DAS PROSODISCH-PHONETISCHE ANNOTATIONSSYSTEM PROPHANO

Sören Wittenberg und Oliver Jokisch

Institut für Akustik und Sprachkommunikation der TU Dresden
soeren.wittenberg, oliver.jokisch@ias.et.tu-dresden.de

Kurzfassung: Umfang und Qualität gesammelter Sprachdaten haben einen maßgeblichen Einfluss auf die Entwicklung von Sprachverarbeitungssystemen. Aufgrund der notwendigen Datenvolumen wird eine automatische Segmentierung und Annotation der Sprachdaten angestrebt. In der Praxis ist in der Regel eine manuelle Kontrolle und Korrektur der Daten erforderlich.

Das Prosodisch-Phonetische Annotationssystem PROPHANO stellt weitgehend automatisierte Werkzeuge zur Verfügung, deren Zusammenschaltung zu einer zuverlässigen Segmentierung und Annotation aufgezeichneter Sprachdaten führt und gleichzeitig das für die manuelle Korrektur notwendige Expertenwissen in den Algorithmen abbildet. Somit kann auch der weniger erfahrene Bearbeiter hochwertige Sprachdatenbasen erstellen.

PROPHANO basiert auf Vorarbeiten des Instituts im Bereich hochqualitativer Sprecherdaten für die Sprachsynthese (TC-STAR [1]) sowie bei der webgestützten Datensammlung von SMS-Texten und dementsprechenden Studioaufnahmen [2]. Vorhandene Werkzeuge zur Aufnahme und Phonemsegmentierung, z. B. WiGE [3], werden verwendet und mit einer prosodischen Analysemöglichkeit erweitert.

PROPHANO integriert die Sprachdatenannotation in eine übersichtliche XML-Struktur, welche eine synchrone oder asynchrone Weiterverarbeitung der Sprachdaten erleichtert und für neue Sprachmerkmale erweiterbar ist.

1 Einleitung

PROPHANO stellt ein komplettes Framework zum Aufbau von annotierten Sprachkorpora bereit. Es ist komponentenbasiert aufgebaut, wodurch die Verarbeitungskette sehr flexibel gehalten werden kann. Ziel ist es alle für die Erstellung einer Sprachdatenbasis notwendigen Komponenten und das in Algorithmen und Regeln notierte Expertenwissen in einem Rahmensystem zusammenzufassen, bei welchen sich die Komponenten aber jederzeit durch leistungsfähigere austauschen lassen und dem Bearbeiter Zugriffspunkte für manuelle Korrekturen zur Verfügung stehen. Abgesehen von der Texterstellung und Sprachaufnahme ist eine manuelle Bearbeitung nicht zwingend.

2 Systembeschreibung

Die Grundstruktur des Systems entsprechend Abbildung 1 geht auf Vorarbeiten des Instituts im Bereich hochqualitativer Sprecherdaten für die Sprachsynthese (TC-STAR [1]) zurück. Es besteht aus den vier Hauptverarbeitungsblöcken *Datenerfassung*, *Segmentierung und phonetische Annotation*, *Linguistisch-prosodische Transkription* und *Informationszusammenführung*.

