Sprachaufnahmen über das WWW – Diskussion der aktuellen Technologie und eine Prototyp-Implementation

Christoph Draxler
Institut für Phonetik und Sprachliche Kommunikation
Universität München
Schellingstr. 3
80799 München
draxler@phonetik.uni-muenchen.de

Abstract

Die Signalqualität von Sprachaufnahmen ist abhängig von der geografischen Verteilung dieser Aufnahmen. Aufnahmen vor Ort erfolgen in hoher technischer Qualität, aber sowohl Sprecher als auch die Aufnahmeausrüstung müssen sich am selben Ort befinden. Telefonaufnahmen erlauben die geografische Verteilung von Sprachaufnahmen, aber ihre Qualität ist auf die Bandbreite des Telefonkanals beschränkt.

Sprachaufnahmen über das WWW entkoppeln die Signalqualität von der geografischen Verteilung der Aufnahmen: Sprachsignale werden mit einem Standard WWW-Browser vor Ort aufgenommen und als Datenpakete über das WWW an einen Sprachserver übertragen. Das BAS hat mit *WebRecorder* ist eine Prototyp-Implementation für Sprachaufnahmen über das WWW entwickelt, die im Projekt RVG-J zum ersten Mal in der Praxis eingesetzt werden wird.

1 Einführung

Der Aufbau großer Sprachdatenbanken ist zeitaufwändig und teuer. Bei Sprachaufnahmen in einem Aufnahmestudio oder im Feld beim Sprecher ist die technische Qualität der Aufnahmen in der Regel hoch, da alle Aufnahmeparameter überwacht werden können. Allerdings ist der Aufwand für solche Aufnahmen ebenfalls hoch, da entweder die Ausrüstung zum Sprecher oder der Sprecher zur Ausrüstung gebracht werden muss (so hat z. B. das BAS für Sprachaufnahmen Schulklassen aus ganz Deutschland nach München eingeladen).

Bei Sprachaufnahmen über das Telefon rufen Sprecher von unterschiedlichen Aufnahmeorten aus einen Telefonserver an. Die Signalqualität dieser Aufnahmen ist allerdings gering, denn diese ist durch den Übertragungskanal (analoges, ISDN- oder GSM-Netz, nur Monosignale) vorgegeben.

Sprachaufnahmen über das WWW entkoppeln die Signalqualität der Sprachaufnahmen vom Aufnahmeort und vereinen somit die Vorteile beider Verfahren: natürliche Sprachaufnahmen bei gleichzeitig hoher Signalqualität.

2 Technologie

Das WWW ist ein sog. Client-Server System: der Client, ein Webbrowser, fordert Seiten von einem Webserver an. Der Server lädt die Seite aus dem lokalen Dateisystem und schickt das Dokument zur Ausgabe an den Client.

2.1 HTML und XML

WWW-Dokumente sind in HTML *(Hypertext Markup Language)* geschrieben. HTML ist eine *Anwendung* von SGML, d.h. die Menge der erlaubten Tags ist vorgegeben (Ragett et al., 1998). XML *(Extensible Markup Language)* dagegen ist eine *Untermenge* von SGML. XML erlaubt die freie Definition von Tags für unterschiedliche Dokumentklassen und Anwendungen. Die zulässigen Tags und ihre Beziehungen untereinander sind in einer DTD *(document type description)* bzw. einem XML-Schema spezifiziert (Connolly and Thompson, 2000). Damit ist es möglich, Auszeichnungssprachen für einzelne Anwendungsgebiete zu spezifizieren.

2.2 VoiceXML

VoiceXML ist eine Auszeichnungssprache für sprachgesteuerte Dienste und Anwendungen. VoiceXML wurde ursprünglich für Internet basierte Telefonieanwendungen wie Verzeichnisdienste, Informationssysteme, Anrufbeantworter usw. entwickelt (McGlashan et al., 2001).

Ein VoiceXML-Dokument definiert den Ablauf einer sprachgesteuerten Internetanwen-

dung. Ein Benutzer ruft einen VoiceXML-Dienst an und startet damit den VoiceXML-Interpreter. Dieser Interpreter steuert die Sprachsynthese und -erkennung zur Kommunikation mit dem Benutzer und er implementiert den Zugriff auf die für den eigentlichen Dienst notwendigen Daten. In klassischen VoiceXML-Anwendungen wird das Endgerät ausschließlich für die Sprachein- und -ausgabe verwendet, die Sprachsignale werden über das Telefonnetz übertragen und der Interpreter läuft auf den Servern des Diensteanbieters.

