SPRACHSTEUERUNG FÜR MOBILE ORTSBASIERTE INTERNET-DIENSTE

Carsten Günther (HDM-I), Mathias Jöst (HDM-I),

Marion Mast (EML), Markus Klehr (EML)

Heidelberg mobil International GmbH (HDM-I)

European Media Laboratory GmbH (EML)

Schloss-Wolfsbrunnenweg 33

D - 69118 Heidelberg

{guenther|joest}@hdm-i.com
{mast|klehr}@eml-d.de

Kurzfassung: Mit der Verbreitung mobiler Internet-fähiger Geräte und neuen Tarifstrukturen steigt die Nutzung des mobilen Internet rasant an. Damit einher geht die steigende Bedeutung weiterer Technologien, die zudem wesentlich voneinander profitieren können: zum einen die ortsbasierten Dienste zum situationsgerechten Finden und Auswählen von Informationen und zum anderen die Sprachtechnologien für eine nutzerfreundliche Bedienerführung mobiler Dienste und Geräte. In diesem Beitrag wird die Integration von Voice-Search-Technologien in eine laufende Stadt-Portal-Lösung vorgestellt (speziell „vascmo“). Das System beruht auf einer Client-Server-Infrastruktur, wobei auf dem Server ein Transkriptionssystem läuft, um eine robuste kontinuierliche Spracherkennung für touristische Suchanfragen sicher-zustellen.

1 Einleitung

Mit der Verbreitung von Smartphones – allen voran Apple's iPhone – haben in den vergangenen zwei Jahren mobile Dienste Einzug in den Massenmarkt gehalten. In der aktuellen Entwicklung sind dabei zwei generelle Trends zu beobachten, zum Einen die „Mobilisierung“ von stationären Internet-Auftritten und zum Anderen die Proliferationen von mobilen Anwendungen und dedizierten Downloadplattformen für die unterschiedlichen Geräteplattformen und Betriebssysteme. Für beide Entwicklungen finden sich mittlerweile Tausende von unterschiedlichen mobilen Diensten für nahezu jede Facette digitaler Dienste, z.B. in den Bereichen Spiele, Handel, Wissen, Unterhaltung oder Reisen. Analog zu klassischen Desktopanwendungen sind diese Dienste zumeist auf eine graphische Interaktion ausgerichtet, wogegen doch gerade in mobilen Szenarien das natürlichste Interaktionsmedium die Sprache wäre. Nur wenige Sprachapplikationen wie beispielsweise die mobile Sprachsuche von Google oder Yahoo sind für mobile Szenarien bisher umgesetzt worden.

2 Überblick des Anwendungsfalls (Use-Case)

Im Folgenden diskutieren wir die Erweiterung des Internet-Reisedienstes „vascmo“ (www.vascmo.de) um die Aspekte der Sprachsteuerung und Sprachsuche. Der Reisedienst „vascmo“ ist eine neue Plattform für die einheitliche Bereitstellung ortsbasierter Informationen mit touristischem Fokus für alle Städte Deutschlands. Als konvergenter Dienst ist „vascmo“ sowohl auf stationären Ausgabegeräten wie einem Desktop-PC oder Internet-Terminals als auch insbesondere für mobile Endgeräte verfügbar. Abbildung 1 zeigt

verschiedene Ansichten, wobei die Desktop-Ansicht bereits den mobilen Auftritt vorwegnimmt, um eine einfache Nutzer-Konversion von der stationären zur mobilen Nutzung zu erzielen.

Ein wesentlicher Aspekt für die Akzeptanz mobiler Dienste, vor allem solcher mit einem Ortsbezug, ist der Abdeckungsgrad und die Aktualität der Inhalte. Nichts ist für einen Nutzer enttäuschender als auf Basis einer falschen Empfehlung einen Ort besucht zu haben, der dann im schlimmsten Fall nicht mehr existiert oder gerade nicht geöffnet hat. Aus diesem Grund wird „vascmo“ mit Inhalten eines zentralen Datenaggregationsdienstes gespeist. Der Dienst bezieht Inhalte von überregionalen, regionalen und lokalen Anbietern und aggregiert bzw. reichert diese mit weiteren Kontextinformationen an. Um einen nutzerfreundlichen Zugriff auf diese Datenfülle im mobilen Umfeld zu ermöglichen, wird die Integration einer Sprachschnittstelle im Folgenden beschrieben. Im Hinblick auf einen sprachbasierten Dienst stellen große Datenmengen dynamischer Inhalte erhebliche Anforderungen an das Sprachmodell und den Spracherkenner.