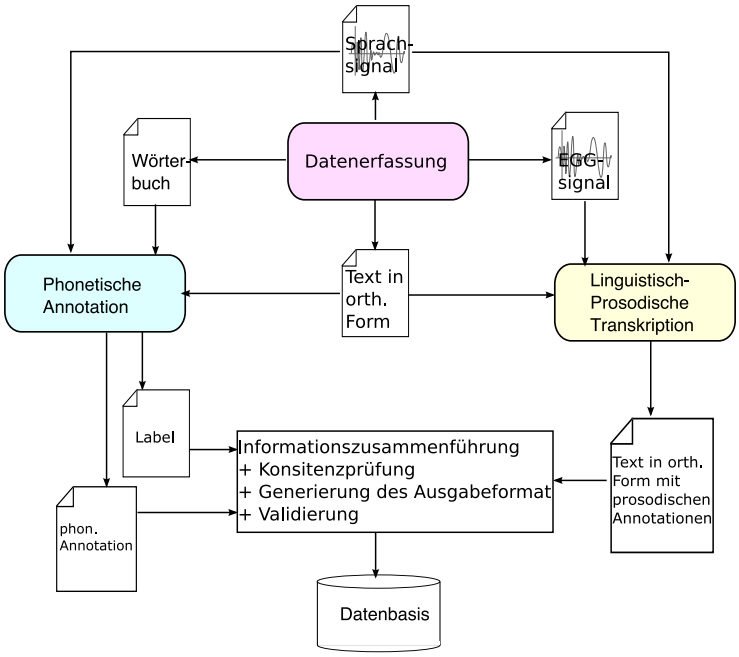

Abbildung 1 - Grobstruktur des Systems

2.1 Datenerfassung

Die Datenerfassung beinhaltet die Erstellung von Texten und Durchführung der Aufnahmen. Die erstellten Texte werden statistisch analysiert. Die dabei entstehende Wortliste wird mit einem vorhanden Wörterbuch (z.B. CELEX für UK Englisch) verglichen und um die phonetischen Beschreibungen erweitert. Aussprachevarianten sind erlaubt. Fehlende Einträge werden durch Experten bzw. andere vertrauenswürdige Wissensquellen ergänzt. Zusammensetzungen mit Bindestrich werden ebenfalls in dieses Wörterbuch aufgenommen, auch wenn die einzelnen Bestandteile bereits enthalten sind. Zur Vermeidung unerwünschter Aussprachevarianten werden Akronyme und Varianten wie ”oh” anstatt ”zero” normalisiert.

Die Sprachaufnahmen erfolgen mit der Fremdanwendung *WiGe-Rec* [3]. Diese präsentiert dem Sprecher die im Vorfeld generierten Texte und führt ihn durch den gesamten Aufnahmeprozess. Der Aufnahmeleiter kontrolliert währenddessen, dass sich der Sprecher strikt an den gegebenen Text hält. Damit kann die notwendige manuelle Nachkontrolle bezüglich Einfügungen, Auslassungen und Ersätzungen verringert werden.

2.2 Segmentierung und phonetische Annotation

Die phonetische Annotation der Sprachaufnahmen erfolgt mit einem phonembasierten HMM-Spracherkenner (forced Alignment). Die im Wörterbuch hinterlegten Aussprachevarianten werden berücksichtigt. Neben den so auf Wortebene ermittelten Phonemsequenzen (einschliesslich zeitlicher Position) werden die Silben- und Wortgrenzen bestimmt. Zur Verbesserung der Qualität der automatischen Annotationen sollte eine manuelle Korrektur erfolgen. Nur so können sprechertypische Aussprachevarianten erfasst werden und die Bestimmung der Lautgrenzen mit großer Genauigkeit erfolgen.

Als Phonemsymbolik wird *Speech assessment methods phonetic alphabet (SAMPA)* [5] mit den folgenden Ergänzungen verwendet:

- l= (addle, ”{d l=)
- n= (admission, @dm”ISn=)

2.3 Linguistisch-prosodische Transkription

Die signalbasierte linguistisch-prosodische Transkription verwendet verschiedene Ansätze, deren Leistungsfähigkeit in [4] dargestellte wurde. So ergänzen sich unteranderem ein auf einem Neuronalen Netzwerk basierender Ansatz mit einem auf dem Fujisaki Modell basierenden. Die resultierenden Annotationen für Phrasengrenzen und -akzente sind zweistufig. Es wird zwischen vollintonierter und intermediärer Phrase bzw. Haupt- und Nebenakzent unterschieden.

Die Verschaltung der in PROPHANO verwendeten Komponenten ist in Abbildung 2 dargestellt. Die Verwendung von Elektroglottograph-Signalen (EGG) ist nicht zwingend, kann aber die Genauigkeit der Pitch-Markierungen verbessern.

