Mobiles Aufnahmesystem für Sprachtechnologie-Anwendungen

Uwe Koloska, Diane Hirschfeld, Michael Albrecht

voiceINTERconnect GmbH, 01067 Dresden
http://www.voiceinterconnect.de/
{koloska,hirschfeld}@voiceinterconnect.de

Anwendungen der Spracherkennung und Sprachsynthese erfordern zumeist große Mengen an Datenmaterial. Der Beitrag stellt die Alternativen zur Beschaffung eines Sprachkorpus gegenüber und beschreibt die konkrete Implementierung eines Workflows zur Aufnahme und Verarbeitung großer Mengen von Sprachdaten. Außerdem wird beispielhaft die Aufnahme eines Trainingswortschatzes für einen Kommandoworterkenner skizziert.

Motivation

Anwendungen der Spracherkennung und Sprachsynthese erfordern zumeist große Mengen an Datenmaterial verschiedener Sprecher, das zudem unter Bedingungen aufgenommen sein sollte, die den realen Bedingungen des Anwendungs- bzw. Produktszenarios möglichst nahe kommen.
Für die Aufnahme von Trainingsdaten für einen kleinen und schnellen Worterkenner wurde sehr schnell eine speziell definierte Menge von Worten von verschiedenen Sprechern benötigt.

Alternativen

Aufgrund der dargestellten Kriterien wurden mögliche Alternativen für eine konkrete Umsetzung evaluiert. Bei verfügbaren Korpora stehen dem Vorteil der schnellen Verfügbarkeit und der schon vorgenommenen Aufbereitung einige schwerwiegende Nachteile gegenüber: Die Daten sind nicht auf das aktuelle Problem zugeschnitten, oft in ungeeigneten Datenformaten (z. B. unterschiedlicher Phonemsatz bei den Labeln) oder die gerade benötigte Information (wie Silbenlabel oder Wortklassen – je nach Anwendung) ist nicht enthalten und meistens ist der Erwerb mit hohe Kosten verbunden.
Die Alternative ist eine eigene Aufnahme und Bearbeitung der benötigten Daten. Unbestreitbarer Vorteil ist die genaue Abstimmung der Daten und daraus abgeleiteten Informationen auf die Anwendung, was aber den Nachteil des erhöhten Zeit- und Ressourcenbedarfs hat. Um in eigener Regie einen Sprachkorpus aufzunehmen und zu bearbeiten, ist eine geeignete Infrastruktur notwendig, die entweder fertig bezogen oder selbst zusammengestellt werden kann. Da die Datenbeschaffung und Aufbereitung eine Kernkompetenz in der Sprachverarbeitung darstellt, ist der Aufbau eines geeigneten Systems in Verbindung mit einer Aufbereitungsstrategie eine wichtige Voraussetzung für zukünftige Aufgaben.

Abbildung 1 Das mobile Aufnahmesystem. Vorne das Rack mit der Audiohardware hinten das iBook.

Kriterien

Bei der Auswahl einer passenden Aufnahmeumgebung galt es verschiedene Randbedingungen einzuhalten, um die Flexibilität und Mobilität zu sichern:

- *Aufnahme und Verarbeitung auf digitaler Ebene*, da die Daten in digitalisierter Form gebraucht werden und Zwischenschritte im Datenfluß, die den Arbeitsaufwand erhöhen und die Signalqualität beeinträchtigen können, minimiert werden sollten.
- Außerdem kann dadurch eine *gleichbleibend hohe Qualität der Aufnahmen* gewährleistet werden.
- *Kompakte und robuste Komponenten* sind wichtig, weil an wechselnden Orten (z. B. in einem fahrenden Auto) aufgenommen werden soll.
- Zur Forderung nach größtmöglicher Unabhängigkeit gehört auch eine einfache *Möglichkeit der Datensicherung*
- und eine *einfache Erweiterung des Massenspeichers*.
- Besonderes Augenmerk wurde auf die Forderung nach einer *Einbindung des Datenflußes in eine automatische Umgebung* zur Verwaltung und Bearbeitung von Sprachdaten gelegt, was konkret heißt, daß die verwendeten Programme sich einfach in einen bestehenden Workflow einfügen müssen (durch offene Quellen, Skriptsteuerung oder Einbindung von Bibliotheken).
- Sehr wichtig ist auch eine *gute Reproduzierbarkeit der Aufnahmesituation*, um z. B. nachträgliche Aufnahmen von fehlenden Worten zu ermöglichen.