2.3 Applets

Applets sind Anwendungsprogramme, die in eine Webseite integriert sind. Sie erweitern Webseiten um Funktionalität, die in HTML nicht verfügbar ist, z.B. dynamische Grafiken o.ä. Applets werden wegen der Plattform Unabhängigkeit dieser Sprache üblicherweise in Java implementiert. Der Programmcode wird vom Server heruntergeladen und in der *Sandbox*-Umgebung ausgeführt, die den Zugriff auf Systemressourcen überwacht. Applets können die volle Rechenleistung eines Clientrechners nutzen und somit auch rechenaufwändige Aufgaben wie Signalverarbeitung implementieren.

2.4 Datenübertragung

Das Standardprotokoll zur Datenübertragung im WWW ist `http` *(hypertext transport protocol)*. `http` spezifiziert keinerlei zeitliche Randbedingungen für die Über-tra-gung von Datenpaketen, gewährleistet aber deren vollständige Übertragung.

Echtzeit-Übertragungen von Multimedia Inhalten über das Internet wie z.B. Internet-TV oder Voice-over-IP basieren auf `rtsp`, dem *real time streaming protocol*. Die Multimediadaten werden hierbei stark komprimiert, was nur mit verlustbehafteten Verfahren möglich ist. Für Sprachdatensammlungen sind derart komprimierte Sprachsignale von geringem Interesse.

3 WebRecorder

WebRecorder ist eine am BAS entwickelte Prototyp-Implementation für Sprachaufnahmen über das WWW auf der Basis von VoiceXML. Im Gegensatz zu klassischen VoiceXML-Diensten läuft der VoiceXML Interpreter in einem Applet auf einem WWW Clientrechner und nutzt dessen Rechenleistung. Der Clientrechner nimmt die Sprachsignale in hoher Qualität auf konvertiert sie verlustfrei in Datenpakete und überträgt diese via `http` in einem Hintergrundprozeß an den Server.

Die Steuerung der Aufnahmesitzung erfolgt durch das Applet. Beim Programmstart initialisiert es auf dem Server eine Aufnahmesitzung und fordert ein VoiceXML-Formular für die Sprachaufnahme an. Für jede Aufnahme wird ein eigenes Formular verwendet. Sobald das Sprachsignal aufgenommen wurde, werden die Signaldaten in Form eines mehrteiligen Dokuments (sog. Multipart document) an den Server übertragen. Die einzelnen Bestandteile des Dokuments enthalten administrative Angaben, z.B. Dateinamen, Angaben zur Aufnahmesitzung, usw. und die eigentlichen Sprachdaten. Der Server speichert diese Angaben und schickt entweder ein neues VoiceXML Formular oder die Nachricht, dass die Aufnahmesitzung beendet sei.

3.1 Interface

Die Benutzerschnittstelle von WebRecorder ist grafisch. Eine Ampel auf der linken Seite gibt den Aufnahmestatus an: keine Aufnahme, bereithalten, Aufnahme. Das obere Textfeld enthält Anweisungen an den Sprecher, z.B. „Lesen Sie den Satz“, „Beantworten Sie die Frage“, usw. Der Prompttext steht im großen Textfeld in der Mitte (Abb. 1). Die grafischen Möglichkeiten von Java (AWT oder Swing) erlauben auch die Ausgabe von Bildprompts.

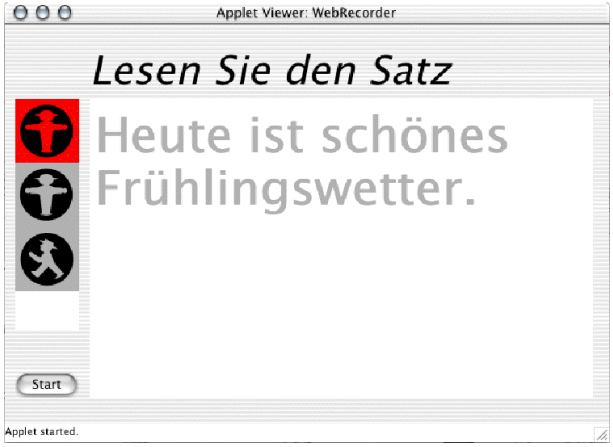

Abbildung 1: WebRecorder in AppletViewer

3.2 Signalaufnahme

Für die Implementation der Sprachaufnahme stehen zwei Alternativen zur Auswahl: JavaSound und QuickTime for Java (Maremaa and Stewart, 1999). JavaSound ist ein Standard Java-API, aber es wurde es erst ab der Java Version 1.3 verfügbar und wird nur von sehr wenigen Plattformen unterstützt. QuickTime for Java ist für Windows und Macintosh verfügbar, es ist zu Java ab Version 1.1 kompatibel, es wird von vielen Audiokomponenten unterstützt und mit der Standard QuickTime-Distribution installiert. Die aktuelle Version von WebRecorder basiert daher auf QuickTime for Java.

