wirken zwar störend, aber auch die Sprachqualität der Sprachsynthese ist immer noch eingeschränkt, da Ressourcen oft knapp bemessen sind.

Da Sprache unser natürliches Kommunikationsmittel ist, werden sprachliche Ausgaben im Fahrzeug aber unmittelbar mit natürlichen Aussagen verglichen, ohne dabei die besonderen Umstände im Automobil zu berücksichtigen. Dabei sind Speicherplatz und Rechenkapazität die größten einschränkenden Faktoren, da dadurch der Footprint für den Datenbestand der Sprachausgaben als auch die Möglichkeiten der Signalmanipulation begrenzt werden.

Die Erwartungshaltung des Hörers — welcher in den meisten Fällen mit dem Fahrer identisch sein wird — geht sogar soweit, dass die Sprachqualität unabhängig von der Fahrsituation

und der Fahrgeschwindigkeit gleich bleibend hoch ist. Es gibt jedoch noch kein System was auch nur annähernd die Anpassung der menschlichen Kommunikation an die Umgebungsbedingungen modelliert (wie z. B. den Lombardeffekt [5]). Mit Maßnahmen wie der Absenkung der aktiven Audio-Quelle und akustische Isolierung des Fahrzeuges wird versucht eine Verbesserung der Wahrnehmung der Sprachausgaben zu ermöglichen und somit die Hörererwartung zu erfüllen.

Die wichtigsten Aspekte des vielschichtigen Begriffs Sprachqualität sind *Verständlichkeit* und *Natürlichkeit*. Das sind grundsätzlich verschiedene Kategorien. Eine Äußerung ohne Sprachmelodie ist fast so verständlich wie mit Sprachmelodie, klingt aber sehr unnatürlich. Ein hemmungslos nuschelnder Mensch ist dagegen zwar natürlich, aber nicht verständlich.

3 Vergleich automotive-tauglicher TTS-Systeme

Die in diesem Abschnitt vorgestellte Untersuchung wurde mit dem Einverständnis der Hersteller durchgeführt. Die Darstellung der Ergebnisse erfolgt aus wettbewerbsrechtlichen Gründen anonymisiert, wie dies ja auch bei anderen Untersuchungen üblich ist.

3.1 Systemauswahl durch Paarvergleichs-Untersuchungen

Für die Untersuchungen wurden nur Systeme ausgewählt, die auch kommerziell verfügbar sind und die Sprachen der in Mercedes Fahrzeugen aktuell verbauten Sprachbedienung *Linguatronic* mit einer weiblichen Stimme abdecken (Englisch, Deutsch, Französisch, Spanisch, Italienisch und Niederländisch). Für die Durchführung der Systemauswahl war die Bereitstellung von Synthesebeispielen in deutscher Sprache nötig, die wahlweise durch ein Demo-System oder durch bereitgestellte Audio-Dateien erfolgen konnte. Der Inhalt wurde dabei vorgegeben und enthielt Navigationsansagen, wie sie auch im Fahrzeug vorkommen.

Mit diesen Einschränkungen wurden sechs TTS-Systeme von fünf Herstellern mit bis zu drei unterschiedlichen Stimmen und/oder unterschiedlichen Footprints pro Stimme und eine im Fahrzeug aufgenommene natürliche Stimme (hier als Pre-Recorded bezeichnet) verwendet. Die endgültige Auswahl erfolgte mittels Paarvergleichs-Test, indem alle vorhandenen Stimmen direkt gegeneinander verglichen wurden. 44 Testteilnehmer (ausgeglichen in männliche und weibliche Hörer) wurden gebeten am Online-Test teilzunehmen. Die überwiegende Mehrheit war dabei ohne Erfahrung im Bereich der Sprachsynthese, aber sehr wohl mit den Stimmen der Sprachausgaben im Mercedes vertraut. Die Hälfte der Testhörer verwendete Kopfhörer und jeweils ein Viertel HiFi-Lautsprecher bzw. Laptoplautsprecher, wobei die Ergebnisse nach Hörergruppen sortiert keine signifikanten Abweichungen zeigten.

