Spoken Dialogue Systems
Sprachdialogsysteme

Seminar (MSc)

Wintersemester 2011/2012

Timo Baumann

baumann@informatik.uni-hamburg.de

Übersicht heute

- Amtl. Bekanntmachungen & Bürgersprechstunde
- Besprechung der Leseaufgabe
- Referat über Sprachsynthese
- hoffentlich anschließende Fragen und Anmerkungen

Termine

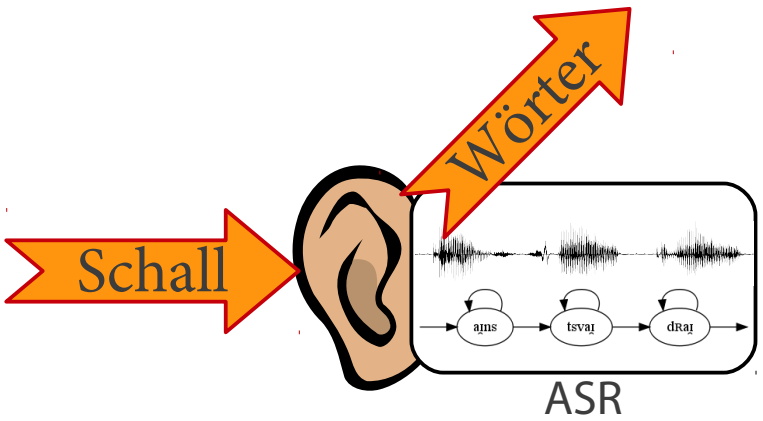
Datum	Thema	Referent
17.10.2011	Einführung	–
31.10.2011	Basistechnologien: Spracherkennung	Timo
07.11.2011	Leseaufgabe	--
14.11.2011	Basistechnologien: Sprachsynthese	Timo
21.11.2011	Semantic Frame-based NLU	Alexander
28.11.2011	Dialogablaufsteuerung	Arne
05.12.2011	entfällt	–
12.12.2011	Praktische Dialoggestaltung	Tim
19.12.2011	Evaluation von Dialogsystemen	Steffen
09.01.2012		
16.01.2012	Turn-Taking	Dimitri
23.01.2012	Inkrementelles Dialogmanagement	Wolfram
30.01.2012	Rückblick und Zusammenfassung	alle

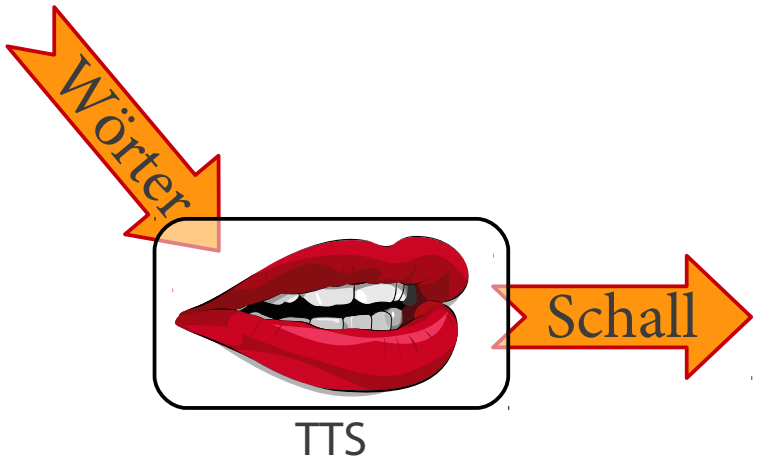
McTear 2004, Kapitel 1-3

- Platz für Anmerkungen und Fragen

Ein- und Ausgabe für Dialogsysteme

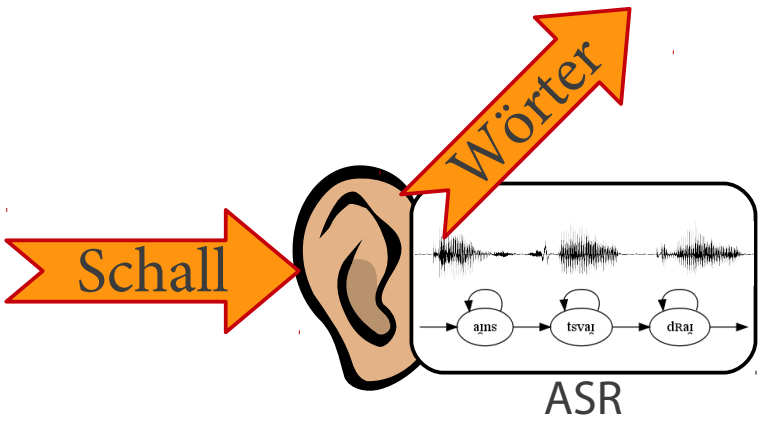
- neulich:
  Spracherkennung
- heute:
  Sprachsynthese

