GUT UND GÜNSTIG? – NUTZUNG DES GOOGLE SPEECH API IN SPRACHDIALOGSYSTEMEN

Stefan Schmidt

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin
s.schmidt@tu-berlin.de

Kurzfassung: Eine Untersuchung der Erkennungsleistung des Google Speech API (GS-API) für die deutsche Sprache, ergab eine Wortfehlerrate (*WER*) von 27 %. 1010 Äußerungen, aus der Domäne Smart Home Environment, mit insgesamt 3.317 Wörtern wurden von 20 Versuchsteilnehmern eingesprochen. In der Auswertung zeigte sich der Effekt, dass die *WER* bei weiblichen Sprechern um 7,8 % höher liegt als bei Männern. Die Antwortzeit des API liegt bei 600 – 400 ms pro Wort einer Äußerung. Weiterhin wird gezeigt wie sich das GS-API sowohl aus einem Webbrowser – als auch einer selbst implementierten Anwendung – heraus nutzen lässt. Im Rahmen der dieser Arbeiten wurde ein Java-basierter Client implementiert, welcher der Allgemeinheit zur Verfügung gestellt wird.

1 Einleitung

Mit Systemen, wie sie von Nuance oder Loquendo vertrieben werden, existieren gute und erprobte Lösungen zur Spracherkennung. Diese können u.a. in Plattformen wie Prophecy (Voxeo) oder OptimTalk VoiceBrowser (OptimSys) eingebunden werden. Solche Plattformen erlauben eine relativ einfache Implementierung von Sprachdialogsystemen, z. B. auf der Basis von VoiceXML. Sie bergen aber auch Nachteile in sich, so müssen für den Einsatz in der Lehre die entsprechenden Lizenzen in ausreichender Menge erworben werden. Auch für den Einsatz in der Forschung sind sie nur begrenzt geeignet, da eine Erweiterung um neue Fähigkeiten, z. B. beim Dialogmanagement, nur sehr eingeschränkt möglich ist. Weiterhin darf der Aufwand zum Aufsetzen und Pflegen der Plattformsysteme nicht unterschätzt werden. Daher wäre, insbesondere in dem Kontext von Forschung und Lehre, die Verfügbarkeit von frei nutzbaren Spracherkennern sehr hilfreich.

Mit CMU Sphinx existiert ein freier Spracherkenner, der jedoch auf entsprechende Sprachmodelle angewiesen ist. Ein zuverlässiges, freies Modell der deutschen Sprache ist für Sphinx derzeit aber nicht verfügbar. Auch das viel versprechende, in InproTK ([1]) genutzte Sprachmodell muss lizenziert werden.

Mit Apples Siri und dem Google Speech API (GS-API) existieren Lösungen, die über frei zugängliche Schnittstellen verfügen (auch wenn diese nur indirekt dokumentiert sind) und auch die deutsche Sprache unterstützen. Ob das GS-API eine hinreichend kleine Wortfehlerrate (*WER*) hat und welche Antwortzeiten zu erwarten sind, wenn es nicht für die Erkennung von Suchanfragen an eine Suchmaschine, sondern im Kontext eines multimodalen Dialogsystems genutzt wird, soll in diesem Beitrag untersucht werden.

Hauptteil dieses Beitrags ist Abschnitt 2, in welchem eine empirische Untersuchung und deren quantitative Ergebnisse dargestellt werden. In Abschnitt 3 werden die Möglichkeiten zur Nutzung des GS-API kurz beschrieben. Eine Diskussion der Ergebnisse (4) sowie ein kurzer Ausblick (5) schließen den Beitrag ab.

2 Empirische Untersuchung des Google Speech API

Um die Performanz des GS-API, u. a., hinsichtlich der WER und der Antwortzeit für das Deutsche zu ermitteln, wurde eine Untersuchung mit 6 Frauen (Alter: 28–58 Jahre, Ø 36,17, SD 11,2) und 14 Männern (Alter: 28–58 Jahre, Ø 35, SD 7,34) durchgeführt. Alle zwanzig Teilnehmer sind deutsche Muttersprachler.

Um die Übersicht zu wahren, werden die technischen Details zur Nutzung des GS-API gesondert in Abschnitt 3 beschrieben.