Bei der Beurteilung im Paarvergleich mussten sich die Testteilnehmer in 5 Kategorien (Höranstrengung, Natürlichkeit, Verständlichkeit, Annehmlichkeit, Deutlichkeit) für das jeweils ihrer Meinung nach bessere Hörbeispiel entscheiden. Die Entscheidung konnte dabei individuell für jede Kategorie getroffen werden (vergleiche linkes Bild in Abbildung 2. Die Summation aller gewonnen Vergleiche einer System-Stimme-Kombination ist für alle Kategorien im Netzdiagramm in Bild 1 dargestellt.

Ein Vergleich der Summationsergebnisse mit den Ergebnissen einzelner Paare ergab keine wesentlich andere Einteilung. System 13, 14 und 15 wurden demnach besser bewertet als die Pre-Recorded Stimme und zeigen einen deutlichen Abstand zu den anderen Systemen. Bei dem System 15 handelt es sich um ein System mit einem größeren Footprint als bei Automotive-Systemen üblich und teilweise sind die Äußerungen bereits im Sprachdatenmaterial enthalten. Dieses System wurde als Referenz mit getestet und gibt an, welche Ergebnisse möglich sind, wenn man nicht den Einschränkungen des KFZ unterworfen ist. Die Systeme 13 und 14 pro-

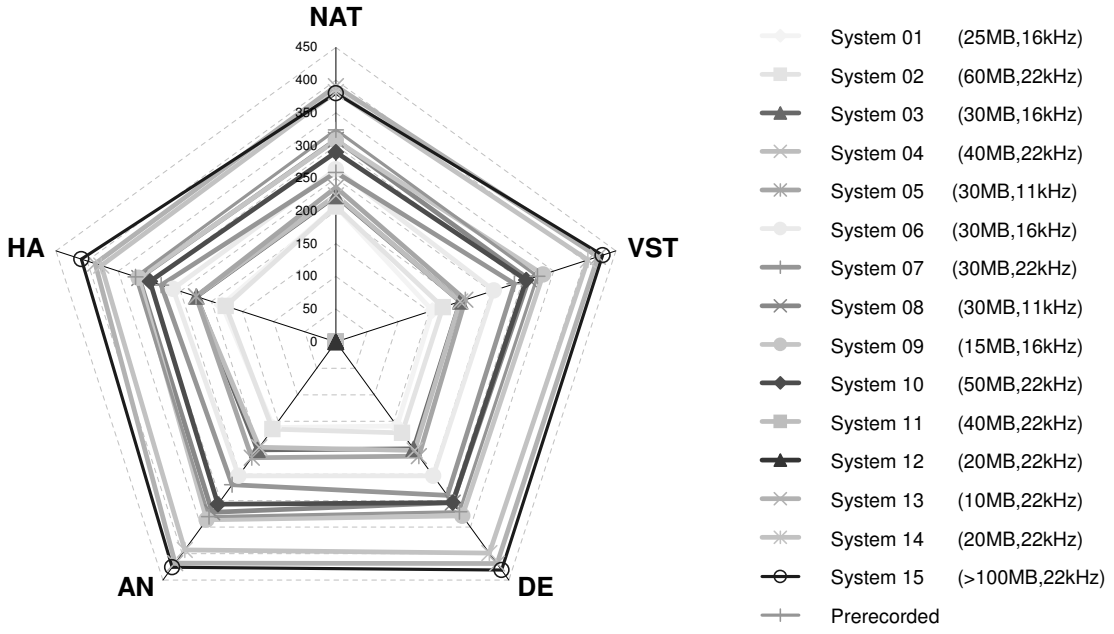

Abbildung 1 - Ergebnisse des Paarvergleich-Tests (Summation der gewonnenen Vergleiche pro System)

HA: Höranstrengung, NAT: Natürlichkeit, VST: Verständlichkeit, AN: Annehmlichkeit, DE: Deutlichkeit

fitieren offenbar von den speziellen Sprachdaten und dem Bekanntheitsgrad der Stimme. Dies konnte nicht weiter geprüft werden, da hier nur die Sprachbeispiele zur Verfügung gestellt wurden.