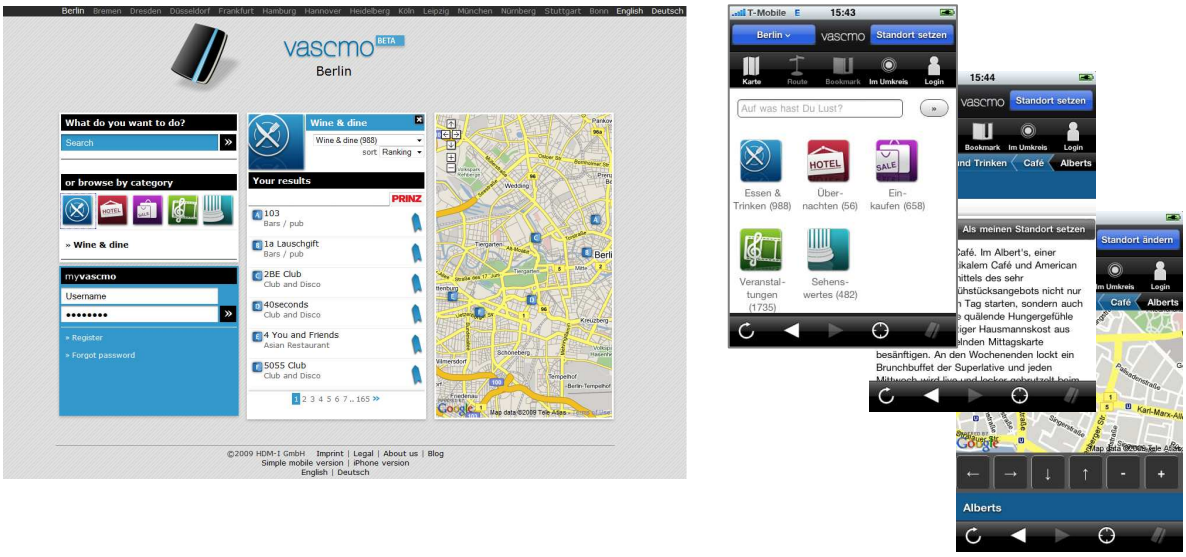

Abbildung 1: Ansichten des „vascmo“-Portaldienstes (Desktop vs. Mobil)

In der derzeitigen Umsetzung ermöglicht die Spracherweiterung des „vascmo“-Dienstes eine namentliche Suche nach Veranstaltungen und Points of Interest (POIs) im Datenbestand. Zu diesem Zweck installiert der Nutzer eine mobile Applikation auf seinem Smartphone, welche die Brücke zwischen dem reinen Internet-Dienst „vascmo“ und der neu entwickelten Sprachfunktionalität auf dem mobilen Gerät schlägt. Durch Betätigen eines Knopfes kann der Nutzer das Telefonmikrophon aktivieren und die Audio-Aufnahme starten. Nach Beendigung der Aufnahme werden die Audio-Datenpakete an die Server-Infrastruktur gesandt und die Sprachverarbeitung aktiviert. Das Ergebnis der Suchanfrage erhält der Nutzer in Form einer Ergebnisansicht innerhalb des „vascmo“-Portals.

3 Spracherkennngskomponente

Die Erweiterung des „vascmo“-Dienstes um die Sprachschnittstelle muss sprecherunabhängig und auf Grund der Domäne für ein großes Vokabular funktionieren. Der Anwendungsbereich ist zwar beschränkt auf ortsbasierte Informationen mit touristischem Fokus, allerdings kann der Benutzer eine Suchanfrage frei formulieren und ist dabei nicht auf die Eingabe einzelner (Schlag-)-Wörter beschränkt.

3.1 Technologie

Die Erkennung erfolgt wie in Abbildung 2 zu sehen in zwei Schritten mit einem sprecherunabhängen (SI) und einem sprecheradaptierten (SD) Modell [1].