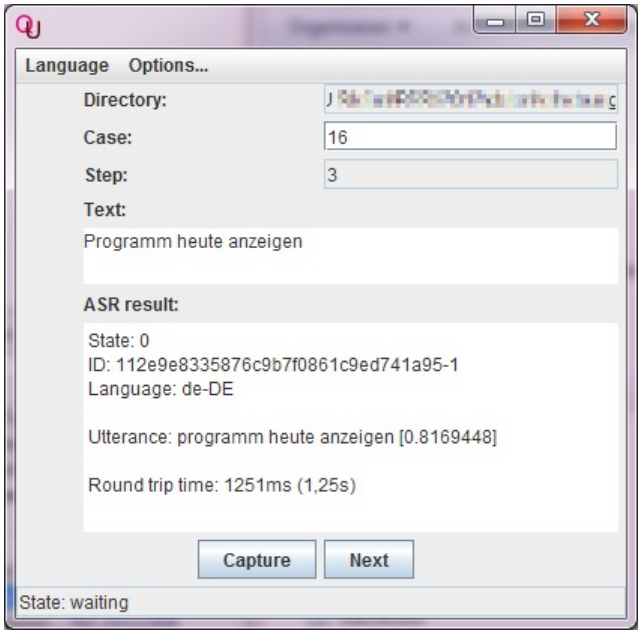


Abbildung 1 - Screenshot der im Versuch genutzten Anwendung.

Jeder der Teilnehmer war aufgefordert 51 vorgegebene Äußerungen auf deutsch über ein Headset einzusprechen. Um die Aufnahme zu starten, klickt der Nutzer einen Button ("Capture") in der Experimentanwendung (s. Abbildung 1). Nach dem Einsprechen der Äußerung klickt er nochmals auf den gleichen Button. Die Aufnahme wird daraufhin sofort an das GS-API gesendet und die Antwort dem Nutzer angezeigt. Die Gesamtzeit welche benötigt wird, um die Daten an das API zu senden, dort zu verarbeiten und die Antwort mit dem Ergebnis zurückzusenden wird gemessen und festgehalten. Diese Zeitdauer wird im Folgenden als RTT (Round Trip Time) bezeichnet.

Tabelle 1 - Zusammensetzung des Korpus K_V, welcher von den Teilnehmern eingesprochen wurde.

Länge	Anzahl	Länge	Anzahl	Länge	Anzahl	Länge	Anzahl	Länge	Anzahl
1	151	5	87	9	8	14	1	21	1
2	289	6	56	10	7	15	2	23	1
3	251	7	14	11	2	16	6		
4	105	8	21	13	1	17	1		

Der Korpus von Äußerungen (K_V), der für die Untersuchung verwendet wurde, basiert auf Transkriptionen der in [5] und [6] beschriebenen Studie mit dem multimodalen Smart Home System INSPIRE [3, S. 32-34 und 128-130]. K_V enthält alle voneinander verschiedenen Äußerungen

welche die Nutzer in dem damaligen Versuch an das multimodale System gerichtet haben. Das bedeutet, wenn die gleiche Äußerung mehr als einmal getätigt wurde, ist sie trotzdem nur einmal in K_V enthalten. Auf Basis der 662 verschiedenen Äußerungen konnten 1010 verwertbare Anfragen an das GS-API generiert werden. Jeder Teilnehmer erhielt 51 zufällig ausgewählte Äußerungen, wobei sicher gestellt wurde, dass im Laufe der Untersuchung jede Äußerung mindestens einmal und maximal zweimal Verwendung fand. Ca. 10 Datensätze konnten nicht verwendet werden, da sich dort in der Nachverarbeitung Probleme bei der Aufnahme (Aussetzer) zeigten. Im Zuge der Nachverabeitung wurden alle aufgenommen Äußerungen gegen die vorgegeben Phrasen geprüft und Letztere ggf. angepasst. So wurde der resultierende Korpus K_R erzeugt, welcher für jede Anfrage an die GS-API, neben dem Wortlaut der tatsächlichen Äußerung und dem der Hypothese (mit dem höchsten Konfidenzwert) des GS-API , die jeweilige Länge (Anzahl Wörter) und die Antwortzeit enthält.

