VERSTÄRKUNGSLERNEN ZUR PROSODIEVORHERSAGE IN EINEM SPRACHPRODUKTIONSSYSTEM

Markus Schnell

Infineon Technologies AG, München
markus.schnell@infineon.com

Abstract: Text-to-Speech-Systeme werden neben dem Vorlesen von Zeitungsartikeln, E-Mail oder Internetseiten auch für die Versprachlichung von Daten aus Datenbanken oder in Dialogsystemen eingesetzt. In diesen Fällen erzeugt der Computer nicht nur das Sprachsignal, sondern auch den zugrundeliegenden Text. Dies bietet die Möglichkeit die Prosodie deutlich zu verbessern, da man den in der Zwischenstufe erzeugten Text mit Markierungen zu Akzentuierung und Phrasierung versehen kann. Solche Systeme zur Sprachproduktion, auch Concept-to-Speech genannt, waren bisher reine Expertensysteme, das heisst, die Regeln zur Prosodieerzeugung wurden manuell kodiert. Der vorliegende Artikel verfolgt einen anderen Ansatz. Anstatt die Regeln manuell einzugeben, werden sie mit einem maschinellen Lernverfahren gelernt. Bei dem eingesetzten Verfahren handelt es sich um das Verstärkungslernen (reinforcement learning). Beim Verstärkungslernen wird nach jeder Aktion des Systems eine Bewertung der aktuellen Situation vorgenommen. Aus der Gesamtbewertung erschließt das Lernsystem die Anteile der einzelnen Aktionen am Erfolg bzw. Misserfolg, und passt seine Strategie entsprechend an. Um das Verfahren zu demonstrieren wird das Concept-to-Speech-System Demosthenes vorgestellt, das auf dem Text-to-Speech-System DRESS beruht. Der Artikel erläutert insbesondere die Anwendung des Verstärkungslernens in der Prosodiekomponente des Systems. Ein Präferenztest zeigt eine deutliche Bevorzugung der Prosodie des Concept-to-Speech-Systems gegenüber dem Text-to-Speech-System.

1 Einleitung

Vorleseautomaten oder *Text-to-Speech-Systeme* werden neben dem Vorlesen von Zeitungsartikeln, E-Mail oder Internetseiten auch für die Versprachlichung von Daten aus Datenbanken oder in Dialogsystemen eingesetzt. Das ist zum Beispiel der Fall bei Systemen zur Wetterauskunft, bei Museumsführern, in Anrufbeantwortern oder in Navigationsgeräten. In diesen Fällen erzeugt der Computer nicht nur das Sprachsignal, sondern auch den zugrundeliegenden Text. Verwendet man hier herkömmliche Text-to-Speech-Systeme können Schwierigkeiten auftreten, die sich durch den Einsatz eines *Concept-to-Speech-Systems* vermeiden lassen. Ein Text-to-Speech-System muss alle Informationen mit Hilfe einer Textanalyse aus dem geschriebenen Text ermitteln. Das ist aber nicht vollständig möglich. Da die Verwendung von Auskunftssystemen impliziert, dass es sich nicht um beliebige Texte handelt, sondern um Texte, die vom System selbst erzeugt wurden, ist die Idee von Concept-to-Speech, die beim Erstellen der Texte vorhandene Information auch bei der Generierung des Sprachsignals zu verwenden. Anstatt einen reinen Text als Schnittstelle zwischen Textgenerierung und Sprachsynthese zu verwenden, werden noch weitere Informationen weitergegeben.

Ein solches System verspricht Vorteile bei den folgenden drei Punkten:

Transkription: Alle Wörter, die für den Text generiert werden, müssen auch gesprochen werden. Notiert man für jedes Wort sowohl Schreib- als auch Sprechweise, kann auf die

Graphem-Phonem-Umsetzung mit Hilfe eines Regelsystems verzichtet werden. Die Angabe der Sprechweise ermöglicht auch die Unterscheidung von Wörtern mit identischer Schreibweise, aber unterschiedlicher Aussprache, zum Beispiel „Bug“ als Teil eines Schiffes (/bu:k/) oder „Bug“ als Fehler im Computerprogramm (/bak/). Zur Transkription gehören auch die Silbengrenzen, so dass zum Beispiel „beinhalten“ „be-in-hal-ten“ und nicht „bein-hal-ten“ getrennt wird.