Ein weiteres Phänomen zeigt auch der Vergleich der Systeme 7 und 8. Hier wurde das System, bei dem die Sprachbausteine nur mit 11 kHz abgetastet wurde besser bewertet als das System mit den 22 kHz abgetasteten Bausteinen. Im Gespräch mit dem Hersteller wurde herausgefunden, dass das kleinere System viel öfter verkauft wurde und die Sprachbausteine teilweise manuell von kleineren Fehlern bereinigt wurden.

3.2 Systembewertung durch ACR-Untersuchungen

Für eine genauere Untersuchung wurden die besten vier Systeme von drei Herstellern mit den maximal zwei besten Stimmen verwendet. Im Rahmen der zweiten Stufe mussten allerdings Demo-Systeme vorliegen, um ein Tuning seitens der Hersteller zu verhindern. Da jedoch die Standardeinstellungen der Demo-Systeme verwendet wurden, konnten die Default Parameter von den Herstellern eingestellt werden. Um allerdings eine Anpassung an eine bestimmte Domäne und damit die Anpassung der Sprachqualität zu vermeiden [6], war im Voraus nicht bekannt, dass die Beispieläußerungen bei diesem Test aus der Domäne Verkehrsmeldungen (TMC) kommen.

Im Rahmen des Tests, der als ACR-Hörtest ausgelegt war, sollten die Kategorien Gesamtqualität, Höranstrengung, Verständlichkeit, Natürlichkeit und Deutlichkeit beurteilt werden. Für die Beurteilung des MOS–Wertes zur Einschätzung der Gesamtqualität wurde die in der ITU-T P.800(08/96) empfohlene 5-Punkte Skala verwendet. Für die Bestimmung der Höranstrengung wurde ebenfalls die in der ITU empfohlene Skala verwendet, allerdings ist die Wertzuordnung entgegen der ITU–Empfehlung umgekehrt. Dadurch wird eine geringere Höranstrengung auch durch kleinere Werte quantifiziert (vergleiche rechtes Bild in Abbildung 2).