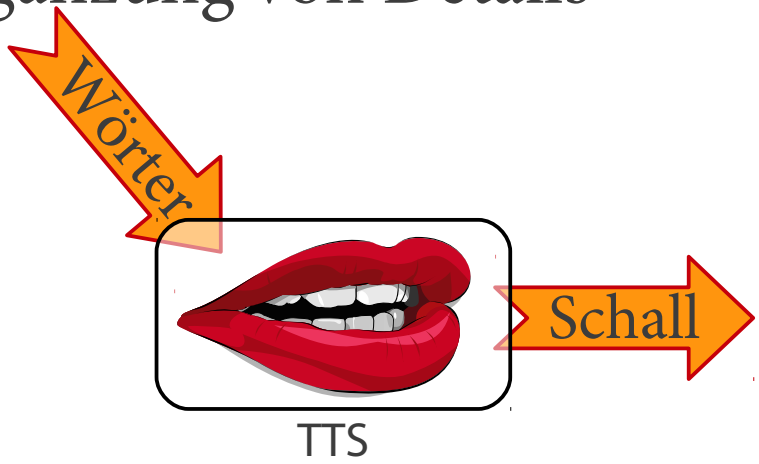




Ein- und Ausgabe für Dialogsysteme

- Spracherkennung
  - Reduktion des Signals auf die Wörter
  - ➔ Informationsverlust
- Sprachsynthese
  - Wörter allein beschreiben das Signal unzureichend
  - ➔ Natürlichkeit erst durch Ergänzung von Details





Geschriebene vs. gesprochene Sprache

- Was wird gesprochen aber nicht geschrieben?
  - Häsitationen, Filled Pauses, Conversational Grunts
  - Korrekturen, Tippfehler, Aussprachefehler
  - Betonung (Satz-, Wort-), Rhythmus
  - Prosodie: Intonation, Dauer, Lautheit, Pausen
  - Sarkasmus, Ironie, Emotion
  - extralinguistisches Verhalten: Gesten, Mimik, …

Geschriebene vs. gesprochene Sprache (Timos Liste)

- Text hat keine Sprechmelodie
  - Satzzeichen helfen nicht auf Diskursebene
- Text enthält keine Angaben über Wortbetonungen
- Homographen die keine Homophone sind: *Montage*
-
- Abkürzungen, Datums-, Zahlenangaben im Text …

Linguistische Vorverarbeitung
Text-to-Speech Concept-to-Speech

- textbasierte linguistische Vorverarbeitung
  - „Das Hindernis bitte umfahren.“
- konkretes Wissen in der Vorverarbeitung

Linguistische Vorverarbeitung
Text-to-Speech Concept-to-Speech

- textbasierte linguistische Vorverarbeitung
  - „Das Hindernis bitte umfahren.“
  - im Sinne von *über* oder *vorbei*?
  - „**úm**fahren“, oder „um**fáh**ren“?
  - der Text gibt diese Information nicht her
- konkretes Wissen in der Vorverarbeitung

Linguistische Vorverarbeitung

Text-to-Speech

- textbasierte linguistische Vorverarbeitung
  - „Das Hindernis bitte umfahren.“
- Erraten der korrekten Sprechweise
- Strategie: Vermeide Fehler durch schwach ausgeprägte Variation

Concept-to-Speech

- konkretes Wissen in der Vorverarbeitung
- Konstruktion aus dem angebenen Kontext
- kann durch ausgeprägte Variation überzeugen

Abwägung Text-to-Speech / Concept-to-Speech

- eigentlich verfügt ein Sprachdialogsystem über die für CTS notwendige Information
- einfachere Anbindung von TTS
  - reine Text-Schnittstelle,
    leichte Integration externer Datenquellen
  - domänenunabhängig (keine Wissensrepräsentation)
- potentiell viel bessere Qualität von CTS
  - Gegebenheit, Topic, Fokus im Diskurskontext, …