4 Diskussion

Die Aufnahme von Sprachsignalen über das WWW ist ein völlig neues Konzept. Daher müssen die sozialen, administrativen und technischen Konsequenzen diskutiert werden.

4.1 Sprecheridentität und Zustimmung zur Aufnahme

Clientrechner im Internet haben eine weltweit eindeutige Kennung, sie können daher im Netz identifiziert werden. Für den Benutzer eines Rechners gilt dies nicht notwendigerweise.

Eine Möglichkeit, die Identität eines Sprechers zu gewährleisten ist die explizite Anmeldung zu Sprachaufnahmen und die damit verbundene Vergabe eines Passworts. Vor Beginn einer Aufnahmesitzung muss sich ein Sprecher einloggen und seine Zustimmung zur Sprachaufnahme bestätigen, z.B. durch Drücken eines „ich akzeptiere“-Buttons.

4.2 Audiokomponenten

Die Audiokomponenten moderner PCs sind, was die Soft- und Hardwareschnittstellen angeht, weitgehend standardisiert. Die Hardware selber ist weit weniger standardisiert und sehr häufig ist keine Information über die technischen Details einer verwendeten Soundkarte verfügbar. Das gilt in der Regel auch für das Mikrofon und die Audioausgabe. Es ist daher nicht realistisch anzunehmen, dass der Benutzer technische Angaben zu den verwendeten Audiokomponenten liefern kann.

Um unter diesen Bedingungen zumindest einen Teil der Ausrüstung kontrollieren zu können, kann man den Sprechern Mikrofone für die Sprachaufnahmen schicken (das ist natürlich noch keine Garantie, dass diese auch verwendet werden). Auch ist es möglich, Sprecher und ihre Ausrüstung mit Testaufnahmen zu prüfen und erst dann für die eigentlichen Aufnahmen zuzulassen, wenn die technische Qualität ausreichend ist – diese Prüfung kann bereits auf dem Clientrechner erfolgen.

4.3 Aufnahmeumgebung

Clientrechner befinden sich in den verschiedensten Umgebungen: in Privaträumen, Büros, an öffentliche Orten oder gar in mobilen Umgebungen. Es ist daher wichtig, vom Benutzer eine minimale Beschreibung der aktuellen Umgebung zu verlangen. Auch sollte ausreichend viel Umgebungsgeräusch aufgenommen werden. Dies kann entweder vor Beginn einer Aufnahmesitzung oder am Ende erfolgen, oder, wie z.B. in SpeechDat-Car oder RVG-J, bei jeder einzelnen Äußerung (Draxler et al., 1999), (Draxler and Schiel, 2002).

4.4 Sicherheit

Applets installieren keine Software dauerhaft auf dem Clientrechner. Allerdings benötigen sie Zugriff auf Systemressourcen, z.B. die Audiokomponenten und temporären Speicher. Der Browser muss daher so konfiguriert werden, dass Applets dieser mindestnotwendige Zugriff auf den lokalen Rechner erlaubt wird.

4.5 Übertragungsgeschwindigkeit

Internetverbindungen mit geringer Übertragungskapazität benötigen einen Datenpuffer auf dem Client. Dieser Puffer speichert die Daten temporär solange bis sie auf den Server übertragen werden konnten. Dieser Puffer wird auch benötigt um Verbindungsunterbrechungen zu überbrücken.

Die Verbindung zum Server muss solange aufrechterhalten werden wie noch nicht alle Daten übertragen wurden. Dies kann bei langsamen Verbindungen länger dauern als die eigentliche Aufnahmesitzung. In diesem Fall muss der Benutzer informiert werden, dass die Datenübertragung noch nicht abgeschlossen ist. Dies gilt auch bei Unterbrechungen der Übertragung; hier ist es aus ökonomischen Gründen sinnvoll, eine unterbrochene Übertragung einfach fortsetzen zu können.