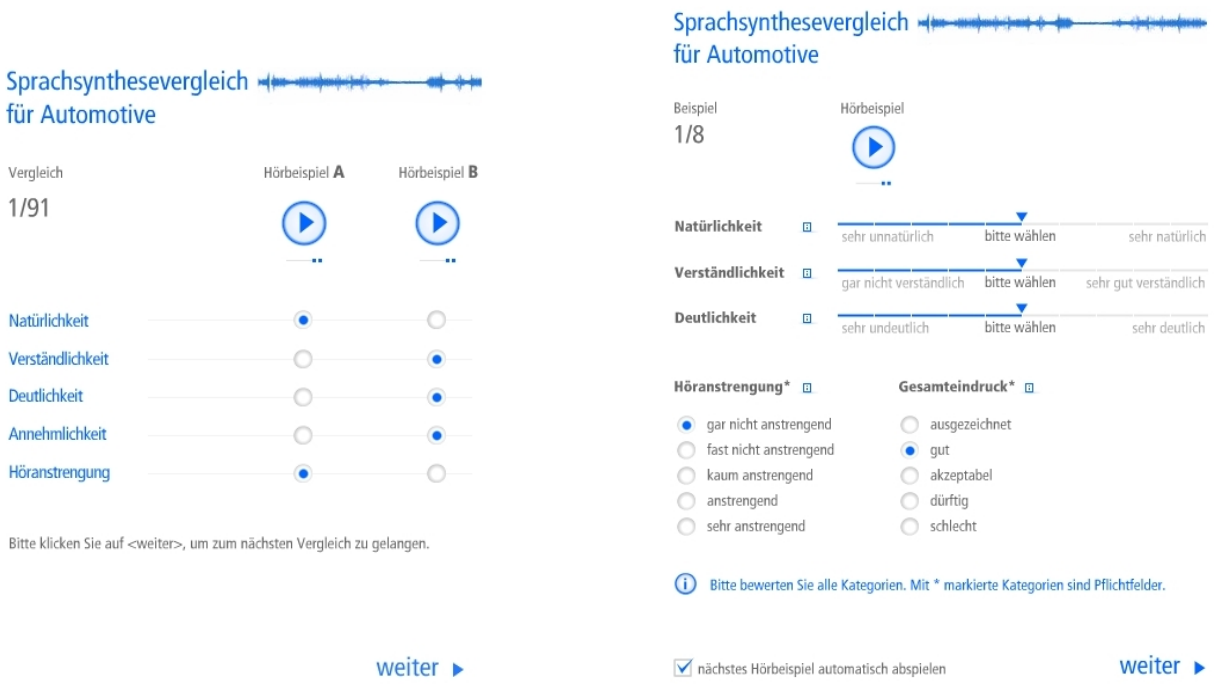

Abbildung 2 - Beispiel der Online-Testumgebung für den Paarvergleichstest (links) und den ACR-Hörtest (rechts)

Für die Bewertung der drei Kategorien Verständlichkeit, Natürlichkeit und Deutlichkeit wurde den Testteilnehmern ein Schieberegler präsentiert, der an den Enden gegensätzlich beschriftet und auf die Mitte voreingestellt war (vergleiche rechtes Bild in Abbildung 2). Die vom Hörer eingestellten Werte schwanken also um einen Mittelwert, der dem Erwartungswert des Hörers entspricht. Die Verschiebung des Reglers misst dementsprechend seine Beurteilung (positiv oder negativ abweichend) gegenüber seiner eigenen Referenz. Intern (für den Hörer nicht sichtbar) wurde eine unipolare Intervallskala mit 10 äquidistanten Intervallen benutzt (Mittelwert 5). Für eine Diskussion, ob bei einer solchen Präsentation die Beurteilung durch den Hörer Intervalleigenschaften aufweist oder nicht, wird auf die Literatur verwiesen [7]. Wie allgemein üblich werden hier die Mittelwerte angegeben und somit Intervalleigenschaften vorausgesetzt.

Wie die Ergebnisse des ersten Teils vermuten lassen, beeinflusst offensichtlich der Bekanntheitsgrad einer Stimme die Bewertungen. Deshalb wurde der zweite Teil der Untersuchungen mit Probanden durchgeführt, die zu einem überwiegenden Teil keine Erfahrung mit der Linguatronic hatten. Durch die Verwendung kommerzieller Systeme kann jedoch nicht ganz ausgeschlossen werden, dass einzelne Probanden bestimmte Stimmen bereits kennen und/oder bevorzugen. Die Probanden setzten sich diesmal zu $3/4$ männlichen und $1/4$ weilichen Hören zusammen. Der Online-Test wurde dabei zu jeweils $1/3$ mit Kopfhörer, mit HiFi-Lautsprechern bzw. mit Laptop-Lautsprechern (eigene Angaben der Testpersonen) durchgeführt. Eine stichpunktartige Ergebnisanalyse in Abhängigkeit der Hörergruppe ergab keine signifikanten Unterschiede zum Gesamtergebnis. Deshalb wird auch nur Letzteres hier dargestellt.

Die Bewertungen für den Gesamteindruck (MOS) und die Höranstrengung sind in Abbildung 3 als Balkendiagramm dargestellt. Sofort ist zu erkennen, dass es im Gegensatz zum Paarvergleich keine großen Differenzen zwischen einzelnen Systemen gibt. Dennoch ist eine eindeutige Rangordnung erkennbar, die bei beiden Kriterien identisch ist. Dies ist auf die hohe Korrelation der beiden Kategorien zurückzuführen (vergleiche dazu auch [8]).