2.1 Wortfehlerrate

Auf dem Korpus K_R wurde die WER anhand des in [7] beschriebenen Verfahrens, das auf der Levenshtein-Distanz basiert, berechnet. Die einzelnen Ergebnisse hinsichtlich Ersetzung (Substitution), Einfügung (Insertion) und Löschung (Deletion) von Wörtern durch den automatischen Spracherkenner (ASR) zeigt Tabelle 2. Es ergibt sich eine totale WER von 27 %, wobei auffällt, dass die WER bei Frauen um 7,8 % höher liegt als bei Männern.

Tabelle 2 - WER bezogen auf den gesamten Korpus sowie getrennt nach dem Geschlecht der Teilnehmer.

		Substitution		Insertion		Deletion		
	Wörter	absolut	%	absolut	%	absolut	%	WER (%)
Total	3.317	577	17,4	84	2,5	235	7,1	27,0
Männer	2.281	395	17,3	58	2,5	107	4,7	24,6
Frauen	1.036	182	17,6	26	2.5	128	12,4	32,4

Der Unterschied in der WER ergibt sich aus dem viel höheren Anteil von Deletions bei Frauen. Die Nullhypothesen, dass die WER und die Häufigkeit von Deletions unabhängig vom Geschlecht des Sprechers sind, mussten, jeweils aufgrund eines Chi-Quadrat-Test, verworfen werden (vgl. Zeile K_R in Tabelle 4).

Tabelle 3 - WER bezogen auf den gesamten Korpus sowie getrennt nach dem Geschlecht der Teilnehmer – nach der Korrektur von zusammengesetzten Hauptwörter.

		Substitution		Insertion		Deletion		
	Wörter	absolut	%	absolut	%	absolut	%	WER (%)
Total	3.317	557	16,8	64	1,9	235	7,1	25,8
Männer	2.281	379	16,6	42	1,8	107	4,7	23,1
Frauen	1.036	178	17,2	22	2,1	128	12,4	31,7

Im Zuge der Datenauswertung fiel auf, dass das GS-API scheinbar recht häufig zusammengesetzte Hauptwörter (z.B. Spielfilm) als einzelne Wörter (z.B. spiel film) erkannte. Da für jeden dieser Fälle in der WER-Berechnung jeweils eine Substitution und eine Insertion (“Spielfilm” wird

Tabelle 4 - Ergebnisse des Chi-Quadrat-Unabhängigkeitstests für die Korpora K_R und K_K. Getestet wurde jeweils die H_0, dass das Auftreten eines Erkennungsfehlers unabhängig vom Geschlecht des Sprechers sei.

	p-Wert aus χ^2-Test, $\alpha = 0.05$			
	summierte Fehler	Substitutions	Insertions	Deletions
K_R	$4,67 \cdot 10^{-9}$	0,9	0,95	$2,78 \cdot 10^{-15}$
K_K	$2,61 \cdot 10^{-9}$	0,72	0,68	$2,78 \cdot 10^{-15}$

durch “Spiel” ersetzt und zusätzlich “Film” eingefügt) gezählt wird, wurde eine zweite Analyse durchgeführt. Dort wurde der korrigierten Korpus K_K verwendet, welcher aus K_R abgeleitet wurde indem die in Tabelle 5 gelisteten Ersetzungen in den API-Antworten vorgenommen wurden. Die Resultate sind in Tabelle 3 gelistet. Es konnte zwar eine Verringerung der WER um 1,2 % erreicht werden, ein Chi-Quadrat-Test auf Unabhängigkeit konnte jedoch keinen signifikanten Unterschied gegenüber der WER in K_V nachweisen.

Die signifikanten Unterschiede zwischen Männer und Frauen hinsichtlich der WER und Deletions blieben bestehen (s. Zeile K_K in Tabelle 4). Dies war ist jedoch leicht dadurch erklärbar, dass die angewandten Korrekturen keine Auswirkung auf die Anzahl der Deletions in den einzelnen Gruppen haben.

Tabelle 5 - Korrigierte Trennung von zusammengesetzten Hauptwörtern in den Resultaten der ASR.