5 Das Projekt RVG-J

RVG-J (Regional Variants of German - Junior) ist eine Erweiterung RVG Korpus. Dieses wurde vom BAS in Kollaboration mit AT&T erstellt (S. and Schiel, 1998), (Draxler and Schiel, 2002).

Das aufgenommene Sprachmaterial besteht aus den ursprünglichen 45 RVG Prompts (Ziffern, Zahlen, phonetisch reiche Wörter, spontane Äußerungen), plus 83 zusätzliche Items aus dem SpeechDat-II Korpus. Die neuen Prompts zur Evozierung von Spontansprache fordern den Sprecher auf, einfache Aufgaben zu lösen, die für die Entwicklung von Sprachtechnologie von Interesse sind, z.B. Nachrichten für Anrufbeantworter aufnehmen, ein Auskunftssystem nach einem Namen fragen, oder eine Wegbeschreibung geben.

In der ersten Phase des Projekts wurden 185 jugendliche Sprecher aus dem Großraum München aufgenommen. In den weiteren Phasen werden diese Aufnahmen auf den gesamten deutschsprachigen Raum einschließlich der Schweiz und Österreich ausgedehnt.

RVG-J ist eine ideale Testumgebung für Webbasierte Aufnahmen. Für diese Aufnahmen wird das BAS zuerst Schulen in ganz Deutschland ansprechen, denn Schulen verfügen über technisch versierte Lehrer, an Technologie interessierte Schüler und schnelle Internetverbindungen. Als Anreiz für die Teilnahme kann den Schulen ein Teil der Ausrüstung gestellt werden. Zu einem späteren Zeitpunkt können dann auch Einzelpersonen an den Aufnahmen teilnehmen.

Das BAS ermuntert Institutionen an RVG-J teilzunehmen. Es stellt die Software zur Verfügung und bietet logistische Unterstützung. Die Annotation der Sprachsignale erfolgt am BAS oder der Partnerinstitution, z.B. mittels der Software WWWTranscribe (Draxler, 1997) Das BAS distribuiert die Daten und beteiligt die Projektpartner an den Lizenzeinnahmen.

6 Zusammenfassung und Ausblick

WebRecorder ist eine Prototyp-Implementation eines Systems zur Sprachaufnahme über das WWW. Die Software befindet sich zur Zeit noch im Entwicklungsstadium und sie wird im aktuellen Projekt RVG-J zum ersten Mal in größerem Maßstab eingesetzt. Die Weiterentwicklung von WebRecorder konzentriert sich auf die Präsentation von Multimedia-Prompts, z.B. Audio und Video und auf die Optimierung des Datentransfers zum Server.

7 Danksagung

Ein Teil dieser Arbeit wurde mit Mitteln des BITS (BAS Infrastructure for Technical Speech processing) Projektes des BMBF (Kennzeichen #01IVB01) gefördert.

References

D. Connolly and H. Thompson. 2000. XML schema. Technical report, http://www.w3c.org/XML/Schema.

Chr. Draxler and F. Schiel. 2002. Three new Corpora at BAS – and a First Step Towards Distributed Recording. In *Proc. of 3rd Intl. Conference on Language Resources and Evaluation*, Gran Canaria.

Chr. Draxler, R. Grudszus, St. Euler, and K. Bengler. 1999. First Experiences of the German SpeechDat-Car Database Collection in Mobile Environments. In *Proc. of Eurospeech*, Budapest.

Chr. Draxler. 1997. WWWTranscribe – a Modular Transcription System Based on the World Wide Web. In *Proc. of Eurospeech*, Rhodes.

T. Maremaa and W. Stewart. 1999. *QuickTime for Java*. Morgan Kaufman, San Diego.

S. McGlashan, D. Burnett, P. Danielsen, J. Ferrans, A. Hunt, G. Karam, D. Ladd, B. Lucas, B. Porter, K. Rehor, and S. Typhonas. 2001. Voice extensible markup language 2.0. Technical report, http://www.w3c.org/TR/voicexml20.

I. Ragett, A. Le Hors, and I. Jacobs. 1998. HTML 4.0 Specification. Technical report, http://www.w3c/TR/REC-html40.

S. Burger S. and F. Schiel. 1998. RVG 1 – A Database for Regional Variants of Contemporary German. In *Proc. of the 1st Intl. Conf. on Language Resources and Evaluation*, pages 1083–1087, Granada.