Doctoral Thesis

A Data-driven Model for the Generation of Prosody from Syntactic Sentence Structures

Author(s):
Hoffmann, Sarah

Publication Date:
2014

Permanent Link:

https://doi.org/10.3929/ethz-a-010337573 →

Dissertation ETH No. 21991

A Data-driven Model for the Generation of Prosody from Syntactic Sentence Structures

A dissertation submitted to the
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
SARAH HOFFMANN
Dipl. Inf.
born December 18, 1978
citizen of Germany

accepted on the recommendation of
Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Bernd Möbius, co-examiner
Dr. Beat Pfister, co-examiner

2014

Abstract

Prosody describes the aspects of speech that go beyond its basic phonetic content. They include the melody and rhythm of speech, loudness and similar properties. Prosody is a vital part of speech that simplifies understanding and often serves as an additional information channel. While modern speech-synthesis systems are able to produce speech with prosody that sounds correct, they still lack the ability to make it lively enough so that the prosody does not immediately betray the artificial nature of the speech. This thesis aims to explore one the functions of prosody in synthesis, namely that of structuring speech. To that end, it proposes a more direct use of the syntax structure of the text for the generation of prosody.

A new hierarchic approach to prosody generation is introduced that produces prosody directly from the syntax structure of the sentence. This is achieved by defining elementary prosody generation functions, called mutators, that describe the local contribution of a single node in the syntax tree to the global prosody of the sentence. We show that these mutator functions can be trained from natural speech and then combined according to the syntax structure of the sentence for the prosody generation. The result is a complex, yet flexible prosody generation model that can produce natural sounding prosody. In contrast to existing methods, the mutator functions directly generate physical prosodic parameters like fundamental frequency and duration, although stylized on a word level. This means that abstract prosody concepts like accent and phrase are represented only indirectly in the model, making it easier to learn new prosody models from arbitrary examples of natural speech and allowing to express more fine-grained the degrees of prosodic expression. The thesis explores different word stylizations

and shows that they can capture the prosody sufficiently.

This thesis also describes in detail the process necessary to prepare natural speech data for the training of new prosody models using the example of audio book data. We discuss normalization of punctuation and text formatting for diverse text sources and efficient parsing of the texts. The thesis then presents an algorithm for phone segmentation of long speech recordings based entirely on forced alignment between text and speech using hidden Markov models. We show the algorithm to be mostly self-contained, requiring only a generalized language independent phone model that can be trained on any other available speech corpus, including on corpora in a different language.

This work concludes with a perceptual evaluation of six different prosody models trained on German and English audio books from the Librivox project. We show that the prosody produced by our model is preferred over a neutral prosody based on a simple accent/phrase model.

Kurzfassung

Prosodie beschreibt den Teil der Sprache, der über den eigentlichen phonetischen Inhalt hinausgeht. Das beinhaltet Melodie und Rhythmus der Sprache, sowie Eigenschaften wie Lautheit. Prosodie macht einen wichtigen Teil der gesprochenen Sprache aus, der die Verständigung erleichtert und oft auch als sekundärer Informationskanal dient. Obwohl moderne Sprachsynthese-Systeme weitgehend in der Lage sind, korrekte Prosodie zu erzeugen, fehlt ihnen nach wie vor die Fähigkeit eine Lebendigkeit und Abwechslung zum Ausdruck zu bringen, die menschliche Sprache auszeichnet. Diese Arbeit beschäftigt sich mit einer der Funktionen von Prosodie für die Synthese, nämlich die gesprochene Sprache zu strukturieren. Dabei geht es darum, die Syntaxstruktur eines Textes direkter in die Generierung der Prosodie einfliessen zu lassen.

Es wird eine hierarchischer Ansatz zur Prosodie-Generierung entwickelt, mit dem Prosodie in direkter Relation zur Sytnaxstruktur des Satzes berechnet wird. Dazu werden elementare Funktionen definiert, sogenannte Mutationsfunktionen, die den lokalen Einfluss eines jeden Knotens im Syntaxbaum des Satzes auf die Prosodie beschreiben. Es wird gezeigt, dass die Mutationsfunktionen anhand natürlicher Sprache trainiert und dann kombiniert werden können, um die Prosodie komplexer Sätze zu erzeugen. Im Unterschied zu existierenden Methoden generieren die Mutationsfunktionen direkt physikalische Prosodieparameter des Sprachsignals wie Grundfrequenz und Lautdauer, jedoch auf Wordebene abstrahiert. Das bedeutet, dass abstrakte linguistische Konzepte wie Akzent und Phrase nur noch indirekt repräsentiert werden, wodurch das Lernen neuer prosodischer Ausdrucksweisen vereinfacht wird. Die Arbeit untersucht verschiedene Wordabstraktionen und Trainingsverfahren für die Mutationsfunktionen.

Desweiteren wird im Detail die automatische Aufbereitung von natürlichsprachlichem Trainingsmaterial im Allgemeinen und Audiobüchern im Speziellen behandelt. Es wird diskutiert, wie der Text bezüglich Satzzeichen und Formatierung normalisiert und dann effizient geparst werden kann. Die Arbeit präsentiert dann eine Methode zur phonetischen Segmentierung von langen Sprachaufnahmen, die vollständig auf Forced-Alignment mit Hidden-Markov-Modellen basiert. Die Methode benötigt nur einen kleinen externen Training-Corpus, der auch in einer anderen Sprache sein kann.

Die Arbeit schliesst mit einer subjektiven Evaluation von sechs verschiedenen Prosodie-Modellen, die an englischen und deutschen Audiobüchern des Librivox-Projekts trainiert worden. Es wird gezeigt, dass das entwickelte Model einem einfachen Phrasen/Akzent-basierten Model vorgezogen wird.