Ein weiterer Bereich sind Abkürzungen. Alle im Text verwendeten Abkürzungen können mit der passenden Sprechweise generiert werden, zum Beispiel „NATO“ als /na:to:/, „CIA“ als /si:?aI?eI/ oder „u.a.“ als /?unt/ /?andEr@/.

Die Integration von Schreib- und Sprechweise verhindert auch Schwierigkeiten bei der Zuordnung der betonten Silben. So wird dem Konzept „der Monat August“ die Schreibweise „August“ und die Sprechweise /?aUg'ust/ zugeordnet. Dem Konzept „der Name August“ wird ebenfalls die Schreibweise „August“, aber die Sprechweise /'?aUgust/ zugeordnet. Für Fremdwörter wie „Bordeaux“ oder Kunstwörter wie „Handy“ ist es ebenso einfach die korrekte Sprechweise zu generieren.

Schwierigkeiten gibt es für Text-to-Speech-Systeme auch bei der Umwandlung von Zahlenformaten. Je nach Kontext muss eine Zahl passend gebeugt werden. So muss der Ausdruck „1.“ in den Sätzen „Sie war die 1.“, „Sie traf ihn am 1.“ oder „Sie traf ihn als 1.“ in „erste“, „ersten“ oder „erstes“ umgewandelt werden. Ebenso kann ein Ausdruck wie „10.12“ je nach Kontext als „zehn Uhr zwölf“, „zwölf nach zehn“ oder auch „zehn Komma zwölf“ gesprochen werden.

Mit einem Concept-to-Speech-System sind Transkriptionsfehler aufgrund falscher Zuordnungen oder unpassenen Regeln ausgeschlossen – es treten keine Mehrdeutigkeiten auf.

Phrasierung: Phrasierung umfasst zum einen die Wahl der Phrasengrenzen, zum anderen die Wahl des passenden Satzmodus. In einem Text-to-Speech-System wird häufig aufgrund des Satzzeichens entschieden, ob die Stimme zum Satzende hin gleich bleiben, steigen oder fallen soll. Dazu muss gewährleistet sein, dass die Satzzeichen auch als Satzzeichen verwendet werden und nicht etwa zu einer Abkürzung gehören.

Auch kann es sein, dass ein normalerweise mit einer bestimmten Melodie verbundenes Satzzeichen in manchen Fällen einen anderen Verlauf aufweist. Ein Beispiel ist die Frage „Wer war das?“, die je nachdem wie die Stimme sich am Satzende verhält, eine unterschiedliche Bedeutung bekommt. Für ein Text-to-Speech-System kann diese Frage jedoch immer nur in einer Variante vorgelesen werden.

Grundsätzlich gilt, dass nicht überall wo im Schriftlichen ein Satzzeichen steht, beim Vorlesen eine Grenze existiert und umgekehrt.

Dadurch, dass bei einem Concept-to-Speech-System die Phrasierung nicht aus den Satzzeichen erschlossen werden muss, sondern als Information zur Verfügung steht, können auch hier keine Fehler aufgrund einer Analysestufe entstehen.

Akzentuierung: Gleiche Sätze können je nach Kontext unterschiedlich akzentuiert werden. Beim folgenden Beispiel haben die zwei Sätze eine etwas unterschiedliche Bedeutung:

1. Sie haben *eine* Reise gewonnen.
2. Sie haben eine *Reise* gewonnen.

Beispiel 1 legt den Akzent auf „eine“, was es als Zahlwort ausweist, und damit klar ist, dass es sich nicht um zwei, sondern um genau eine Reise handelt. Beispiel 2 legt den Akzent auf „Reise“, also auf das, was gewonnen wurde.

Die genannten Punkte – Transkription, Phrasierung und Akzentuierung – lassen sich bei Text-to-Speech-Systemen zwar durch Regeln oder statistische Ansätze verbessern, aber sie lassen sich nicht systematisch korrekt generieren. Bei einem Concept-to-Speech-System ist dies jedoch der Fall.

2 Das Concept-to-Speech-System Demosthenes

Aufgrund dieser Überlegungen wurde aus dem Text-to-Speech-System DRESS [1] ein Concept-to-Speech-System erstellt, das auf Textanalyse und Graphem-Phonem-Umsetzung verzichtet und stattdessen eine Textgenerierungskomponente aus verschiedenen Komponenten enthält [2]. Ein Auftrag veranlasst das System aus Schablonen mit Hilfe der Komponente Planer/Formulator eine Textstruktur zu generieren, die dann zur Vorhersage der symbolischen Prosodie verwendet wird. Aus dieser mit prosodischen Merkmalen versehenen Textstruktur erstellt die Komponenten Artikulator ein Syntheseformat, das als Eingabe in den Syntheseteil von DRESS dient und alle Angaben zu Transkription, Phrasierung und Akzentuierung enthält. DRESS selbst generiert aus dem Syntheseformat dann das akustische Sprachsignal.
Abbildung 1 zeigt den Aufbau des Text-to-Speech-Systems DRESS und Abbildung 2 den Aufbau des umgebauten Systems.