Erkanntes Wort	verwendete Ersetzung
musik sammlung	musiksammlung
programm information	programminformation
spiel film	spielfilm
zurück rufen	zurückrufen

2.2 Round Trip Time

Neben der Wortfehlerrate wurde auch die RTT gemessen. Diese Zeitspanne reicht vom Start des Versands der Audiodaten an das GS-API bis zum vollständigen Erhalt der Antwort. Die Untersuchungen wurden an 3 unterschiedlichen Standorten durchgeführt, welche auch jeweils unterschiedliche Verbindungen in das Internet bereitstellten:

Direkt Direkte Anbindung des Client an das universitätsinterne LAN.

Proxy Anbindung über WiFi ein VPN mit zusätzlicher Verzögerung durch einen HTTP Proxy.

DSL Anbindung über Wifi an einen gewöhnlichen DSL Anschluss bei exklusiver Nutzung durch den Client.

Für die weitere Auswertung wurden nur Äußerungen der Länge eins bis sechs Wörter herangezogen da, die Fallzahl bei größere Längen zu gering war. Wie zu erwarten ist, steigt die RTT mit zunehmender Länge der zu erkennenden Äußerung an, was in Abbildung 2 dargestellt ist.

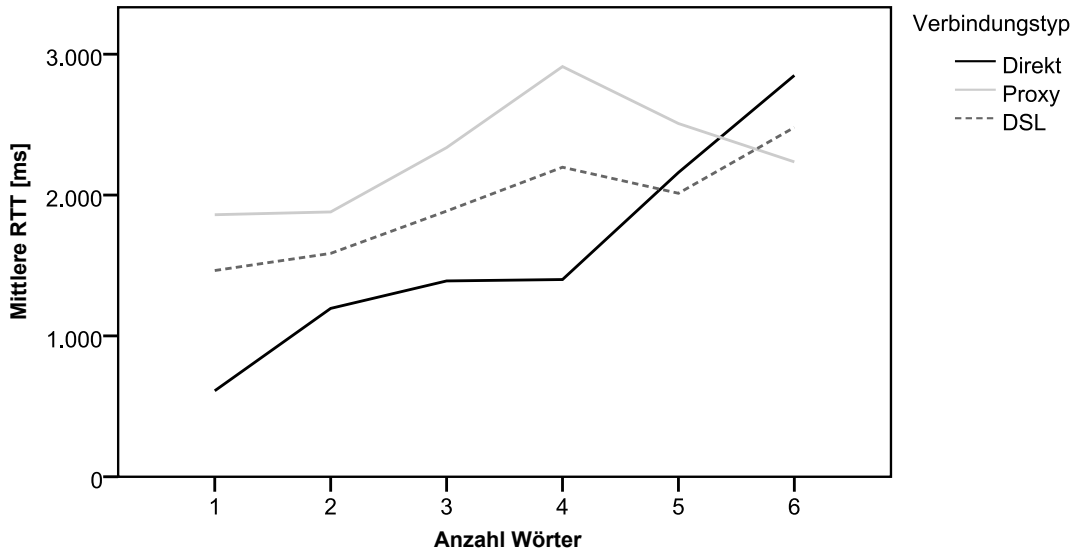


Abbildung 2 - Vergleich, der benötigten RTT bei ansteigender Wortanzahl in der Äußerungen, für die verschiedenen Verbindungen.

Eine Zerlegung der RTT in die Zeit die für den Datentransfer benötigt wird und die tatsächliche Verarbeitungszeit durch die ASR war nicht möglich.

Tabelle 6 listet die durchschnittliche RTT pro Wort (RTT_W). Sei N die Anzahl der Wörter in der Äußerung A dann gilt: $RTT_W = \frac{RTT_A}{N}$. Es zeigt sich also, dass bei Äußerungen der mit sechs Wörtern, je nach Verbindungsqualität, eine RTT (die sich für den Nutzer als Wartezeit niederschlägt) zwischen 2 und 3 Sekunden auftritt.

