Elektronische Beiträge zur Lehre in der Sprachkommunikation

Rüdiger Hoffmann, Ulrich Kordon
Technische Universität Dresden, Institut für Akustik und Sprachkommunikation

1. Einführung

Automatische Sprachkommunikation (Spracherkennung, Sprachsynthese, Sprachcodierung) gehört zu den wesentlichen Gegenständen der Ausbildung in der Informations- und Kommunikationstechnik. Natürliche Sprache, die selbst eines der wichtigsten Medien darstellt, eignet sich besonders für die Lehre mit multimedialen Hilfsmitteln. Innerhalb des europäischen Socrates/Erasmus-Programmes widmet sich das Thematische Netzwerk „Speech Communication Sciences“ der Koordinierung und Unterstützung solcher Aktivitäten mit besonderer Betonung des Internet [1]. Der vorliegende Beitrag beschreibt Arbeiten der Professur für Sprachkommunikation der TU Dresden, die in enger Verbindung mit dem Netzwerk entstanden sind.

2. Projekte

Seit 1997 entstanden die folgenden Ergebnisse:

- Multimediale Präsentation der Geschichte der Sprachsynthese unter besonderer Berücksichtigung der Entwicklungen an der TU Dresden (als CD-ROM);
- einzelne Demonstrationsmodule zur Unterstützung der Lehre hauptsächlich in den Fächern Psychoakustik und Signalverarbeitung;
- Internet-Tutorial über Aufbau und Funktionsweise eines Sprachsynthesesystems [2, 3] (unter Mitwirkung des Lehrstuhls Kommunikationstechnik der BTU Cottbus).

Das zuletzt genannte Projekt wird in Abschnitt 3 genauer vorgestellt. Aus den vorliegenden Erfahrungen entstand schließlich die Idee für ein weiteres Projekt, das die Schaffung einer breit anwendbaren Rahmensoftware für Tutorials im Bereich der Sprachtechnologie zum Ziel hat. Auf dieses Projekt wird in Abschnitt 5 eingegangen.

3. Das Tutorial zur Sprachsynthese

3.1. Aufbau eines Text-to-Speech-Systems

Moderne Sprachsynthesesysteme zeichnen sich dadurch aus, daß sie schriftsprachlichen Text (gewöhnliche ASCII-Zeichenfolge) in Sprache umsetzen, weshalb sie auch als Text-to-Speech-Systeme (TTS-Systeme) bezeichnet werden. Bei der Umsetzung des Textes in das akustische Signal muß eine Vielzahl komplexer Operationen erfolgen, die grob in den Blöcken linguistische Verarbeitung, Graphem-Phonem-Umsetzung, Prosodiesteuerung, Bausteinauswahl und akustische Synthese zusammengefaßt werden.

Sucht man im Internet nach vorhandenen Möglichkeiten, die Funktionsweise und das Zusammenwirken dieser Blöcke zu veranschaulichen, findet man zwar viele TTS-Systeme, in die man Text eingeben kann und die daraufhin das Synthesesignal liefern, jedoch keine Möglichkeit, in das Innere der Systeme zu sehen. Das hängt sicher damit zusammen, daß die meisten Systeme kommerziell genutzt werden. Wir haben uns deshalb entschlossen, im Rahmen des Tutorials die internen Schnittstellen unseres TTS-Systems (Dresdener Sprach-Synthese DreSS) offenzulegen.

Zu diesem Zweck wurde DreSS in Einzelkomponenten zerlegt und in dieser Form in das Tutorial eingebaut. Auf diese Weise kann der Nutzer einen beliebigen Text eingeben, das Ergebnis der linguistischen Vorverarbeitung betrachten und im weiteren die schrittweise Anreicherung mit prosodischer und phonetischer Information verfolgen. Am Ende steht natürlich die akustische Ausgabe. Ein Flußdiagramm, dessen einzelne Funktionsblöcke und Datenbasen angeklickt werden können, dient der verbalen Beschreibung der Komponenten.

3.2. Interaktive Bausteinselektion

Die Erzeugung synthetischer Sprache ist immer mit einem Kompromiß aus Aufwand und Qualität verbunden. Heute ist ein Verfahren üblich, das kurze Sprachsegmente miteinander verkettet, die aus natürlichem Sprachmaterial extrahiert wurden. Diese Sprachsegmente können Einzellaute, aber auch ganze Wörter sein. Verbreitet ist die Verwendung von Zweilaut-Kombinationen (Diphonen), auf die wir uns nachstehend der Einfachheit halber beschränken wollen.