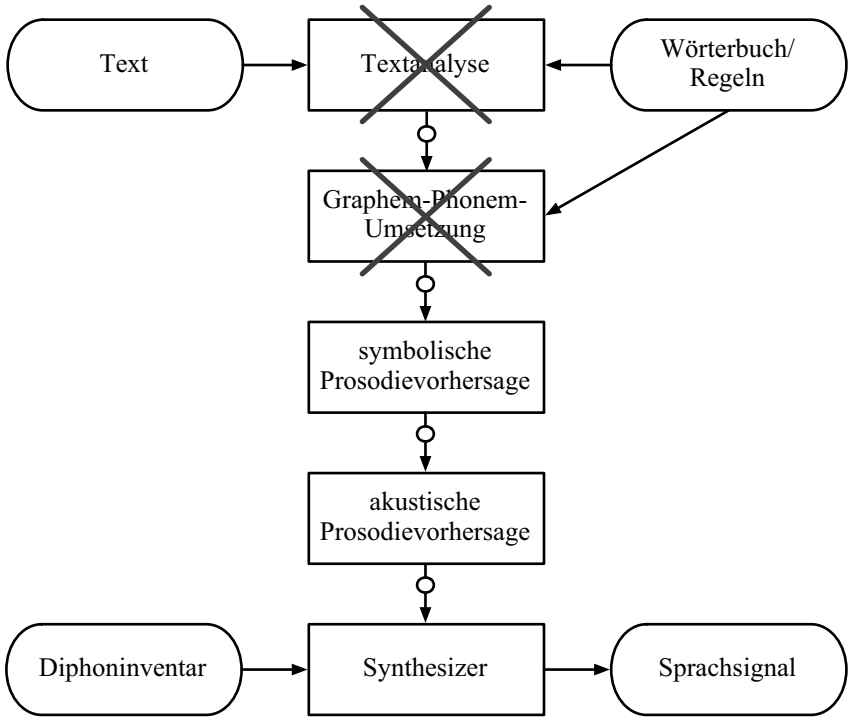

Abbildung 1 - Der Aufbau des Text-to-Speech-Systems DRESS.

Mit dem umgebauten System soll nun ein Weg gefunden werden, die symbolische Prosodie vorherzusagen. Dazu soll zunächst ein Überblick über bisherige Concept-to-Speech-Systeme und die dort verwendeten Methoden gegeben werden.

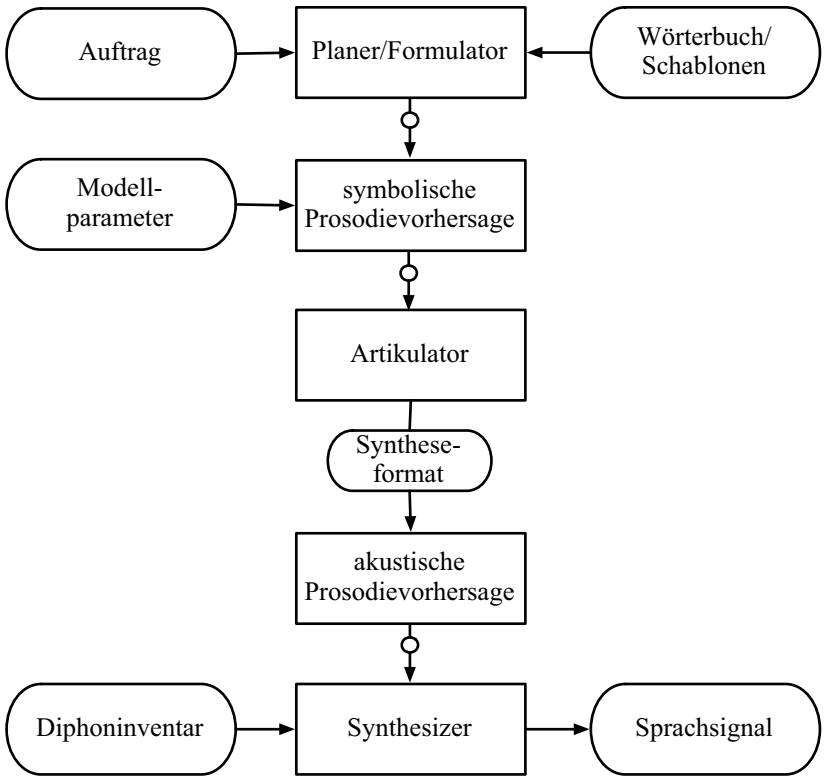

Abbildung 2 - Der Aufbau des Concept-to-Speech-Systems Demosthenes.