Die Ergebnisse der weiteren Kategorien sind in Abbildung 4 dargestellt. Die ersten fünf

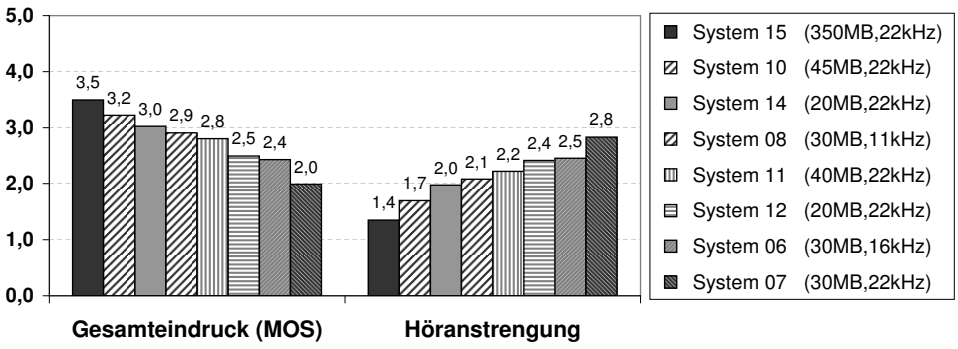

Abbildung 3 - Ergebnisse (arithmetische Mittelwerte) der ACR-Test-Bewertungen in den Kategorien Gesamteindruck (MOS) und Höranstrengung

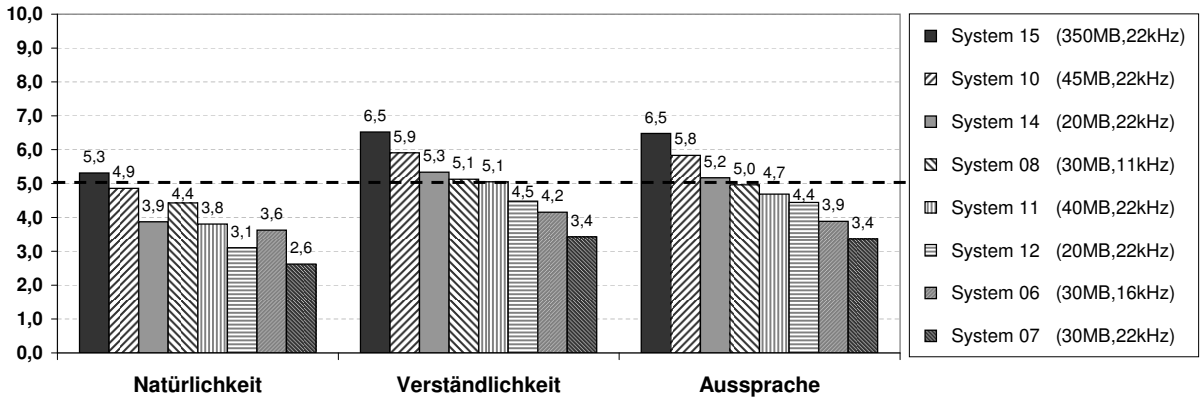

Abbildung 4 - Ergebnisse (arithmetische Mittelwerte) der ACR-Test-Bewertungen in den Kategorien Natürlichkeit, Verständlichkeit, und Aussprache

System-Stimm-Kombinationen liegen in der Kategorie Verständlichkeit und Deutlichkeit knapp über der mittleren Hörererwartung, aber in der Kategorie Natürlichkeit knapp darunter (außer Referenz-System 15 mit höherem Footprint). Die Verständlichkeit ist also auf dem Niveau, wie es die Nutzer akzeptieren.

Die Systeme 6 und 7, welche eine sehr ausgeprägte Akzentsteuerung und damit verbunden eine starke Grundfrequenzänderung besitzen, wurden dagegen weniger gut bewertet. Andere Untersuchungen mit inhaltlich anderen Fokussierungen sehen genau diese Systeme eher im besseren Bereich. Um Umkehrschluss bedeutet dies, dass anscheinend eine ausgeprägte oder gar übertriebene Prosodiesteuerung eher ungünstig für die Bewertung von Sprachäußerungen im Automotive-Umfeld ist.