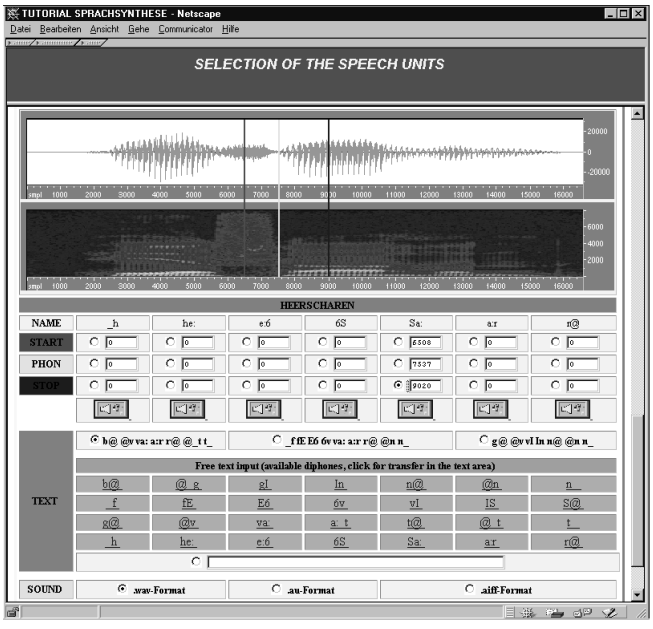

Abbildung 1: Beispiel für die Auswahl eines Diphons. Im oberen Teil sind Zeitfunktion und Spektrogramm eines der vordefinierten Wörter (‚Heerscharen‘) dargestellt. Mit Hilfe der Schaltflächen im mittleren Teil hat der Nutzer die Grenzen des Diphons ‚scha‘ markiert. Mit Hilfe des unteren Teils kann nun der Nutzer ausgewählte Diphone kombinieren, um die Qualität des Selektionsprozesses zu beurteilen.

Das Herstellen eines Diphoninventars zur Verwendung in einem TTS-System ist eine Aufgabe, die einerseits sehr zeitraubend ist, andererseits erfahrenes Personal erfordert. Die notwendige Erfahrung kann man nur durch Training

erhalten. Deshalb enthält unser Tutorial einen speziellen Abschnitt über das Schneiden von Diphonen. Zielstellung war, dem Nutzer die Möglichkeit zu geben, selbst Diphone zu schneiden und die Qualität dieser Selektion durch das Anhören synthetisierter Wörter zu beurteilen. Natürlich kann man niemandem zumuten, alle etwa 1200 Diphone des Deutschen anzufertigen. Deshalb wurde so vorgegangen, daß das Inventar des Standardsprechers von DreSS als Ausgangspunkt genommen wird. Zusätzlich werden von dem gleichen Sprecher vier Beispielwörter bereitgestellt, aus denen der Nutzer Diphone schneiden kann, die dann schrittweise die betreffenden Diphone des Basisinventars ersetzen. In Abbildung 1 ist gezeigt, wie diese Lösung umgesetzt wurde.

4. Probleme

Sowohl bei der Implementierung von vorlesungsbegleitenden Demonstrationen zur Signalverarbeitung und Psychoakustik als auch bei der Gestaltung des Tutorials zur Sprachsynthese wurden wir mit dem Problem konfrontiert, daß interaktive Komponenten in HTML-Pages eingebettet werden müssen, die u. a. die folgenden Leistungen unterstützen sollen:

- Leistungsfähiges graphisches Nutzer-Interface,
- Sprachein- und –ausgabe,
- verschiedene Algorithmen von einfacher Mathematik bis hin zu komplexen Programmen,
- Möglichkeiten zur Handhabung großer Datenmengen.

Diese Forderungen lassen sich kaum komplett auf der Client-Seite erfüllen. So kann man z. B. einen Spracherkenner samt seiner komplexen Datenbasen nicht zum Client herunterladen. Andererseits haben Programme, die komplett auf dem Server laufen, nur beschränkte graphische und Interaktions-Möglichkeiten. Um diese Probleme zu überwinden, wurde die nun dargestellte Architektur vorgeschlagen [4].