3 Relevante Literatur

Ein Concept-to-Speech-System wurde erstmals von Young und Fallside beschrieben [3]. Dieses System verwendet die syntaktische Struktur um Grenzen und Grundfrequenzverläufe abzuleiten. Spätere Systeme erzeugen die akustischen Prosodiemerkmale nicht mehr direkt, sondern verwenden eine symbolische Zwischenstufe mit Akzenten oder Tönen und nehmen Diskurs-, Informations- oder Argumentstruktur als Grundlage zur Prosodievorhersage [4, 5, 6, 7].

Allen diesen Systemen ist gemeinsam, dass sie auf manuell erstellten Regelsystemen beruhen, die von linguistischen Experten erstellt werden müssen. Zwar lassen sich diese Regelsätze verwenden, um arbeitsfähige Systeme zu erstellen, jedoch ist der Anteil, den der Experte beim Entwickeln beitragen muss, sehr groß und führt so zu hohen Entwicklungskosten.

Um die Erstellung von Regelsätzen per Hand zu vermeiden, bietet es sich an, ein lernendes System zu verwenden, was auch die Verwendung des Systems für andere Sprachen oder eine Adaption auf bestimmte Gegebenheiten erleichtert.

Erste Arbeiten in diese Richtung verwenden einen Korpus, um daraus statistische Angaben abzuleiten, die dann in Regeln umgesetzt werden [8] oder verwenden den Korpus direkt, um vollständige Ausgaben zu erzeugen [9].

Das in diesem Artikel vorgestellte Verfahren des Verstärkungslernens hat den Vorteil, dass es nur wenige Beispiele zum Lernen benötigt, und auch noch während des produktiven Einsatzes weiter lernen kann.

4 Verstärkungslernen

4.1 Grundlagen

Verstärkungslernen geht von der Vorstellung eines *Akteurs* (oder Agenten) in einer *Umgebung* aus [10]. Der Akteur kann die Umgebung beobachten und aus der Beobachtung auf ihren *Zustand* schließen. Der beobachtete Zustand kann vom wahren Zustand der Umgebung abweichen. In jedem Zustand entscheidet sich der Akteur für eine *Aktion*, die auf die Umgebung einwirkt. Diese gibt dem Akteur eine *Bewertung* der aktuellen Situation zurück. Es handelt sich dabei nicht um eine Bewertung der einzelnen Aktion, sondern um eine Bewertung des aktuellen Gesamtzustandes. Abbildung 3 zeigt den Zusammenhang zwischen Akteur und Umgebung.

Abbildung 3 - Interaktion von Akteur und Umbegung beim Verst¨arkungslernen.

Ziel des Akteurs ist es, die zukünftigen Bewertungen zu maximieren. Zu einem bestimmten Zeitpunkt t bezeichnen $r_{t+1}, r_{t+2}, \ldots$ die noch folgenden Bewertungen. Diese werden zu einer Gesamtbewertung R_t addiert. Dabei sorgt ein Abschlagsfaktor γ dafür, dass weiter in der Zukunft liegende Bewertungen weniger stark in die aktuelle Gesamtbewertung R_t einfließen:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \; . \tag{1}$$

Da der Akteur nicht in die Zukunft schauen kann, muss er die noch folgenden Bewertungen anhand der bereits gemachten Erfahrungen schätzen. Dazu befindet sich innerhalb des Akteurs eine Tabelle, die für jeden beobachtbaren Zustand und den in diesem Zustand möglichen Aktionen einen Schätzwert für den erwarteten *Nutzen* enthält. Dieser ist die Schätzung für die Gesamtbewertung für eine bestimmte Zustands-Aktions-Kombination.