1. Das sprecherunabhängige Modell verwendet kontinuierliche links-rechts Hidden Markov Modelle (HMM) mit Gauß'schen Mischverteilungen und uniformen Übergangswahrscheinlichkeiten. Jedes HMM hat drei Zustände, außer das HMM für Stille, welches nur aus einem Zustand besteht. Das deutsche Modell benutzt 51 Phone, 47 Phone für Sprache, 2 für Stille und 2 Noise-Phone.
2. Im zweiten Schritt wird zuerst eine Normalisierung der Länge des Vokaltraktes (VTLN) durchgeführt. Das sprecheradaptierte Modell ist trainiert auf Merkmalen in einem linear-transformierten Merkmalsraum, der durch die Anwendung der fMLLR (Maximum Likelihood Linear Regression im Merkmalsraum) Transformation auf die mit VTLN normalisierten Merkmale pro Sprecher entsteht.

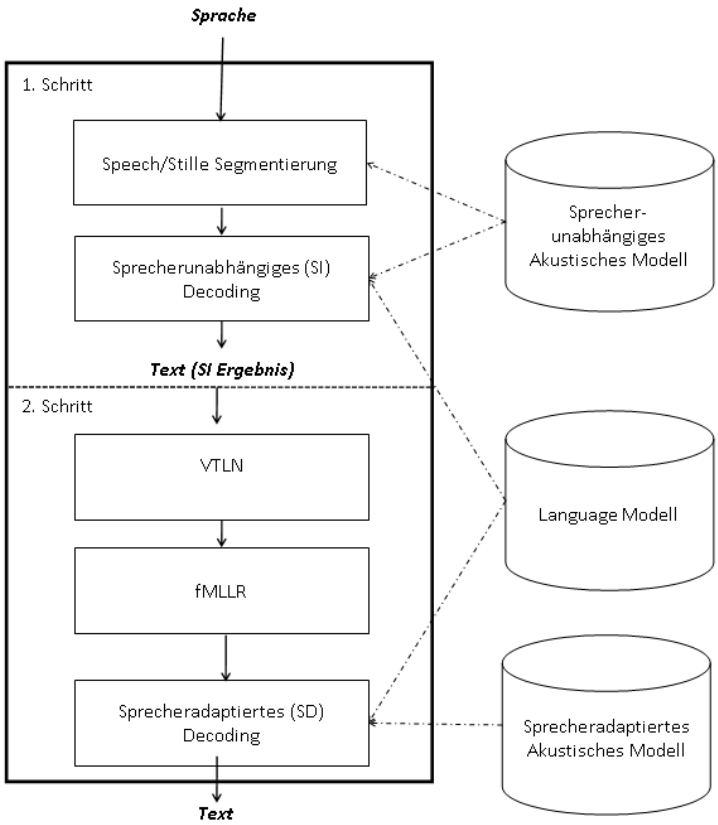

Abbildung 2: Architektur der Spracherkennungskomponente

3.2 Akustisches Modell

Das akustische Modell ist mit einigen hundert Stunden Sprache (8 kHz Telefonkanal, Sprecher: männlich, weiblich) aus unterschiedlichen Anwendungen wie z.B. Call-Center-Dialogen, VoiceMails und Self-Service-Anwendungen trainiert.

3.3 Language Modell

Als Language Modell (LM) wird ein 4-Gramm-Modell mit modifizierter Kneser-Ney-Glättung [2] verwendet. Zum Training des Language Modells wurde eine Textdatensammlung

durchgeführt und Sätze aus den Anwendungsbereichen des „vascmo“-Dienstes gesammelt (aufenthaltsortsbezogene Suche nach Veranstaltungen und POIs). Aus diesen Sätzen wurden Schablonen (Templates) erzeugt, aus denen mit den aktuellen Kategorien des Anwendungsbereichs der Korpus generiert wurde. Mit diesem initialen Korpus wurde das Language Model für ein erstes System gebildet (VS01 in Tabelle 1), um Life-Daten mit dem sprachaktivierten „vascmo“-Dienst zu sammeln. Auch diese weiteren Daten wurden zur Korpuserweiterung verwendet.