➔ teilweise Nutzung von Aussprache-Markup für TTS

Aufbau von Sprachsynthesesystemen

1. linguistische Vorverarbeitung
   - Umwandlung in „Targets“, jeweils:
     Lautidentität, Lautdauer, Grundfrequenzverlauf,
     (Lautheit, Prosodielabel, …)
2. Target → Sprachsignal entweder
   a. musterbasiert: vgl. Dynamic Time Warping
      - Diphonsynthese, konkatenative Synthese
   b. parameterbasiert: HMMs in der Vorlesung
      - Vocoding, Parameter aus HMMs generiert

weiterer Inhalt des Referats

- Linguistische Vorverarbeitung
- Diphonsynthese
- Signalangleichung bei musterbasierten Verfahren
- konkatenative Synthese
- Vocoding und HMM-Synthese

Linguistische Vorverarbeitung

- Eingabe: entweder Text oder Konzepte
- häufig folgende Schritte

• Tokenisierung	Hallo !	Wie geht ’s ?
• Text Normalisierung	hallo !	wie geht ’s ?
• POS-Tagging	/ITJ /$.	/KOKOM /VVFIN /PPER /$.
• G2P	/halo:	vi: ge:t s /
• Prosodieerzeugung	H* L-	L+H* H-%
• akustische Parameter		

Diphonsynthese

- Aneinanderreihung von Sprachschnipseln
- Einheiten von jeweils der Mitte eines Targets bis zur Mitte des nächsten:

  _h+ha:+a:l+lo:+o:_+_v+vi:+i:g+ge:+e:t+ts+s_
- Ziel:
  - Verkettung in der stabilen Phase
  - Abdeckung von Koartikulationseffekten

Diphonsynthese

- Diphone sind von einem Sprecher aufgenommen
  - eingebettet in einen Trägersatz oder ein Trägerwort
  - zum Beispiel immer betont in der Mitte eines Wortes
- ~40 Laute → theoretisch 1600 Lautübergänge!
- Optimierung der Diphonabdeckung
  durch relativ wenige Trägersätze
  - vgl. IpdS-Sätze
  - ist das sinnvoll?

Signalverarbeitung

- Ziele:
  - Verkettungsstellen unhörbar machen
  - Angleichung von Lautheit und Grundfrequenz
  - Manipulation von Grundfrequenz und Dauer
- PSOLA: Pitch-Synchronous Overlap and Add
  - Lautidentität und -charakteristik bleibt erhalten
  - erlaubt Anpassung der Grundfrequenz
  - erlaubt Anpassung der Lautdauer
- Lautheit durch Amplitudenskalierung

PSOLA

1. Markierung der Signalperioden (am Nulldurchgang)
   - nicht trivial, weil Sprache ein komplexes Signal ist
2. Fensterung im Bereich jeder Periode
3. Überlappen und Addieren der Perioden

- Verändern der Periode → Grundfrequenzanpassung
- Wiederholung/Auslassung von Perioden → Daueranpassung
- Signalcharakteristik bleibt weitestgehend erhalten
  - allerdings Artefakte bei starker Manipulation

Valbret et al, 1992

Diphonsynthese – Bewertung

- allerdings ändert sich im natürlichen Sprachsignal einiges in Abhängigkeit von Lautheit und Sprechhöhe
- es bleiben an den Verkettungsstellen unnatürliche Sprünge im Spektrum
  - das Hörorgan kann diese nicht interpretieren
    → Stimme klingt unnatürlich
- Vorteile: kleine Datenbasis (Aufnahme und Speicherung), geringer Berechnungsaufwand

Konkatenative Synthese

- große Auswahl für Mapping Target → Lautmaterial
  - besser zur Targetsequenz passendes Lautmaterial
  - dadurch weniger (oder keine) Signalverarbeitung nötig
- Einheitenauswahl als Suche in einem Sprachkorpus
  - untereinander passende Einheiten → Verkettungskosten
  - zum Target passende Einheiten → Targetkosten
- Suchalgorithmus nach Wahl, potentiell teuer
- hoher Rechenaufwand, sehr hoher Speicherbedarf
- deutlich höhere Natürlichkeit des Resultats

Konkatenative Synthese

- weiterhin nur „Wiedergabe“ des Lautmaterials, aber keine Abstraktion/Modellierung der Eigenschaften des Lautmaterials
- z.B. Adaption an Emotionen bleibt unmöglich (es sei denn Sprachkorpus enthält alles Material in emotional gefärbten Varianten)
- Adaption an Heiserkeit/Schnupfen/Weinen/…
- Kombination aller Stellschrauben

