TEXT-TO-SPEECH IM RUNDFUNK – DER PROGRAMMIERBARE MODERATOR?

Thomas Schierbaum

Institut für Rundfunk GmbH, München

schierbaum@irt.de

Abstract:

Maschinelle Sprachausgabe gewinnt im Rundfunk für die kostengünstig zu produzierende Darbietung von Serviceinhalten über Telefondienste und Radiozusatzangebote immer mehr an Bedeutung. Der Vortrag gibt einen Einblick in die Anwendungsgebiete und die daraus resultierenden Anforderungen. Ferner sollen derzeit laufende Projekte sowie die ersten Ergebnisse dieser Arbeiten vorgestellt werden.

1 Maschinelle Sprachausgabe im Rundfunk

In vielen Situationen werden wir schon heute durch maschinelle Sprache angesprochen. Bei der Telefonansage, bei Durchsagen auf dem Bahnsteig, bei elektronischen Spielzeugen oder als Stimme des Navigationssystems im Auto.

Der Rundfunk setzt diese Entwicklungen bereits zur kostengünstigen Präsentation von Informationen in Serviceangeboten über Rundfunk- oder Telefonkanäle ein. Eine Vorreiterrolle spielt dabei VERA (Verkehr in Real Audio), eine Entwicklung des Westdeutschen Rundfunks und eingesetzt bei zahlreichen ARD-Rundfunkanstalten. VERA ist also keine Frau, spricht aber so. Autofahrer erhalten damit rund um die Uhr Verkehrshinweise über Telefon (z.B. SWR-Stauhotline: 150.000 Anrufe/Monat), Mittelwelle (z.B. WDR-Sender Langenberg und Bonn) oder Digital Radio (DAB) – und dort in besonders guter Tonqualität. Dazu werden zunächst kodierte Verkehrsdaten aus der Verkehrsredaktion und dem Traffic Message Channel TMC ausgewertet. VERA verfügt über einen „Sprachschatz“ von etwa 7500 zuvor gesprochenen und aufgezeichneten Meldungsteilen (Audiosamples). Daraus werden dann die kompletten Meldungen nach dem Baukastenprinzip zusammengesetzt und abgespielt. Auch bei NDR, HR, SWR und MDR wird VERA zur Generierung von gesprochenen Verkehrsinformationen eingesetzt. Der Bayerische Rundfunk nutzt für Verkehrsinformationen im digitalen Verkehrskanal ein Text-to-Radio-System auf Basis echter Sprachsynthese. Außerdem werden die beiden Sprachinformationskanäle „BR News + Wetter“ und „BR Business“ automatisch aus „B5aktuell“ generiert. In Sachsen-Anhalt versorgt ARVID (Akustischer Regionaler Verkehrs-Informations-Dienst) die Hörer mit synthetisch generierten Verkehrs- und Wetterinformationen. Sehr erfolgreich ist auch in den USA ein Wetterinformationsdienst der Regierungsbehörde NOAA (National Oceanic and Atmospheric Administration). Die Vorhersagen werden auf Basis von Sprachsynthese über regionale FM-Sender verbreitet.

2 Anforderungen für den Einsatz im Rundfunk

Maschinelle Sprachausgabe lässt sich also in zwei Grundprinzipien unterteilen: Die Wiedergabe zuvor gespeicherter Aufzeichnungen (= Voice Response) und die Umsetzung der Textinformationen mittels Sprachsynthese (= Text-to-Speech TTS). Die Voraufzeichnung von Ansagen weist naturgemäß die höchste Qualität und Natürlichkeit des Stimmenklangs auf, da

in diesem Fall die Meldungsteile von einem geschulten Sprecher/in zuvor gesprochen und gespeichert wurden. Eine flexible Umsetzung beliebiger Eingaben wie bei neuen Straßennamen oder bei Personensuchmeldungen erfordert eine zeitaufwändige Vorbereitung und schließlich wieder den Einsatz eines Rundfunksprechers.
Deutliche mehr Flexibilität bei der Umsetzung unterschiedlicher Textinhalte bietet hingegen die Sprachsynthese. In diesen Systemen werden heutzutage Lautpaare, also Diphone, die zuvor im Sprachlabor gewonnen und extrahiert wurden, abgespeichert und nach einer Analyse des eingegebenen Textes neu zusammengesetzt.
Synthetische Stimmen entsprachen bisher häufig nicht den hohen Qualitätsanforderungen des Rundfunks. Im vergangenen Jahr wurde daher im IRT zusammen mit den Firmen SIKOM und INTERLINX untersucht, wie mit heutiger Technik und bei vertretbaren Kosten eine maximale Qualität bei der Ausgabe synthetischer Sprache zu erreichen ist.
Zu Beginn des Projektes wurden Lösungsansätze für häufig auftretende Problemstellungen bei der Nutzung von Sprachsynthese im Rundfunk entwickelt:
- Berücksichtigung der Anforderungen im Rundfunk;
- modulare Einsetzbarkeit unterschiedlicher TTS-Systeme wie Elan, Rethorical, AT&T;
- Einsatz eines umfangreichen Ausnahmelexikons für Abkürzungen, Aufzählungen, Namen, Strassen, Orte;
- Editor zur Auswahl, Bearbeitung und Wiedergabe einzelner Formen des Foneminventars;
- prosodische Optimierung (= Intonation) zur Steigerung der Anmutung.

