ETH zürich

ETH Library

Messung und Modellierung der Segmentdauer für die Synthese deutscher Lautsprache

Doctoral Thesis

Author(s):
Huber, Karl

Publication date:
1991

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-a-000622667

Rights / license:
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

DISS. ETH Nr. 9535

Messung und Modellierung der Segmentdauer für die Synthese deutscher Lautsprache

ABHANDLUNG
Zur Erlangung des Titels
DOKTOR DER TECHNISCHEN WISSENSCHAFTEN

der
EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE
ZÜRICH

vorgelegt von
KARL HUBER
Dipl. El.-Ing. ETH
geboren am 15. Dezember 1957
von Hohenrain LU

Angenommen auf Antrag von:
Prof. Dr. W. Guggenbühl, Referent
Prof. Dr. G. Moschytz, Korreferent

1991

i. G. Guggenbühl

Kurzfassung

Der Hintergrund dieser Arbeit ist die *Sprachsynthese* basierend auf der Verknüpfung von *Diphonen*. Die *segmentelle* (oder lautliche) Komponente wird dabei getrennt von der *suprasegmentellen* (oder *prosodischen*) Komponente betrachtet. Als akustische Parameter der *Prosodie* werden die *Grundfrequenz*, die *lautliche Dauer* und die *Leistung* angesehen. Diese Arbeit befasst sich mit der Dauer der lautlichen Segmente. Sie umfasst die Messung der Dauer in natürlicher Lautsprache, das Anpassen eines statistischen Modells an die Daten und die Prädiktion von Dauerwerten für die Synthese lautlicher Sprache.

Eine erste Schwierigkeit ist die Messung der Dauer der Segmente in natürlicher Lautsprache. Als Einheiten der Messung werden nicht Laute, sondern die Grundelemente der Synthese – die Diphone – verwendet. Zur Messung dieser Einheiten in natürlicher Lautsprache wird ein Verfahren eingesetzt, das auf *Dynamischer Programmierung* beruht. Es geht aus von der Voraussage der in der Äusserung zu erwartenden phonetischen Sequenz und allfälligen *allophonischen Varianten*. Ausserdem wird für die Messung eine Lautdauerprädiktion eingesetzt, die einen Bereich vorgibt, in dem ein Messwert zu erwarten ist. Diese Voraussage der Dauer macht den Messprozess effizienter und sicherer.

Für die statistische Analyse der Messdaten wird ein *Verallgemeinertes Lineares Modell* (*Generalized Linear Model*) eingesetzt. Dies ist eine Erweiterung der linearen Regression. Die Erweiterung besteht erstens darin, dass *kategorielle Grössen* (neben kontinuierlich veränderbaren Grössen) als *erklärende Variablen* einbezogen werden können, zweitens kann die *Zielgrösse* einer andern als der *Normalverteilung* genügen.

Ein Hauptpunkt beim Erstellen eines Lautdauermodells ist die Formulierung und Beurteilung der Faktoren, die die Lautdauer beeinflussen. Die Literatur zur Lautdauer, die heute schon recht umfangreich ist, vermittelt ein gutes Bild von diesen Einflüssen. Aus der Kenntnis der möglichen Einflüsse lässt sich ein einheitliches Schema für den Einbezug der Einflussgrössen aufstellen. Das statistische Modell bietet seinerseits Hilfestellung in der Beurteilung der Wirkung einzelner Einflüsse.

Es werden Resultate der Analyse von dreissig längeren deutschen Sätzen vorgestellt. Unter anderem ist ein bemerkenswertes Resultat die Wirkung des Sprechrhythmus. Die Analyse zeigt, dass ein Segment innerhalb einer rhythmischen Einheit die Tendenz hat, kürzer zu werden, je weiter entfernt sie von der Akzentstelle liegt.

Schliesslich wird die Art und Weise erläutert, wie die Dauervoraussage in der Synthese eingesetzt und wie sie zur Erstellung von Dauerbereichen in der Lautdauermessung benutzt wird.

Die folgenden Punkte werden als zentrale Beiträge der vorliegenden Arbeit angesehen:

- Die Behandlung der segmentellen Einheiten ist nicht wie sonst üblich lautbezogen, sondern erstreckt sich abhängig vom phonetischen Kontext über einen kürzeren oder längeren Bereich. Das Problem mit den unsicheren Lautgrenzen wird damit umgangen.
- Die Methode zur automatischen Ausmessung der segmentellen Dauer basiert selber auf einer Prädiktion der Dauer. Es werden dieselben segmentellen Einheiten verwendet wie bei der Synthese. Analyse und Synthese bilden eine abgeschlossene Einheit.
- Die Anwendung eines Verallgemeinerten Linearen Modelles der Statistik für die Beschreibung der Segmentdauer ist erstmalig und zeigt eine Reihe von Vorteilen gegenüber andern Modellen.

Abstract

Research on *segmental duration* in speech has been carried out on the background of a speech-synthesis project. In this project, speech production (the *segmental component*) is based on *diphone concatenation.* It was assumed that the *suprasegmental level* (*prosody level*) may be treated independently from the segmental component (the sound characteristics). The acoustic parameters of prosody are *fundamental frequency*, *duration*, and *intensity.* Research described in this dissertation is concerned with duration of speech segments. It comprehends duration measurements in natural speech, the fitting of a statistical model to the measurement data, and prediction of duration in speech synthesis.

Duration measurement uses the *diphone inventory* as a reference. The procedure starts from a prediction of *allophonic elements* that are likely to occur in the test utterance. These allophones are converted to reference elements. A duration range is assigned to each reference element using the prediction device developed in the course of the project. All the input information is represented in a *directed acyclic graph*, which is mapped on the test signal by means of *dynamic programming techniques.*

The class of *Generalized Linear Models* is found to be well suited for statistical modelling of duration. Generalized Linear Models allow for *explanatory variables* that are *categorical* (apart from continuously valued). They also allow for a *response variable* that does not need to be of *normal distribution.* One of the main points in a duration model is the formulation and evaluation of factors that affect duration. There is a considerable quantity of literature reporting on such factors. A general scheme is built up, which accounts for potential factors in a duration

model. The statistical model provides means for the evaluation of the significance of the individual factors.

The model is fitted to data obtained from 30 lengthy German sentences. One of the interesting results concerns rhythmic effects. Analysis reveals that within a rhythmic unit, duration tends to decrease for positions farther from the accentuated syllable.

Finally, it is described how duration prediction is used in speech synthesis as well as in duration measurement.

The issues considered significant about this work are the consequent design of all components (measurement, modelling) in the context of speech synthesis and the use of adequate algorithms for measurement, modelling, and prediction of segmental duration.