3.4 Language Modell für VoiceSearch in „Heidelberg“ und „Hannover“

Das Language Modell für „vascmo“ in Heidelberg und Hannover umfasst 2700 POIs verteilt über 20 Kategorien. Mit diesen Kategorien und POIs wurde auf der Grundlage der erstellten Schablonen ein Korpus generiert. Das aktive Vokabular wurde aus diesen Sätzen, den Kategorien und POIs abgeleitet und evolutionär erweitert (siehe Tabelle 1). Die Aussprachevarianten für das Vokabular werden automatisch generiert. Eine Schwierigkeit stellt jedoch die automatische Generierung von Aussprachen für Eigennamen und Fremdwörter da, die häufig in POIs vorkommen, da diese häufig von den Ausspracheregeln abweichen. Zur Erstellung eines Vokabulars für den Produktionsbetrieb ist eine Überprüfung der automatisch generierten Aussprachevarianten erforderlich.

LM	Vokabular	Anzahl Sätze
VS01	4000	20.000
VS02	4600	450.000
VS04	4600	1.200.000
Contact Center	150.000	7.200.000

Tabelle 1: Language Modelle

Zum Vergleich wurde ein Language Modell für den Anwendungsbereich Contact Center herangezogen, welches auf einem viel größerem Vokabular basiert, jedoch nicht die gleiche Variabilität in den Äußerungen aufweist.

4 Lösungsbeschreibung

4.1 Architektur des Gesamtsystems

Das Gesamtsystem zur Bereitstellung von sprach- und ortsbasierten mobilen Diensten auf Basis des „vascmo“-Portals setzt sich aus vier Kernkomponenten zusammen:

- einem mobilen Client,
- dem Portaldienst,
- einem Datenaggregationsdienst und
- der Spracherkennungskomponente.

Abbildung 3 zeigt das Zusammenspiel der Kernkomponenten, welche in den folgenden Abschnitten näher beschrieben werden.

4.2 Der mobile Client

Auf der Seite des mobilen Endgerätes wird der lokale Browser durch eine lokal installierte Komponente erweitert. Diese Erweiterung gestattet es, eine Sprachaufnahme zu starten und zu beenden und die aufgenommene Audiodatei via http an einen zentralen Server zu senden. Als Antwort nimmt die Komponente eine URL entgegen zur Identifizierung des gesuchten Inhalts, der dann im lokalen Browser zur Anzeige gebracht wird. Neben der Übertragung des Audiofiles wird zudem die Position des Nutzers übertragen, um auf diese Weise eine ortsbasierte Suche zu unterstützen, ohne dass der Nutzer explizit seinen Standort eingeben muss. Diese Funktion wird nur für die mobilen Endgeräte unterstützt, die ein eingebautes GPS-Modul besitzen.