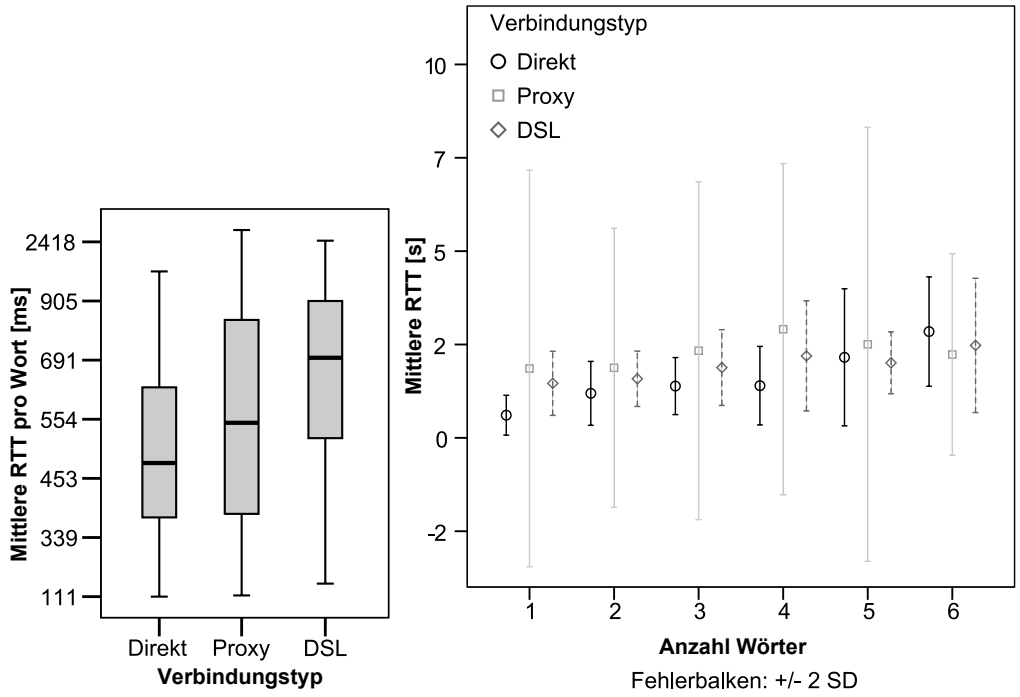
Tabelle 6 - Mittlere RTT_W (ms) bei Äußerungen der Längen 1 bis 6 Wörter in Abhängigkeit von der Verbindungsart.

	RTT_W (ms) für $N = 1 \ldots 6$					
Verbindungstyp	1	2	3	4	5	6
Direkt	599	592	460	347	430	473
DSL	1.849	934	775	725	499	371
Proxy	1.452	787	625	546	400	411

Abbildung 3a und auch Tabelle 6 suggerieren zunächst, dass die Nutzung hinter einem Proxy keine großen Nachteile mit sich bringt, tatsächlich trat dort aber im Maximum eine RTT von 11,3 Sekunden auf. Solche Ausschläge zeigen sich deutlich im Diagramm 3b, in welchem neben der RTT auch deren Standardabweichung, für jeden Verbindungstyp und die Äußerungenlängen 1 bis 6, eingetragen ist.

3 Nutzung des Google Speech API

Das GS-API kann aus Webseiten heraus mittels des Browsers Google Chrome – durch ein im HTML Code entsprechend markiertes <input />Tag – oder aber aus einer beliebigen Anwendung heraus, durch das Absenden eines entsprechenden HTTP-Post-Request angesprochen werden.



(a) Boxplot der mittleren RTT_W je Verbindungstyp

(b) Balken

Abbildung 3 - Boxplot und Darstellung der Standardabweichung bei den ermittelten RTT.

Die Attribute und URL Parameter, welche dabei verwendet werden können, sind in Tabelle 7 beschrieben.

Über eine feste URL[1] ist das GS-API via Internet, ohne die Notwendigkeit einer Authentifizierung, erreichbar. In eigenen Anwendungen kann diese URL individuell parametrisiert werden, um einen Post-Request mit dem Sprachsignal an die ASR zu senden. Das Signal kann wahlweise im FLAC (Free Lossless Audio Codec) oder Speex Format kodiert sein, wobei die Abtastrate zwischen 8000 Hz und 44.000 Hz liegen muss.