Textinhalt	gewünschte Ausgabe	tatsächliche Ausgabe
1	Eins	Eins
1m	Ein Meter	Eins Meter
1.	Erstens	Erstens
1. Mai 2003	Erster Mai	Erstens Mai

Tabelle 1 – Beispiel eines Syntheseproblems

3 Emphasis Studio

Mittlerweile liegen die Entwicklungsergebnisse in Form eines gemeinsamen Produktes „Emphasis Studio“ vor. Die Software besteht aus einem Präprozessor zur linguistischen Optimierung sowie aus einer marktüblichen TTS-Software (= TTS-Engine). Über die grafische Benutzeroberfläche können etwa zwei Millionen Einträge des Ausnahmelexikons des linguistischen Präprozessors verwaltet und phonetisch manipuliert werden. Nach der Übernahme einer Textinformation aus dem integrierten Texteditor oder über eine Softwareschnittstelle werden die Wortformen vom Präprozessor analysiert und mit Hilfe des Ausnahmelexikons in eine für das TTS-System universelle Sprache umgesetzt. Neben den festen Begriffen können auch Aufzählungen, Abkürzungen und Datumsangaben erkannt und zugeordnet werden. Eine Sprachsynthese orientiert sich immer an der Orthografie, was im Fall von Fremdwörtern oft zu falschen Ergebnissen führen kann. Daher verfügt „Emphasis Studio“ über einen einfach zu bedienenden Editor für das International Phonetic Alphabet, kurz IPA. Diese Funktion erlaubt einer Wortform eine Lautschrift aus dem Phoneminventar zu hinterlegen und die durch Abhörvorgänge zu optimieren.

Kategorie	Wortkette	Betonungsvorschrift
Achtung	„Achtung Autofahrer“	Volume: +40%; Speed: - 20%; usw.
Fahrtrichtung	Autobahn A(Nr.) , (Ort1) in Richtung (Ort2)	Volume: +10%; Speed: +10%; usw.

Tabelle 2 – „Emphasis“-Grammatiken

Der wichtigste Schritt hin zur optimalen Ausgabequalität ist die prosodische Komponente, also die Intonation, da diese sich auf die Gesamtanmutung der maschinellen Sprachausgabe auswirkt. Häufig klingen TTS-Systeme sehr modulationsarm und ohne Dramaturgie. „Emphasis Studio“ verfügt daher über eigene Grammatiken. Eine „Emphasis-Grammatik“ kann mehrere Wortketten enthalten und wird nach Kategorien geordnet in der Datenbank abgespeichert. Auf diese Weise werden Zeichenketten im Text erkannt und die Sprachausgabe durch die Betonungsvorschriften, wie Sprechgeschwindigkeit, Tonhöhe und Betonungsverläufe moduliert. Die Vorgaben des Präprozessors werden unter Verwendung von XML-Daten (eXtensible Markup Language) in der international genormten Auszeichnungssprache Speech Synthesis Markup Language (SSML) an eine TTS-Engine übergeben. Die Sprachausgabe erfolgt entweder in eine WAVE-Audiodatei oder als permanenter Audiostream.
In der aktuellen Version von emphasis Studio wird derzeit die TTS-Engine rVoice der Firma Rethorical eingesetzt. Mit dieser Software lassen sich im derzeitigen Produktvergleich sehr gute Ergebnisse in Bezug auf Sprachqualität und Betonungsverläufe erzielen.