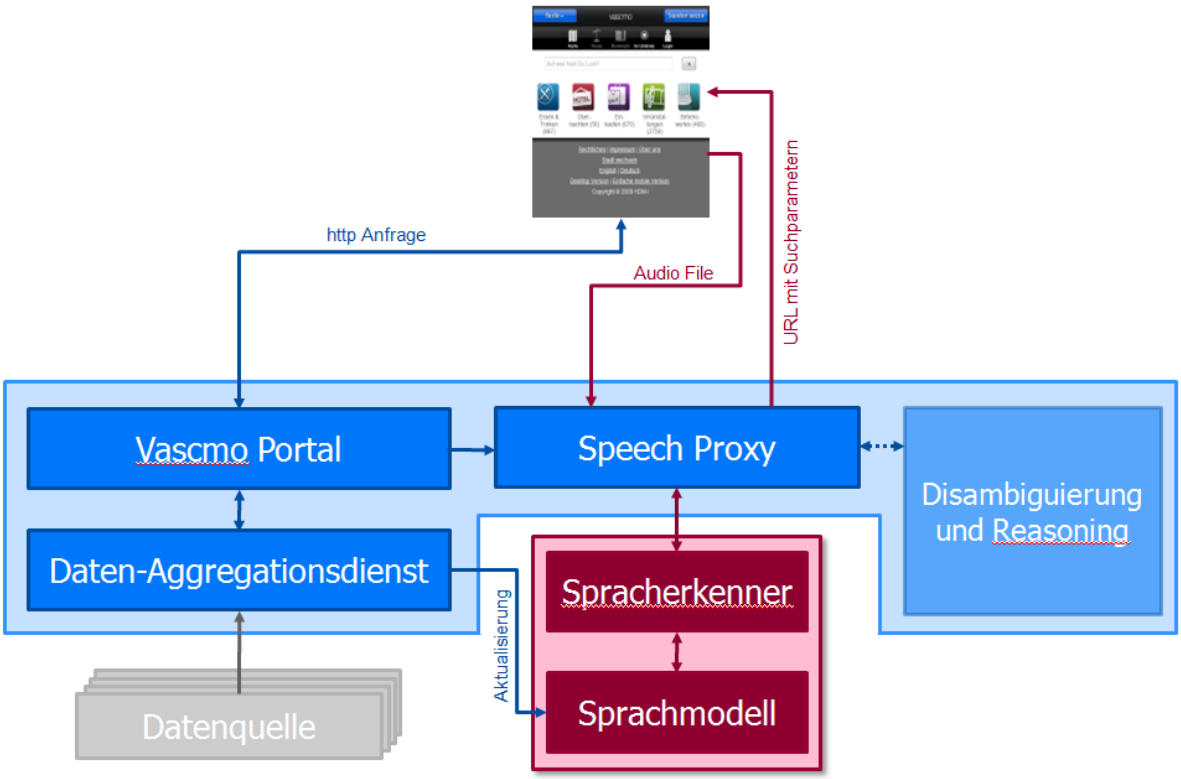

Abbildung 3: Gesamtarchitektur

4.3 "vascmo"-Portal und Speech Proxy

Als zentrale, d.h. Requests empfangende, verarbeitende und auslösende, Einheit fungiert der Portaldienst „vascmo". Diese wurde um eine Speech Proxy Komponente erweitert, die serverseitig das Audiofile entgegennimmt und an den Spracherkenner weiterleitet. Das Ergebnis der Spracherkennung wird anschließend innerhalb des Speech Proxy auf die vom „vascmo"-Portal unterstützen Suchanfragen abgebildet und eine entsprechende Suchanfrage im URL-Schema aggregiert und an die mobile Komponente übermittelt. Die generierte URL enthält zudem als weiteren Parameter ebenfalls die Ortsinformation des mobilen Nutzers. In der derzeitigen Ausbaustufe des Systems werden klassische Suchanfragen beispielsweise nach dem nächstgelegenen Ort XY bezogen auf den aktuellen Aufenthaltsort unterstützt. Im Hinblick auf die Unterstützung komplexerer Suchanfragen wie beispielsweise nach eine Veranstaltung an einem Ort zu einer bestimmten Zeit ist eine Disambiguierungs- und Reasoning-Komponente vorgesehen, die dann auf Basis der Dialoghistorie und semantischer

Informationen das Ergebnis der Spracherkennung auf die korrekte URL-Anfrage abbilden kann.

4.4 Datenaggregationsdienst

Als Datenquelle für die Bereitstellung der Inhalte des „vascmo“-Portals und für die Bildung des domänenspezifischen Vokabular dient ein Datenaggregationsdienst, der Inhalte aus verschiedenen externen Quellen aggregiert wie beispielsweise Veranstaltungsdatenbanken oder Redaktionsdiensten für touristische Inhalte. Die Aktualisierung der Daten erfolgt je nach Datenquelle in unterschiedlichen periodischen Abständen. Mit jeder Aktualisierung wird ein Abgleich des Vokabulars im Sprachmodell angestoßen, um Suchanfragen nach aktuellen Inhalten im „vascmo“-Portal zu ermöglichen. Dies ist vor allem für die dynamischen Veranstaltungsdaten wichtig.

4.5 Architektur der Spracherkennungskomponente

Abbildung 4 stellt die interne Struktur der Spracherkennungskomponente dar. Der Speech Proxy schickt an den Queue Server eine Kontrolldatei mit einer Beschreibung des durchzuführenden Transkriptionsauftrags. In dieser Kontrolldatei sind Informationen über die zu transkribierende Audiodatei, die verwendete Sprache und das einzusetzende Sprachmodell enthalten.

Der Queue Server überträgt die Kontrolldatei in die Warteschlange. Die Warteschlange besitzt ein Prioritätsmanagement, wodurch z.B. Transkriptionsaufträge von Premiumkunden bevorzugt ausgeführt werden können.

Die einzelnen Transkriptionsserver registrieren sich an der Auftragswarteschlange (Queue). Sobald ein neuer Auftrag verfügbar ist, wird die Kontrolldatei vom Transkriptionsserver aus der Warteschlange geholt und ausgewertet. Danach erstellt der Transkriptionsserver eine lokale Kopie der Audiodatei, die er über das http-Protokoll empfangen hat, und generiert dafür eine automatische Transkription.