Für die in Abschnitt 2 beschriebene Untersuchung wurde ein Java-Client implementiert, welcher eine einfache Parametrisierung des notwendigen HTTP-Request erlaubt, FLAC-kodierte Audioaufnahmen an das GS-API sendet und die resultierende API-Antwort aufbereiten kann. In Abbildung 4 ist die Antwort der GS-API auf eine Anfrage dargestellt. Das Resultat des Erkenners wird von dem GS-API JSON (JavaScript Object Notation) codiert geliefert. In der Antwort enthält das Feld *status* ggf. einen Fehlercode (0 = kein Fehler, weitere Codes siehe [8]), der Wert von *id* ist für jede einzelne Anfrage eindeutig. Das nachfolgende Array (*hypotheses*) enthält die n-best Liste der Hypothesen des Spracherkenners. Zu jeder Hypothese wird die erkannte Äußerung (*utterance*) geliefert, jedoch nur für die Erste ein Konfidenzwert (*confidence*). Für alle weiteren Hypothesen in der n-best Liste ist dieser Wert immer auf 0 gesetzt.

```
{"status":0,"id ":"3943e18923393c0169a8659f5ef8db6a-1",
"hypotheses":[{"utterance":"das pferd frisst keinen
gurkensalat",
"confidence":0.90929186}]}
```

Abbildung 4 - Antwort der GS-API an den Client im JSON Datenformat.

[1] `https://www.google.com/speech-api/v1/recognize?xjerr=1&client=chromium&` (vgl. [2])

Tabelle 7 - Bekannte Request-Parameter der GS-API, entnommen aus [2, 4].

<input /> Attribut	URL Parameter	Erläuterung
x-webkit-speech	—	Mit diesem proprietären Attribut markierte Eingabefelder können durch den HTML Browser mittels Spracheingabe gefüllt werden (derzeit ausschließlich in Chromium/Chrome).
x-webkit-grammar	*lm*	Dient der Angabe einer URL, welche auf eine zu verwendende Grammatik (SRGS) verweist. Momentan werden offenbar nur zwei innerhalb von Google Chrome vordefinierte Grammatiken (*builtin:search*, *builtin:translate*) tatsächlich durch die GS-API verarbeitet.
lang	*lang*	Angabe der Sprache im Locale-Format (bspw. *de_DE* oder *en_US*).
—	*client*	Optionaler Name des Client.
—	*xjerr*	Steuert die Fehlerrückgabe. Der Werte *1* veranlasst die Rückgabe von Fehlern im JSON-formatierten Resultat, anstatt in Form von HTML Fehlercodes.
—	*maxresults*	Maximale Anzahl der durch das API zu liefernden Hypothesen des Erkenners.

4 Diskussion und Zusammenfassung

Für die in Abschnitt 2 vorgestellte Studie zur Untersuchung der GS-API hinsichtlich der WER und der Antwortzeiten wurden Äußerungen verwendet, wie Sie in der Interaktion von Nutzern mit einem multidmodalen System auftreten. Typische Äußerungen in der Interaktion mit INSPIRE sind z. B. “Bitte den Fernseher einschalten”, “Nachricht löschen” oder “Vierten Titel abspielen”. Für Äußerungen dieser Art – die nicht typischen Suchanfragen an eine Suchmaschine entsprechen – wurde eine totale WER von 27 % gemessen. Ein interessantes Phänomen ist die Differenz von 7,8 % in der WER für Frauen und Männer (32,4 % und 24,6 %). Der Unterschied entsteht durch den viel höheren Anteil von Deletions bei den Äußerungen von Frauen. Die Ursache dafür konnte im Rahmen dieser Arbeit noch nicht weiter untersucht werden. Es kann spekuliert werden, ob die verwendete ASR besser auf männliche Stimmen trainiert ist, aber auch ob die am Test teilnehmenden Frauen stimmliche oder sprachliche Besonderheiten aufwiesen.

Die Auswertung zeigt weiterhin, dass die Antwortzeit bzw. die RTT des GS-API stark von der Art der Verbindung des Client zum Internet abhängt. Bei einer direkten Verbindung über ein LAN sind Antwortzeiten von zwischen 400 und 600 Millisekunden pro Wort in einer Äußerung realisierbar (vgl. Tabelle 6). Im Gegensatz dazu führte die Nutzung über einen Web-Proxy zwar zu einer mittleren RTT_W von 1.452 ms (ein Wort) bis 411 ms (sechs Wörter), aber ist bei einer Standardabweichung von 2168 ms nicht verlässlich in einem Dialogsystem nutzbar.

