KTH Computer Science
and Communication

Modelling Phone-Level Pronunciation in Discourse Context

Per-Anders Jande

Doctoral Thesis
Stockholm, Sweden, 2006

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framläggs till offentlig granskning för avläggande av filosofie doktorsexamen måndagen den 11 december klockan 13.00 i sal F3, Kungliga Tekniska Högskolan, Lindstedtsvägen 26, Stockholm.

TRITA-CSC-A 2006:25 · ISSN 1653-5723 · ISRN KTH/CSC/A–06/25–SE · ISBN 91-7178-490-X

Abstract

Analytic knowledge about the systematic variation in a language has an important place in the description of the language. Such knowledge is interesting e.g. in the language teaching domain, as a background for various types of linguistic studies, and in the development of more dynamic speech technology applications. In previous studies, the effects of single variables or relatively small groups of related variables on the pronunciation of words have been studied separately. The work described in this thesis takes a holistic perspective on pronunciation variation and focuses on a method for creating general descriptions of phone-level pronunciation in discourse context. The discourse context is defined by a large set of linguistic attributes ranging from high-level variables such as speaking style, down to the articulatory feature level. Models of phone-level pronunciation in the context of a discourse have been created for the central standard Swedish language variety. The models are represented in the form of decision trees, which are readable for both machines and humans. A data-driven approach was taken for the pronunciation modelling task, and the work involved the annotation of recorded speech with linguistic and related information. The decision tree models were induced from the annotation. An important part of the work on pronunciation modelling was also the development of a pronunciation lexicon for Swedish. In a cross-validation experiment, several sets of pronunciation models were created with access to different parts of the attributes in the annotation. The prediction accuracy of pronunciation models could be improved by 42.2% by making information from layers above the phoneme level accessible during model training. Optimal models were obtained when attributes from all layers of annotation were used. The goal for the models was to produce pronunciation representations representative for the language variety and not necessarily for the individual speakers, on whose speech the models were trained. In the cross-validation experiments, model-produced phone strings were compared to key phonetic transcripts of actual speech, and the phone error rate was defined as the share of discrepancies between the respective phone strings. Thus, the phone error rate is the sum of actual errors and discrepancies resulting from desired adaptations from a speaker-specific pronunciation to a pronunciation reflecting general traits of the language variety. The optimal models gave an average phone error rate of 8.2%.

TRITA-CSC-A 2006:25 · ISSN 1653-5723 · ISRN KTH/CSC/A–06/25–SE · ISBN 91-7178-490-X