Befindet sich der Akteur in einem bestimmten Zustand s, folgt auf eine Aktion a eine Bewertung r, die dazu verwendet wird, den Schätzwert $Q(s, a)$ für das Zustand-Aktions-Paar (s, a) zu aktualisieren. Die neue Schätzung $Q'(s, a)$ ergibt sich aus der alten Schätzung $Q(s, a)$ und dem maximalen Nutzen der folgenden Aktionen. Dieser ist näherungsweise der Maximalwert aller Aktionen des Folgezustands s'. Ein Lernfaktor α regelt die Robustheit gegenüber langfristigen Änderungen der Umgebung und der Abschlagsfaktor γ ist der Faktor aus Gleichung 1. Die Formel zur Aktualisierung des Schätzwertes für ein Zustands-Aktions-Paar lautet dann:

$$Q'(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \; . \tag{2}$$

Bevor der Akteur zu lernen beginnt, enthält die Tabelle Zufallswerte. Durch fortwährende Interaktion mit der Umgebung werden diese über die Zeit dem wahren Wert für den Nutzen angenähert.

Die Tabelle ist auch Grundlage für die Auswahl der Aktionen in einem bestimmten Zustand. Eine Strategie ist es, zu Beginn die Aktionen vorwiegend zufällig zu wählen und im weiteren Verlauf immer öfter Aktionen mit maximalem Nutzen zu wählen. Das hat zur Folge, dass zunächst alle Aktionen erforscht und die Schätzwerte für ihren Nutzen angepasst werden. Dies lässt sich zum Beispiel mit der folgenden Methode erreichen: Grundsätzlich wird die Aktion mit dem höchsten Nutzen gewählt, aber mit einer Wahrscheinlichkeit ϵ wird stattdessen zufällig eine andere Aktion ausgewählt. Zunächst hat ϵ einen hohen Wert, der dann später sehr klein oder auf null gesetzt wird. Vor der Wahl jeder Aktion wird eine Zufallszahl $z \in [0, 1]$ erzeugt. Bei n möglichen Aktionen in einem Zustand ist die Wahrscheinlichkeit $P(a)$ für eine bestimmte

Aktion a dann

$$P(a) = \begin{cases} 1, & \text{falls} \quad z < \epsilon \wedge Q(s,a) = \max_{a'} Q(s,a') \\ 0, & \text{falls} \quad z < \epsilon \wedge Q(s,a) < \max_{a'} Q(s,a') \\ \frac{1}{n}, & \text{falls} \quad z \geq \epsilon \end{cases} \tag{3}$$

mit

$$\epsilon_{t+1} = \beta \cdot \epsilon_t \ . \tag{4}$$

4.2 Einsatz bei Demosthenes

Bei Demosthenes wird Verstärkungslernen für die Prosodievorhersage eingesetzt. Die Prosodiekomponente ist der Akteur. Die Umgebung ist der Rest des Systems, wobei sich der Zustand der Umgebung allein aus der Reihenfolge der einzelnen Wörtern bestimmt. Der vom Akteur beobachtbare Zustand besteht aus vier Elementen:

Neue oder bekannte Information: Handelt es sich bei dem Konzept hinter dem aktuellen Wort um ein schon erwähntes oder erschließbares Konzept oder ist es vollkommen neu? Im Textausschnitt „Riesling ist eine Weißweinsorte. Er duftet nach Pfirsich." ist „Riesling" neue Information und „er" bekannte Information.

Kontrastive Information: Beschreibt, ob es sich um einen Kontrast handelt. Im Textausschnitt „Die Angaben sind in Euro, nicht in Mark." sind „Mark" und „Euro" kontrastive Information.

Funktion- oder Inhaltswort: Beschreibt, ob es sich bei dem aktuellen Wort um ein Inhaltswort („Baum, schlafen, rot") oder ein Funktionswort („wird, ein, durch") handelt.

Anzahl Wörter seit dem letzten Akzent: Diese Angabe zählt die Anzahl der Wörter seit dem letzten nicht akzentuierten Wort. Für das Beispiel „Die *Wolken* bedeckten den *Himmel.*", bei dem die Akzente auf „Wolken" und „Himmel" liegen, ergibt sich folgende Zählung: „(0) Die (1) *Wolken* (0) bedeckten (1) den (2) *Himmel* (0)."