5. Eine Rahmensoftware für Tutorials

Bei der Entwicklung der Architektur war besonders das Problem der großen Mengen zu verarbeitender Daten zu berücksichtigen. Wichtig war, Datentransporte so weit wie möglich zu vermeiden. Dieses Ziel wurde dadurch erreicht, daß eine sinnvolle Aufteilung der Verarbeitung zwischen Client und Server vorgenommen wird. Folgende Komponenten wirken nach dem in Abbildung 2 dargestellten Schema zusammen:

- Eine Bibliothek mit Grundalgorithmen (mathematische Standardoperationen, Signaltransformationen usw.), die sowohl beim Client als auch beim Server vorhanden und lauffähig ist;
- ein Programm auf der Server-Seite, das auf dem Server selbst komplexe Programme (wie Spracherkenner oder TTS-System) aufrufen kann;
- ein einfacher Script-Interpreter, der kurze „Mikroskripten“ abarbeitet, indem er Aufrufe der Bibliotheksfunktionen oder der Server-Software vornimmt;
- ein Steuermechanismus, der entscheidet, ob eine Bibliotheksfunktion beim Client oder beim Server aufgerufen werden muß;
- ein Nutzer-Interface, das Anzeige- und Editier-Funktionen für Zeitfunktionen, Spektrogramme usw. sowie auch die Sprachein- und –ausgabe enthält.

Dazu gehört natürlich die Definition der entsprechenden Schnittstellen und Funktionsaufrufe.

6. Schlußbemerkung

Der aktuelle Stand der Arbeiten kann unter www.ias.et.tu-dresden.de/kom/lehre betrachtet werden.

7. Literatur

[1] Bloothoft, G. (Ed.): The Landscape of Future Education in Speech Communication Sciences. Utrecht: Led

Vol. 1: Analysis. 1997.

Vol. 2: Proposals. 1998.

Vol. 3: Recommendations. 1999.

[2] Hoffmann, R.; Kordon, U.; Kürbis, S.; Ketzmerick, B.; Fellbaum, K.: An interactive course on speech synthesis. Proc. ESCA/SOCRATES Workshop MATISSE, London 1999, pp. 61 – 64.

[3] Hoffmann, R.; Ketzmerick, B.; Kordon, U.; Kürbis, S.: An interactive tutorial on text-to-speech synthesis from diphones in time domain. Proc. EUROSPEECH 99, Budapest 1999, Vol. 2, pp. 639 – 642.

[4] Hoffmann, R.; Wolff, M.: Framework design and implementation of web-based tutorials in spoken language engineering. Eingereicht für IEEE ICME 2000.

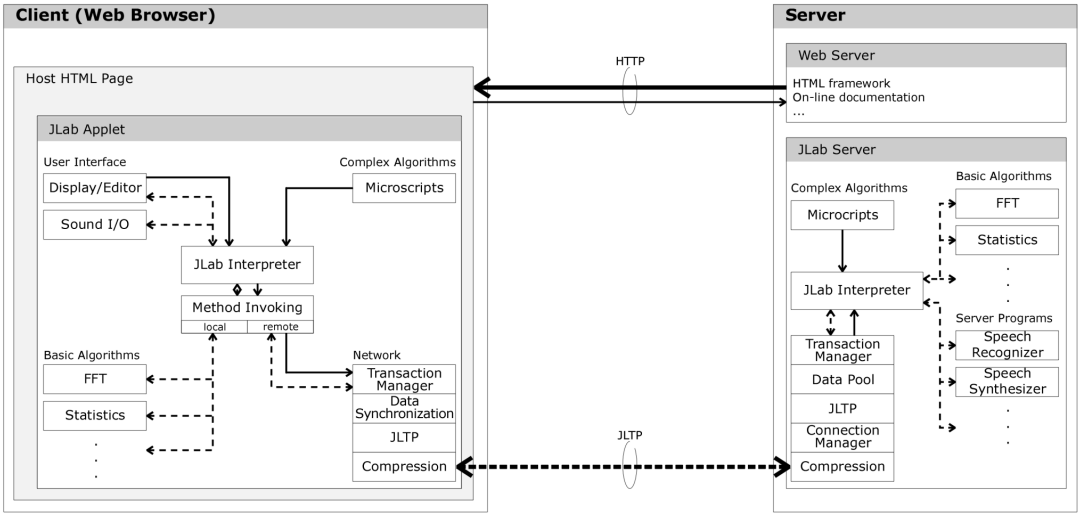

Abbildung 2: Schema der Rahmenarchitektur für Tutorials (aus [4]).