Vocoding und HMM-Synthese

- Quelle-Filter-Modell der Spracherzeugung
  - Stimmlippen im Kehlkopf erzeugen Primärschall
    - periodisch bei stimmhaften Lauten, obertonreich
    - Rauschen bei stimmlosen Lauten
  - Ansatzrohr bildet Resonanzkörper
    - unterschiedlich starke Dämpfung von Frequenzbereichen
    - zusätzliche Schallquellen bei Konsonanten durch Verwirbelungen (Frikative), Plosive
  - Sprachsignal = Primärschall $\otimes$ Resonanz ($\oplus$Zusatzzahl)

Vocoding II

- wir brauchen also
  - Parameterdarstellung des Resonanzkörpers
  - Stimmhaftigkeit → je nach Laut, also bekannt
  - Grundfrequenzverlauf → brauchten wir eh schon
- früher Ansatz: Parameter werden regelbasiert für die Laute (und Lauttrajektorien) erzeugt
  - klingt seeeeehr unnatürlich
- Resonanzkörperbeschreibung wie bei der ASR
  - dann können wir HMMs auf einem Korpus trainieren!

HSMM-Synthese

- heute suchen wir tatsächlich
  $\hat{O} = \text{argmax } O : P(O|W)$ (bzw. $P(O|Ph)$)
  die wahrscheinlichste Observation zur Phonemfolge
- HSMM: Hidden Semi-Markov Modelle: Anzahl der Observationen in einem Zustand ist festgelegt
  - Lautdauermodelle können dies sehr gut bestimmen
  - wir brauchen nicht mehr die optimale Zustandsfolge suchen (sie ist einfach festgelegt)
- Suche reduziert sich auf lineares Gleichungssystem

HMM-Synthese

- Problem 1: die wahrscheinlichste Observation in einem Zustand ist immer genau der Mittelwert
  - Sprünge an Zustandsübergängen
- Abhilfe 1: wie in ASR müssen wir die Kontinuität bei Veränderungen modellieren
  - durch „dynamische Features“, die ins Gleichungssystem mit eingehen; je nach Zustand unterschiedliche Gewichtung der Dynamisierung; meist Δ und $\Delta\Delta$

Tokuda et al, 1995

HMM-Synthese

- Problem 2: immernoch keine Abweichungen von der Ideallinie; zu geringe Varianz im Signal
  - braucht der Mensch, damit es „natürlich“ klingt
- Abhilfe 2: Global Variance Optimization
  - boostet „unwahrscheinliche“ Observationen, sodass die globale Varianz erreicht wird, die Mittelwerte erhalten bleiben
  - variablere Synthese, aber eben natürlicher

Toda und Tokuda, 2007

HMM-Synthese – Bewertung

- höherer Abstraktionsgrad bei ähnlich guter Leistung
- vergleichsweise klitzekleine Modelle (10MB vs. 500)

- Erkennung und Synthese aus ähnlichen Modellen
  - aber: beim Menschen grundverschiedene Organe…
- Nutzbarkeit von ASR-Technologien für Sprecheradaption

Zusammenfassung

- Linguistische Vorverarbeitung
- Diphonsynthese
- Signalangleichung bei musterbasierten Verfahren
- konkatenative Synthese
- Vocoding und HMM-Synthese
  - parameterbasiert
  - Parameter werden aus Korpus gelernt

Vielen Dank für die Aufmerksamkeit.

- Platz für Fragen und Anmerkungen

Termine

Datum	Thema	Referent
17.10.2011	Einführung	–
31.10.2011	Basistechnologien: Spracherkennung	Timo
07.11.2011	Leseaufgabe	--
14.11.2011	Basistechnologien: Sprachsynthese	Timo
21.11.2011	Semantic Frame-based NLU	Alexander
28.11.2011	Dialogablaufsteuerung	Arne
05.12.2011	entfällt	–
12.12.2011	Praktische Dialoggestaltung	Tim
19.12.2011	Evaluation von Dialogsystemen	Steffen
09.01.2012		
16.01.2012	Turn-Taking	Dimitri
23.01.2012	Inkrementelles Dialogmanagement	Wolfram
30.01.2012	Rückblick und Zusammenfassung	alle