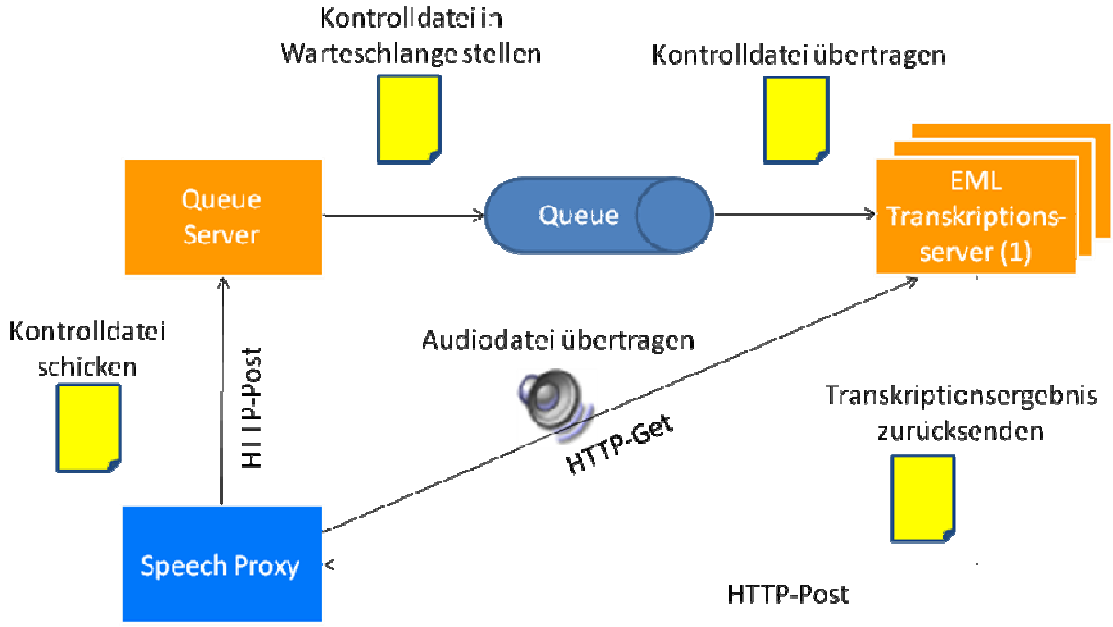

Abbildung 4: Architektur Spracherkennung

Das Ergebnis der automatischen Transkription wird der Kontrolldatei hinzugefügt und an einen http-Server, der in der Kontrolldatei angegeben ist mittels eines http-POST Requests übermittelt. Nach der erfolgreichen Übertragung des Ergebnisses wird die Durchführung des Transkriptionsauftrags der Warteschlange (Queue) bestätigt.

Für hochverfügbare und hochskalierbare Systeme kann der Queue Server redundant ausgelegt werden. Darüberhinaus können Transkriptionsserver dynamisch hinzugenommen und deaktiviert werden.

4.6 Mapping des Spracherkennungsergebnisses auf die URL-Anfrage

Die Übertragung des Spracherkennungsergebnisses auf die entsprechende in einer URL-encodierten „vascmo"-Suchanfrage folgt einem einfachen aber effizienten Abbildungs-schema. Auf Basis des trainierten Wortschatzes sind dem Spracherkenner sowohl die verschiedenen Inhaltskategorien des Portals als auch die Namen und Bezeichnungen der POIs bekannt. Wird in der Suchanfrage eines dieser Schlüsselwörter erkannt, erfolgt eine semantische Auszeichnung als besonderer Bestandteil der verbalen Äußerung. Dieser Teil der Äußerung wird zunächst einer inhaltlichen Prüfung unterzogen. Lässt er sich auf eine Kategorie innerhalb des Portals abbilden, wird innerhalb des Speech Proxys die entsprechende Kategorie-Suchanfrage erstellt.

```
http://m.vascmo.de/mobilePlace/list? viewCategory=' +
```

Ist keine direkte Abbildung auf eine Kategorie-Anfrage möglich, wird eine entsprechende Umkreissuche für den POI angestoßen.

```
http://m.vascmo.de/mobilePlace/list?act=search&cmd=poi&title=' + title;
```