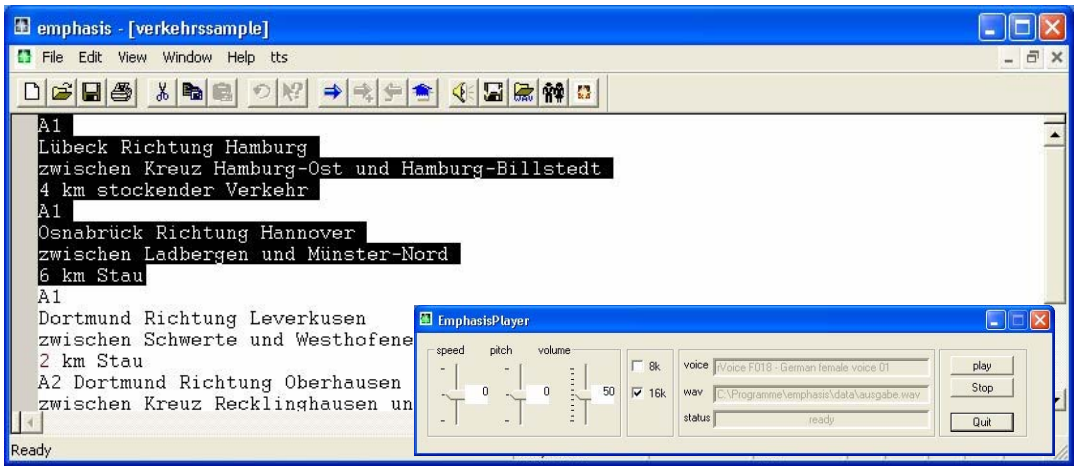

Abbildung 1 – „Emphasis Studio“-Bildschirmoberfläche mit Editor und Player

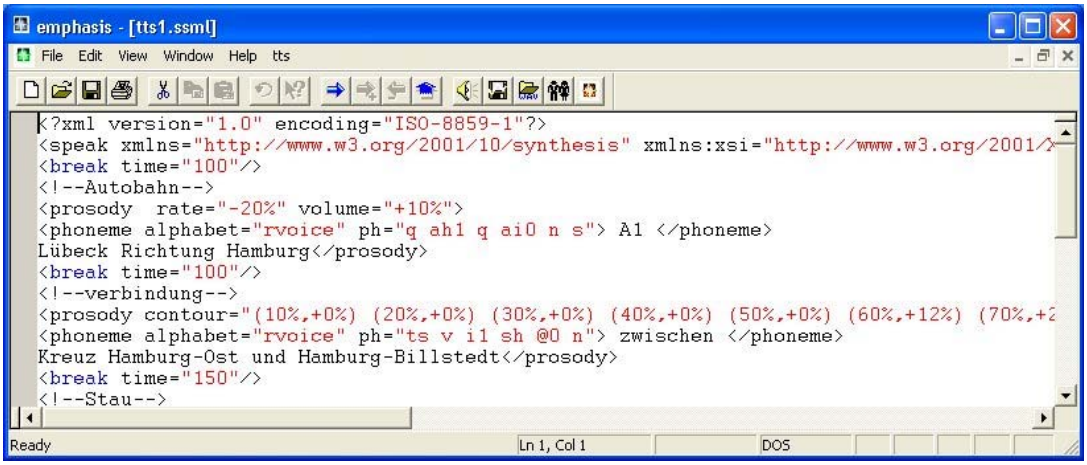

Abbildung 2 – SSML-Ausgabedatei mit Betonungsvorschriften

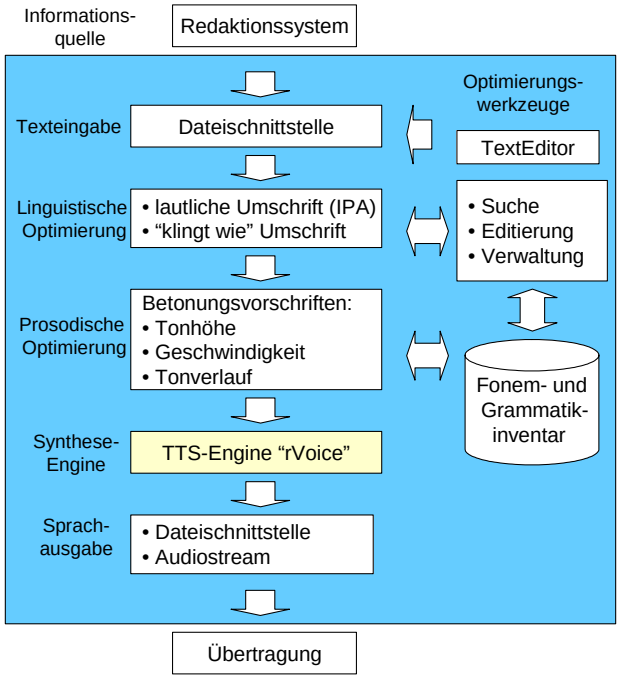

Abbildung 3 – Systemaufbau von „Emphasis Studio“

4 Ausblick

Mit „Emphasis Studio“ steht dem Rundfunk ein modulares Basissystem für eine maximale Zielqualität zur Ausgabe maschineller Sprache zur Verfügung. Zahlreiche Tests haben gezeigt, dass viele Betonungsregeln, die von einem Rundfunksprecher intuitiv umgesetzt werden, zuerst analysiert und im System als Grammatik zu hinterlegen sind. Das ist abhängig vom Einsatzgebiet und der Dramaturgie innerhalb der Meldungskette. Im nächsten Schritt gilt es gemeinsam mit den Rundfunkanstalten mögliche Anwendungsgebiete festzulegen, mit geschulten Sprechern einmalig die Betonungsregeln zu definieren und im linguistischen Präprozessor zu speichern. Ein Hörer sollte später den Unterschied zwischen dem Text von einer Maschine oder von einem Rundfunksprecher spontan nicht mehr erkennen können.
Zu der Horrorvision vom „automatischen Moderator“ im Rundfunk, wie sie im Hörspiel „Die Stimme des Hörers“ von Eran Schaerf aus dem Jahre 2002 mitzuerleben war, wird es jedenfalls nicht kommen. Einem Avatar im Radioprogramm würde es an jeglicher Form von Spontaneität, Kreativität und sprachlicher Wärm fehlen. Aber für kostengünstig zu produzierende Zusatzangebote mit ununterbrochenen Wiederholungen hätte er den richtigen Arbeitsplatz.