Insgesamt zeigt sich das die GS-API geeignet ist, um in Übungen, studentischen Projekten usw. verwendet zu werden. Die GS-API ASR stellt eine Alternative zu andern frei nutzbaren Spracherkennern da, insbesondere für das Deutsche. Ein Manko ist die aktuell noch fehlende Möglichkeit individuelle, auf die jeweilige Anwendung hin zugeschnittene, Grammatiken angeben zu können. Im Rahmen einer im Sommersemester 2012 durchgeführten Veranstaltung des Quality and Usability Labs (TU-Berlin) haben mehrere studentische Übungsgruppen das GS-API

in ihre multimodalen Übungsprojekte erfolgreich integrieren können. Auch dort entstand der subjektive Eindruck, dass die Erkennung bei weiblichen Sprechern höhere Fehlerraten aufweist.

5 Ausblick

Auf Basis der in dieser Arbeit vorgestellten Ergebnisse wird das GS-API in verschiedenen unserer Lehr- und Demonstrationssystem eingebunden werden. Eine Klärung der Frage worin die wesentlich höhere Wortfehlerrate bei Sprecherinnen begründet ist, muss durch eine weitergehende Analyse des vorliegenden Datenmaterials und eine eventuelle nachfolgende Untersuchung erfolgen. Weiterhin können die gewonnen Daten dazu genutzt werden, eine Simulation des Erkenners des GS-API zu trainieren. Solch eine Simulation wiederum, kann im Rahmen der automatischen Usabilityevaluierung von Sprachdialogsystemen genutzt werden.

Literatur

[1] BAUMANN, T., O. BUSS und D. SCHLANGEN: *InproTK in Action: Open-Source Software for Building German-Speaking Incremental Spoken Dialogue Systems*. In: MIXDORFF, H. (Hrsg.): *Electronic Speech Signal Processing 2010*, Bd. 58 d. Reihe *Studientexte zur Sprachkommunikation*, S. 204–211, Dresden, September 2010. TUDpress.

[2] *[chrome] View of /trunk/src/content/browser/speech/google_one_shot_remote_engine.cc*. `http://src.chromium.org/viewvc/chrome/trunk/src/content/browser/speech/google_one_shot_remote_engine.cc?view=markup`. Abgerufen: 27/04/2012.

[3] KÜHNEL, C.: *Quantifying Quality Aspects of Multimodal Interactive Systems*. Doktorarbeit, Technische Universität Berlin, November 2011.

[4] SAMPATH, S. und B. BRINGERT: *Speech Input API Specification. Editor's Draft 18 October 2010*. `http://www.w3.org/2005/Incubator/htmlspeech/2010/10/google-api-draft.html`, October 2010. Abgerufen: 26.04.2012.

[5] SCHMIDT, S., K.-P. ENGELBRECHT, M. SCHULZ, M. MEISTER, J. STUBBE, M. TÖPPEL und S. MÖLLER: *Identification of interactivity sequences in interactions with spoken dialog systems*. In: *Proccedings of Thrid International Workshop on Perceptual Quality of Systems*, S. 109–114. Chair of Communication Acoustics TU Dresden, 2010.

[6] SCHMIDT, S., M. SCHULZ, M. MEISTER, J. STUBBE, M. TÖPPEL, K.-P. ENGELBRECHT und S. MÖLLER: *Identifikation von Interaktivitätssequenzen zur regelbasierten Usability-Evaluierung von Sprachdialogsystemen*. In: MIXDORFF, H. (Hrsg.): *Electronic Speech Signal Processing 2010*, Bd. 58 d. Reihe *Studientexte zur Sprachkommunikation*, S. 188–195, Dresden, Germany, September 2010. TUDpress.

[7] SIMPSON, A. und N. M. ERASER: *Black box and glass box evaluation of the SUNDIAL system*. In: *EUROSPEECH'93*, S. 1423–1426, Berlin, Germany, 1993.

[8] *Android Developers, class documentation of SpeechRecognizer*. `http://developer.android.com/reference/android/speech/SpeechRecognizer.html`. Abgerufen: 27/04/2012.