Da für beide unterstützen Anwendungsfälle der Ortsbezug eminent wichtig ist, wird der URL Anfrage sowohl die aktuelle Lokation in Geokoordinaten als auch der entsprechende Suchradius angefügt. Im Hinblick auf weitere Anwendungsfälle, wie beispielsweise die Anfrage für ein Routing zu einem Zielort, wird das Abbildungsschema zukünftig um Verben erweitert, um Fußgängernavigationsanfragen zu unterstützen.

```
http://m.vascmo.de/mobilePlace/route?cmd=new_route& dest=' + title;
```

5 Spracherkennungsergebnisse

Der „vascmo"-Dienst wurde mit den in Abschnitt 3 beschriebenen Akustischen und Language Modellen getestet. Insgesamt wurden 143 Anrufe mit VoiceSearch-Anfragen ausgewertet. Zum Vergleich wurde aus den in Kapitel 3.4 verwendeten Schablonen eine Grammatik erzeugt und mit einem Standard „state of the art" Erkenner getestet (Tabelle 2:).

Einschränkend ist zu erwähnen, dass bei der Verwendung der Grammatik nicht das gleiche Akustische Modell verwendet werden konnte, so dass der Vergleich nur eine Tendenz anzeigt. Die Ergebnisse legen nahe, dass ein Spracherkennungsprozess, der auf einem Sprachmodell basiert, für eine VoiceSearch-Anwendung wie „vascmo" speziell bei den

domänenbedingten unbekannten Formulierungen bessere Ergebnisse liefert als ein grammatikbasierter, und somit ein robusteres Systemverhalten aufzeigt.

Gegenüber einem ‚ContactCenter' Sprachmodell ohne Anpassung an die Voice-Search Domäne, konnte die Anzahl korrekter Wörter um 50 % gesteigert werden, was die Bedeutung von Sprachmodelltrainingsdaten aus der Domäne und die Anpassung des Vokabulars hervorhebt. Schon ein Sprachmodell (VS01) mit einer kleinen Menge an Trainingsdaten aus der Domäne, ist einem Modell, welches für eine andere Domäne (out-of-the-box ContactCenter) entwickelt wurde, überlegen.

Language Model	**Word correct (143 Sätze, 668 words)**	**WER (143 Sätze, 668 words)**	**Anzahl unbek. Wörter**	**Perplexität**
VS01	55.2	66.5	132	114
VS02	76.2	44.3	23	30
VS04	72.8	48.4	23	43
ContactCenter	49.9	64.5	115	136
Grammatik	40.1	63.9	23	-

Tabelle 2: Spracherkennungsergebnisse

6 Ausblick

Die oben angeführten Ergebnisse weisen die Spracherkennungsgenauigkeit auf Wortebene aus. Für den Nutzer ist jedoch alleinig die Erfolgsrate von Interesse, mit der er die korrekte Information erhält (basierend auf „task success rate" und „task completion rate"). Hierfür ist eine weitergehende manuelle Auswertung erforderlich

Ein weiterer Aspekt, den die Ergebnisse noch nicht reflektieren, ist die Variabilität der möglichen Suchanfragen im Life-Betrieb. So können Anfragen zu einem Thema auf sehr unterschiedliche Weisen formuliert werden. Um sich über Veranstaltungen zu informieren, könnte man etwa folgendes äußern: „Wie kann ich mich heute in Heidelberg amüsieren?", „Was ist heute in Heidelberg los?", „Was ist heute in Heidelberg angesagt?", „Welche Veranstaltungen gibt es heute in Heidelberg?". Hier genügt es nicht, sich auf Schlüsselwörter oder Phrasen zu beschränken. Um diese Äußerungen korrekt zu beantworten, ist eine weitergehende linguistische Analyse erforderlich.

Literatur

[1] Ramabhadran, B.; Siohan, O.; Mangu L.; Westphal, M.; Schulz, H.; Soneiro, A.: The IBM 2006 Speech Transcription System for European Parliamentary Speeches. In: TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain: June 19–21, 2006.

[2] Kneser, R. und H. Ney (1995) Improved backing-off for m-gram language modeling. In: *IEEE ICASSP-95*, Vol. 1, 181-184