3.3 Untersuchung des Einflusses von Offline–Synthese-Tunings

Einige kommerzielle Anbieter bieten Tuning-Möglichkeiten, vor allen für Systeme mit kleinerem Footprint. Tuning wird dabei aber sehr weit interpretiert. Es reicht von einer Optimierung von Intonationsverläufen und dem Hinzufügen von Akzentstellen, über das Hinzufügen zusätzlicher Bausteine aus einem größeren System bis hin zu inhaltsoptimierten Synthesen (Domänen spezifisch) [2, 6]. Solche Verfahren eignen sich vor allem für die Vorbereitung von festen Ausgabeinhalten, die offline generiert und gespeichert werden können.

Leider reichen die Daten nicht aus, um statistische Untersuchungen zu präsentieren. Die ersten Ergebnisse deuten nur eine sehr geringe Verbesserung in den einzelnen Kategorien an. Die ersten wenigen Daten zeigen eine für die vorliegenden kurzen Äußerungen der Domäne TMC eine um 0,2 höhere Gesamtbewertung und eine um 0,1 geringere Höranstrengung. Gleichzeitig sinkt die Natürlichkeit um 0,1. Jedoch waren die Unterschiede nicht signifikant.

4 Qualitätssicherung von Sprachausgaben im KFZ

Die Einführung einer voll synthetischen Sprachausgabe bringt neben der höheren Flexibilität bei einer Erweiterung eines bestehenden Dialog- und/oder Navigationssystems auch eine größere Unabhängigkeit von professionellen muttersprachlichen SprecherInnen für Audioaufnahmen mit sich. Gleichzeitig kann die Qualität von gemischten Systemausgaben erhöht werden, wie z. B. das Vorlesen von Routeninformationen mit einer Mischung von dynamischen und statischen Inhalten.

Die Einführung von synthetischen Sprachausgaben, vor allem wenn diese erst im System generiert werden, erfordert aber eine zusätzliche Qualitätssicherung durch qualitativ und quantitativ bestimmbare Gütekriterien. Objektive Gütekriterien sind direkt messbar bzw. aus messbaren Größen berechenbar wie beispielsweise Erkennungsraten in der Spracherkennung. Dagegen erfolgt die Beurteilung von Synthesesystemen derzeit vorwiegend in subjektiven Kriterien durch Hörtests.

Diese Hörtests müssen jedoch entsprechend gelenkt werden, um den gewünschten Fokus der Beurteilung zu bekommen [3]. Dazu ist die Berücksichtigung der Domäne enorm wichtig. D.h. die gesamte Evaluierung vor einer Systemauswahl sowie die Qualitätssicherung nach einer Systemintegration muss anwendungsbezogen erfolgen. Die Resultate allgemeiner Evaluierungen wie z. B. der Blizzard challenge [9] sind nur bedingt für die Systemauswahl geeignet. Das erfordert aber implizite Untersuchungen zu allen möglichen Domänen der synthetisch generierten Sprachausgaben [10].

Im Rahmen eines Entwicklungsprozesses wäre es also von Vorteil, teilweise automatisierte Testverfahren zur Verfügung zu haben. Dazu werden immer wieder Ansätze diskutiert [11, 12, 13, 14], die aber oft nur auf die Signalform und/oder bestimmte Parameter der Sprachbeispiele abzielen und teilweise auf Vergleiche von synthetischen mit natürlichen Mustern angewiesen sind. Bisher offen ist allerdings die Frage: Welchen Einfluss haben gefundene Abweichungen bzw. gemessene Größen auf die Qualitätsbewertung von sprachlichen Äußerungen und inwieweit oder ab welcher Stärke werden diese überhaupt vom Hörer wahrgenommen.