Realisierung

Die anfangs nur zur Einordnung der angebotenen Fertiglösungen angestellten Recherchen zu preiswerteren und zeitgemäßen Alternativen, führten schnell zu einer eigenen Lösung. Basis ist ein 8 kanaliges A/D-Interface mit Firewire Schnittstelle. Die Alternativen zur Firewire basierten Audiohardware waren: PC-Card (deutlich höherer Preis, Beschränkung auf 4 Eingänge und zwei Ausgänge, AD-Wandlung im Rechner), USB-Audiointerface (geringer Datendurchsatz des USB Busses, ebenfalls nur 4 Eingänge und eingeschränkte Vollduplexfähigkeit), PCI-Bus Lösung entweder via PC-Card nach PCI-Bus Wandler oder mit einem Industrie-PC statt einem Laptop (deutlich höherer Preis und Aufwand als bei den anderen Lösungen).
Bei den Firewire Lösungen waren zum Zeitpunkt der Recherche (Dez. 2001) etliche Geräte angekündigt, deren Preise und Liefertermine aber derart unklar waren, daß wir uns für eine Lösung mit dem schon länger am Markt befindlichen Motu 828 entschieden.
Durch die Konzentration auf Firewire war die Plattformfrage (Windows basierter PC oder Apple Mac) offen. Wegen der stabileren Unterstützung für Firewire und der ausgereifteren Audiosoftware, entschieden wir uns für ein Apple iBook in seiner Combo Variante (mit integriertem CD-Brenner/DVD-Rom Laufwerk), daß daneben noch den unschätzbaren Vorteil hat, auch nach mehrstündigem Betrieb nahezu lautlos zu laufen.
Da das Audiointerface nur Ein- und Ausgänge mit Line-Pegel hat, ist zum Betrieb der drei Mikrofone und des Talkback-Kanals (die Sprecherin erhält hier Informationen wie z. B. Tonhöhe, Worttrenner oder Regieanweisungen) ein Mikrofonvorverstärker mit getrennt schaltbarer Phantomspeisung nötig. Eingebaut in ein Rack mit vier Höheneinheiten besteht das mobile Studio aus dem Rack, einem Koffer für Kabel und Mikrofone und dem iBook. Die Datensicherung erfolgt über den eingebauten CD-Brenner des iBook.

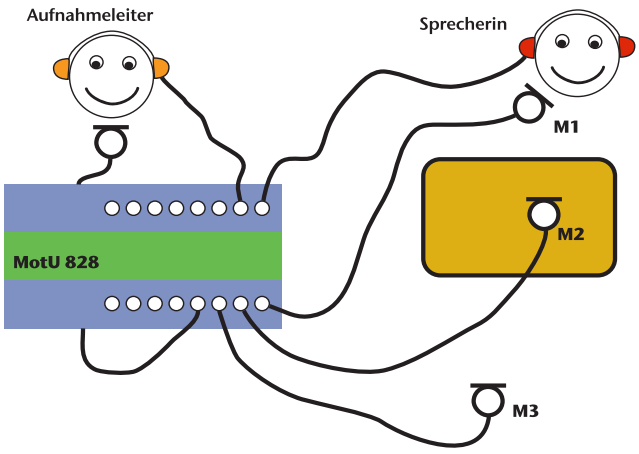

Abbildung 2 Die Situation bei der Aufnahme der Kommandowörter.

Aufnahmen für einen Einzelwortkorpus

Der erste Einsatz des mobilen Audiostudios waren die Aufnahmen für einen zweisprachigen Korpus mit Kommandowörtern. Nach der Festlegung der Wortlisten für die beiden Sprachen Deutsch und Englisch wurde das mobile Studio wie in Abbildung 2 in einem Aufnahmeraum aufgebaut. Die Sprecherin erhält über einen offenen Kopfhörer (damit sie sich selbst hören kann) ein Piepssignal, daß die einzelnen Worte trennt und zur automatischen Wortsegmentierung benutzt wird. Außerdem kann der Aufnahmeleiter auf diesem Kanal Regieanweisungen geben. Das Sprachsignal der Sprecherin wird mit drei unterschiedlich platzierten Mikrofonen aufgenommen und parallel zum Trennsignal auf separaten Spuren gespeichert.
Die Segmentierung der Aufnahmen erfolgt anhand der Markierungsspur vollautomatisch für alle drei Tonspuren. Obwohl die Markerspur auch schlechte oder falsche Worte markiert, erfordern Fehler bei der Aufnahme noch ein manuelles Eingreifen.

Statistik

Eine Übersicht über die Arbeiten und die entstandenen Daten bei den Aufnahmen für einen Trainingskorpus:
Sprachen 2, Deutsch und Englisch
Wortschatz 59 Worte Deutsch, 52 Worte Englisch
Sprecher 23 mit jeweils 5 Durchgängen
Einzelworte 12.775 einzelne Realisierungen
Zeitbedarf pro Sprecher zwischen 30 und 40 Minuten, insgesamt zwei Arbeitstage zu je 8 Stunden
Aufnahmen 3 Mikrofone (Headset, Tisch- und Raummikrofon) gleichzeitig in einem Studio
Nachbearbeitung Aufbereiten, Schneiden, Wortgrenzen markieren und Kontrolle ca. 6 Arbeitstage
Daten Rohdaten, 16 Bit, 48 KHz: 8,36 GB
Wortdateien und Label, 16 Bit, 48 KHz: 4,83 GB
Wortdateien und Label, 16 Bit, 16 KHz: 1,82 GB

Erfahrungen und Ausblick

Der erste Einsatz des Audiostudios war damit (unter Zeitdruck) erfolgreich – die zu Tage getretenen Mängel lassen sich in Zukunft durch die offene Konzeption leicht beheben. Es gibt z. B. noch keinen Automatismus, um die Anlage zu kalibrieren und an den Aufnahmeraum anzupassen. Die gewählten Komponenten haben durchweg überzeugt, besonders das flexible Audiorouting, die roadtaugliche *Verpackung* und das durch die Abwesenheit von Lärm und den unkomplizierten Betrieb überzeugende iBook.
In Zukunft soll mehr von dem Aufnahme- und Verarbeitungsprozeß automatisiert werden und durch eine Einbettung in das WiGE Framework [1] die Erzeugung von Inventaren für Synthese und Erkennung vereinfacht werden.

Literatur

[1] Koloska, Uwe, *Ein interaktives automatisches System zur Inventarerstellung und -optimierung*, Fortschritte der Akustik – DAGA2000, Oldenburg