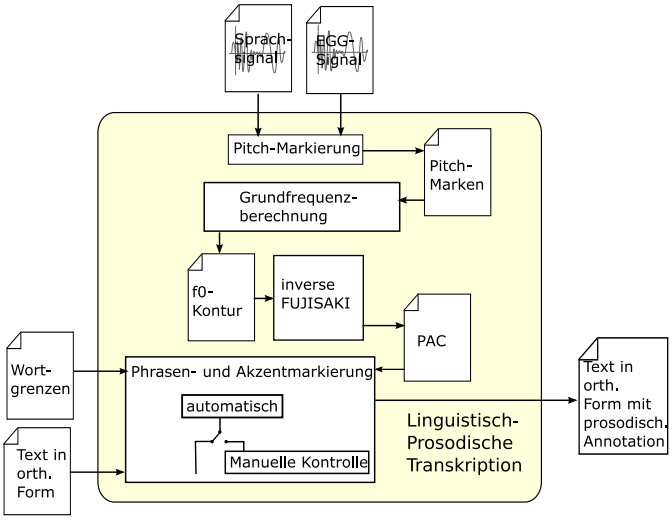

Abbildung 2 - Linguistisch-Prosodische Transkription

2.4 Informationszusammenführung

Die Zusammenführung der aufgezeichneten Sprachdaten und der erstellten Transkriptionen erfolgt im vierten Schritt. Die Daten werden dazu in ein gewünschtes Zielformat konvertiert und auf Konsistenz geprüft. Solange das System aus fest definierten Komponenten besteht, ist diese Prüfung mittels weniger Regeln durchführbar und das System liefert sehr gute Ergebnisse.

Um dem Anspruch von PROPHANO gerecht zu werden, Komponenten einfach austauschen zu können, kommt der Konsistenzprüfung eine sehr wichtige Rolle zu. Sie muß stets die (propritären) Ausgabeformate einzelner Komponenten verarbeiten und alle im System befindlichen Wissensquellen, die unterschiedliche Aspekte einer einzigen sprachlichen Äußerung beschreiben, zusammenführen können. Die syntaktische Beschreibung hat in einer elektronisch verarbeitbaren Form zu erfolgen, damit die notwendigen Transformationen bzw. Konvertierungen rein vom gegebenen Eingabeformat und dem Wunsch des Zielformates gesteuert werden. Das verlangt die Verwendung strukturierter und definierter Ein- und Ausgabeformat. Zur geschlossen Verwendung ist dieses Prinzip auch auf die Beschreibung innerhalb der verschiedenen Wissensquellen zu übertragen wurden, was zu einer strukturierten Annotation führt.

3 Strukturierte Annotation

Die Verwendung von Markup-Sprachen hat sich in vielen Bereichen der Sprachverarbeitung durchgesetzt und es gibt eine Vielzahl von Werkzeugen, Verarbeitungs- und Darstellungssystemen die mit strukturierten Daten arbeiten. Der entscheidende Vorteil ist die entstehende Interoperabilität, die sich aus der eindeutigen Beschreibung z.B. mittels einer Dokumenttyp- oder Schemabeschreibung ergibt. Die Anwendung von Markup-Sprachen führt zu Sprachkorpora mit Annotationen in einer einheitlichen und strukturierten Form, allerdings nur solange die formale Gültigkeit des Dokumentes gegeben ist. Bedenken bezüglich des erhöhten Speicherbedarfs können durch fallende Gigabyte-Preise außer acht gelassen werden.

Im Bereich der Sprachverarbeitung gibt es viele W3C-Empfehlungen zur Beschreibung unterschiedlicher Sachverhalte. Beispiele sind *Voice Extensible Markup Language* (VoiceXML) zur Sprachdialogsteuerung, *Speech Synthesis Markup Language* (SSML) in der Sprachsynthese, *Speech Recognition Grammar Specification* (SRGS) und *Semantic Interpretation for Speech Recognition* (SISR) in der Spracherkennung, sowie zur Darstellung von Aussprachewörterbüchern *Pronunciation Lexicon Specification* (PLS). Da diese Empfehlungen nicht immer allen Anforderungen genügen, gibt es speziell angepaßte Formen wie beispielsweise *MaryXML* [6]. Auch PROPHANO ist diesen Weg gegangen, da es aufgrund seiner Granularität eine sehr feine Beschreibung der Sprachdaten zulässt.

3.1 Wörterbuch

Das während der Datenerfassung entstandene Aussprachewörterbuch wird entsprechend der W3C-Empfehlung mittels *Pronunciation Lexicon Specification* dargestellt. Einzige die für PLS-Prozessoren verlangte Unterstützung des *IPA*-Alphabetes wird nicht bedient, um die ASCII-Darstellung *X-SAMPA* verwenden zu können. Ein Eintrag hat folgende Struktur:

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" alphabet="x-sampa" xml:lang="en-GB">
  <lexeme>
    <grapheme>leg</grapheme>
    <phoneme>leg</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>...
  ...
</lexicon>
```

Aussprachevarianten können als weitere <phoneme>-Beschreibung innerhalb des <lexeme>-Tags plaziert werden. Im Beispiel wird deutlich, welchen Mehrwert die stukturierte Darstellung gegenüber schwerinterpretierbaren, zeilenbasierten Darstellung wie zum Beispiel leg leg bietet.

3.2 Sprachdaten

Die Beschreibung der Sprachdaten erfolgt auf zwei Ebenen. Die erste umfaßt die orthographische und prosodische Beschreibung der gesamten Äußerung mittels *Speech Synthesis Markup Language* (SSML). Damit können allgemeine Parameter wie Pausen, Akzent und Phrasengrenzen annotiert werden. Die Einhaltung der W3C-Empfehlung sichert die Synthese der Äußerungen mittels SSML-konformen TTS-Systemen.

Da SSML nicht umfassend genug für eine vollständige phonetische und prosodische Beschreibung ist, integriert die zweite Ebene die erste und stellt Elemente für eine feinere Transkription bereit. Das folgende Beispiel zeigt einen kleinen Ausschnitt:

```
<?xml version="1.0" encoding="UTF-8"?>
 <prophano version="0.1" xml:lang="en-GB" ...>
   <speak version="1.1" xml:lang="en-GB" ...>
     <lexicon uri="file:///lexicon_laura.pls" xml:id="lexicon_laura"/>
     <voice name="Laura" gender="female" age="32">
       <lookup ref="lexicon_laura">
         <p>
           <s>
             <emphasis level="moderate">Parliament</emphasis> budget <break time="220ms"/>
             ought not to exceed <break strength="x-weak"/> twenty ...
           </s>
         </p>
       </lookup>
     </voice>
   </speak>
   <lexicon uri="file:///lexicon_laura.pls" xml:id="lexicon_laura"/>
   <voice name="Laura" gender="female" age="32">
   <p>
     <s>
       <pros emphasis_level="moderate">
         <w start="0.06s" end="0.60">
           <orth>Parliament</orth>
           <phon>
             <phoneme start="0.06" end="0.15">p</phoneme>
               ...
             <syllable start="0.06" end="0.42">p A:</syllabe>
               ...
           </phon>
         </w>
       </pros>
       <pros break_time="220ms"/>
       <w start="0.60s" end="1.14">
           <orth>budget</orth>
      ...
   <pm> ... </pm>
   <wav xsi:type="base64Binary"> ... </wav>
</prophano>
```

4 Manuelle Korrektur