Als Aktionen sind Akzentuierung und Nichtakzentuierung eines Wortes möglich. Um den Akteur, also die Prosodiekomponente, zu trainieren, wird ein kurzer Beispieltext von einem Sprecher vorgelesen und aufgenommen und die Positionen der Akzente markiert. Diese Markierung dient der Umgebung als Grundlage zur Bewertung: Am Ende jedes Satzes wird die vom Akteur vorgenommene Akzentuierung mit der Beispielakzentuierung verglichen und eine Bewertung zurückgemeldet. Wenn der Satz k Wörter enthält und p die Anzahl der in der Akzentuierung übereinstimmenden Wörter und n die Anzahl der nicht übereinstimmenden Wörter bezeichnet, dann berechnet sich die Bewertung r zu

$$r = \begin{cases} (p-n)/k = (2p-k)/k, & \text{am Satzende} \\ 0, & \text{sonst} \end{cases} \tag{5}$$

5 Präferenztest

Um festzustellen, ob die in der Einleitung genannten Vorteile eines Concept-to-Speech-Systems auch zu einem besseren Hörerurteil führen, wird das trainierte Concept-to-Speech-System mit dem herkömmlichen Text-to-Speech-System verglichen.

Dazu erzeugt das Concept-to-Speech-System mehrere Texte, die es zum einem selbst versprachlicht, zum anderem vom Text-to-Speech-System vorgelesen werden. Diese Aufnahmen

Textbeispiel-Nr.	TTS	CTS
1	0	8
2	1	7
3	0	8
4	0	8
5	0	8

Tabelle 1 - Ergebnisse des Pr¨aferenztests zwischen Text-to-Speech-System (TTS) und Concept-To-Speech-System (CTS).

werden in zufälliger Reihenfolge verschiedenen Hörern vorgelegt. Die Ergebnisse dieses Experiments zeigt Tabelle 1.

Das Ergebnis ist eindeutig: Das Concept-to-Speech-System wird in fast allen Fällen bevorzugt. Weitere Experimente zeigen, dass dies hauptsächlich daran liegt, dass keine groben Fehler mehr auftreten. Beispiele für grobe Fehler sind die Aussprache von „Chardonnay“ als /CardOnaI/ oder die Akzentuierung von „eine deutliche Säure“ auf „eine“, obwohl es sich hier auf jeden Fall um einen nicht akzentuierten Artikel handelt. Keine klare Präferenz ließ sich hingegen für die Akzentuierung auf „deutliche“ oder „Säure“ erkennen, dies macht für den Hörer keinen Unterschied.

6 Schlussbemerkung

Die Vorteile, die für ein Concept-to-Speech-System gegenüber einem Text-to-Speech-System in Transkription, Phrasierung und Akzentuierung angenommen wurden, konnten anhand eines Präferenztests nachgewiesen werden. Die Akzentuierung wurde dabei nicht durch ein manuell erstelltes Regelsystem erzeugt, sondern durch Verstärkungslernen gelernt und vorhergesagt.

Für die Zukunft wäre es denkbar, das Concept-to-Speech-System mit einem Spracherkenner zu verbinden und die Akzentuierung direkt am gesprochenen Beispiel zu lernen.

Literatur

[1] Hoffmann, Rüdiger: A multilingual text-to-speech system. In: The Phonetician, 1999, Vol. 80, S. 5–10

[2] Schnell, M.: Umsetzung semantischer Konzepte in gesprochene Sprache. In: Hoffmann, R. (Hrsg.): Tagungsband der 13. Konferenz Elektronische Sprachsignalverarbeitung (ESSV02), Dresden, 2002, S. 274–281

[3] Young, S. J.; Fallside, F.: Speech synthesis from concept: A method for speech output from information systems. In: Journal of the Acoustic Society of America, 1979, Vol. 66, No. 3, S. 685–695

[4] Davis, J. R.; Hirschberg, J.: Assigning intonational features in synthesized spoken directions. Proceedings ACL88, 1988, S. 187–193

[5] Prevost, S. A.: A semantics of contrast and information structure for specifying intonation in spoken language generation. PhD thesis, University of Pennsylvania, 1995

[6] Günther, C.: Prosodie und Sprachproduktion. Diss., Universität Hamburg, 1999

[7] Marsi, E. C.: Intonation in spoken language generation. PhD thesis, LOT, 2001

[8] Pan, S.: Prosody modeling in concept-to-speech generation. PhD thesis, Columbia University, 2002

[9] Pan, S.; Weng, W.: Designing a speech corpus for instance-based spoken language generation. Proceedings of the 2nd International Natural Language Generation Conference (INLG02), New York, 2002, S. 49–56

[10] Sutton, R. S.; Barto, A. G.: Reinforcement learning: An introduction. The MIT Press, Cambridge, 1998