Durch den Einsatz von Sprachsynthese im KFZ rückt auch die dazugehörige Vorverarbeitung stärker in den Fokus der Qualitätssicherung. Denn z. B. Navigationsdaten und Verkehrsnachrichten müssen entsprechend aufbereitet werden, dass keine Formate wie Datum und/oder Uhrzeit und keine Straßenabkürzungen mehr vorhanden sind. Es gibt durchaus einige Systeme bei denen dies gut gelungen ist [15], allerdings ohne eine Beschränkung in der Rechenleistung.

5 Zusammenfassung und Ausblick

Die vorgestellte Untersuchung zeigt einen Vergleich automotiv-tauglicher Sprachsynthese-Systeme. Dabei zeigte sich, dass es sowohl qualitative Unterschiede zwischen den Systemen, als auch unterschiedliche Bewertungen im Vergleich zu etablierten Challenges gibt. Letztere ist wohl hauptsächlich auf eine begrenzte thematische Fokussierung zurück zu führen. Die Vorliegenden Untersuchungen, wurden zwar mit real vorkommenden Sprachausgaben durchgeführt, jedoch wurde bisher die Geräuschkulisse im Fahrzeug außer acht gelassen. Die Weiterführung

der Untersuchungen soll die Einflüsse von Umgebungsgeräuschen z. B. durch unterschiedliche Geschwindigkeiten und/oder Straßenbeläge mit einbeziehen.

Durch einen exklusiven Einsatz von synthetisch erzeugten Sprachausgaben wird sowohl eine höhere Flexibilität für die inhaltliche Ausführung als auch eine Reduktion des Aufwandes für Sprachaufnahmen erreicht. Gleichzeitig wird aber der Qualifikationsaufwand erhöht, da es derzeit keine gängige und etablierte automatische Messmethoden zur Sprachqualitätsbeurteilung gibt.

Literatur

[1] "http://www.youtube.com/watch?v=zlfmq5iwvji."

[2] A. W. Black, "Perfect synthesis for all of the people all of the time," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pp. 167 – 170, September 2002.

[3] U. Jekosch, *Voice And Speech Quality Perception : Assessment And Evaluation*. Berlin, Germany: Springer Verlag, 2005.

[4] A. Meroth and B. Tolg, *Infotainmentsysteme im Kraftfahrzeug*. Wiesbaden: GWV Fachverlage GmbH, 2008. ISBN: 978-3834802859.

[5] D. D. Silverman, *A Critical Introduction to Phonology: Of Sound, Mind and Body*. Continuum International Publishing Group, 2006.

[6] A. W. Black and K. A. Lenzo, "Limited domain synthesis," in *Proc. Intl. Conference on Spoken Language Processing (ICSLP)*, (Beijing, China), pp. 411–414, 2000.

[7] L. J. Bronwen and P. R. McManus, "Graphic scaling of qualitative terms," *Society of Motion Picture and Television Engineers (SMPTE)*, vol. 95, pp. 1166–1171, Nov. 1986.

[8] S. Werner, *Sprachsynthese und Sracherkennung mit gemeinsamen Datenbasen*. PhD thesis, TU Dresden, 2007.

[9] R. Clark, A. J. R. M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," *BLZ3-2007*, 2007.

[10] S. Möller, U. Jekosch, J. Mersdorf, and V. Kraft, "Auditory assessment of synthesized speech in application scenarios: Two case studies," *Speech Communication*, vol. 34, pp. 229–246, Juni 2001.

[11] L. Cai, R. Tu, J. Zhao, and Y. Mao, "Speech quality evaluation: A new application of digital watermarking," *IEEE Transactions on Instrumentation and Measurement*, vol. 56, pp. 45–55, Februar 2007.

[12] A. W. Rix, "Perceptual speech quality assessment - a review," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Montreal, Canada), pp. III–1056–1059, Mai 2004.

[13] T. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive gmm-based speech quality measurement," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (März), pp. I–125–128, 2005.

[14] D.-S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, (Montreal, Canada), pp. III–1060–1063, Mai 2004.

[15] "SWR3-Stauhotline unter 07221/9282 oder http://www.swr3.de/info/verkehr/."