Durch die Verwendung der Markup-Sprachen können die Daten automatisiert für die manuelle Korrektur aufbereitet werden. Im einfachsten Fall ist es die reine Darstellung in einem Browser, wo mittels xHTML und skalierbarer Vektorgrafiken (SVG) die Darstellung des Sprachsignals mit Segmentgrenzen erfolgt und durch Auswählen eines Segments in eine tiefere Annotationsebene gewechselt werden kann (Beispiel siehe Abbildung 3). Ist die Korrektur der Segmentgrenzen oder Annotationen notwendig, kann durch die Hinzunahme von Java Applets die Bearbeitung ebenfalls direkt im Browser ablaufen. Damit bietet die Anwendung von Markup-Sprachen bei der Erstellung von Sprachdatenbasen die Möglichkeit ohne lokal verfügbare Spezialsoftware Korrekturen an den automatisiert erzeugten Annotationen vorzunehmen. Gepaart mit Servertechnologien könnten die Korrekturen direkt in die Datenbasis eingepflegt werden.

5 Zusammenfassung und Ausblick

PROPHANO ist ein Framework zur Erstellung von annotierten Sprachkorpora. Durch die Verwendung von Markup-Sprachen kann es mittels Datentyptransformationen die Eingabeformate für die verschiedensten Sprachverarbeitungssysteme liefen. Dadurch ist es auch möglich Komponenten anderer Sprachverarbeitungssystem zu integrieren, solange sie automatisiert verarbeitbare strukturierte Daten liefern. Zukünftig soll die PROPHANO zugrundeliegende XML-Struktur dazu genutzt werden, die Suche wie in einer normalen Datenbanken zu ermöglichen und so schnellen Zugriff auf beliebige Elemente der Sprachdatenbasis zu erhalten.

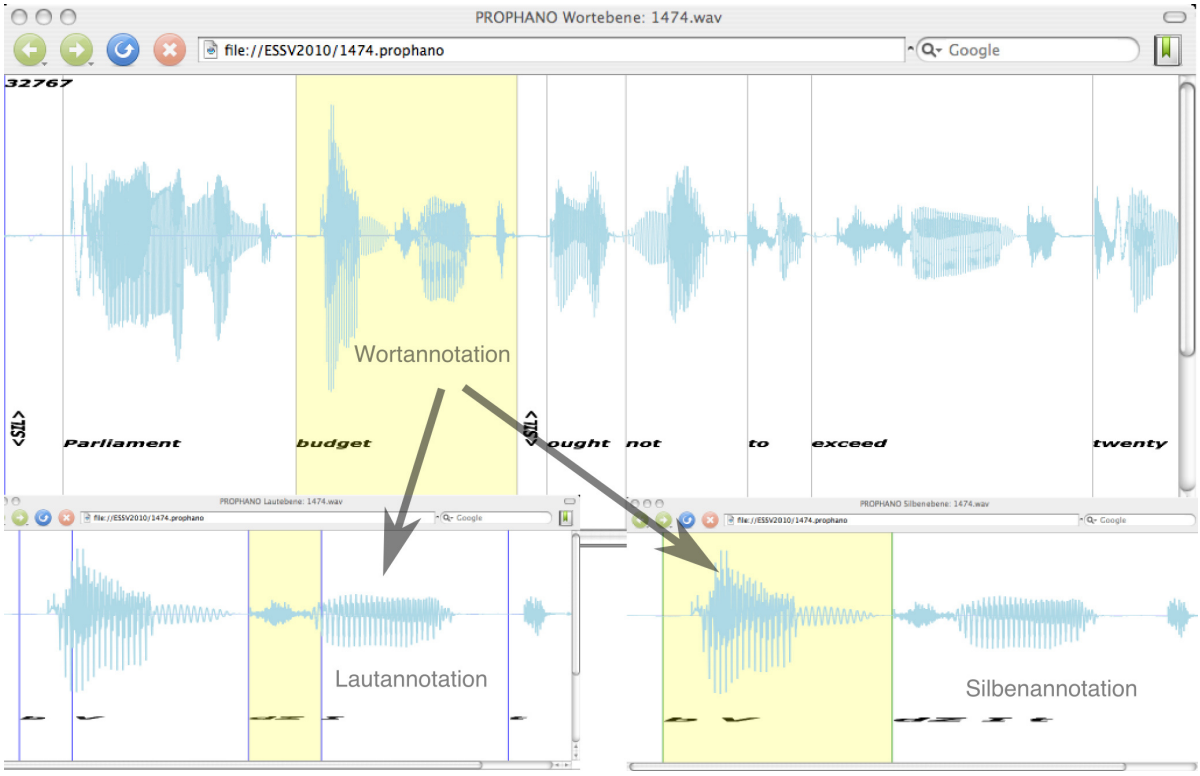

Abbildung 3 - Darstellung und Auswahl der Annotationsebene direkt im Browser (SVG-Grafiken)

Literatur

[1] O. Jokisch, S. Wittenberg, M. Cuevas, H. Hussein, G. Strecha, H. Ding and R. Hoffmann: "Towards an Automatic Process Chain for the Speech Corpora Annotation". Proc. SPECOM 2007, Moscow, October 2007.

[2] H. Ding and O. Jokisch: "An investigation of the pronunciation of English words in German SMS texts", In: Elektronische Sprachsignalverarbeitung, Tagungsband der 20. Konferenz 2009, Studientexte zur Sprachkommunikation 53, S. 396-402.

[3] U. Koloska, D. Hirschfeld: "Integrierte multimodale Sprachdatenverwaltung mit dem WiGE System". In: Elektronische Sprachsignalverarbeitung, Tagungsband der 15. Konferenz 2004, Studientexte zur Sprachkommunikation 30, S. 222-228.

[4] M. Hofmann, O. Jokisch: "Optimization of MFNs for signal-based phrase break prediction". Proc. 3rd Intern. Conference on Speech Prosody (SPEECH PROSODY), vol. II, 636-639, Dresden (Germany), May 2006.

[5] J.C. Wells: "SAMPA computer readable phonetic alphabet". In Gibbon, D., Moore, R. and Winski, R. (eds.), Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B, 1997. http://www.phon.ucl.ac.uk/home/sampa

[6] M. Schröder, S. Breuer: "XML Representation Languages as a Way of Interconnecting TTS Modules". Proc. ICSLP'04 Jeju, Korea, 2004