Pronunciation variation modelling using decision tree induction from multiple linguistic parameters

Per-Anders Jande
KTH: Department of Speech, Music and Hearing/CTT – Centre for Speech Technology

Abstract

In this paper, resources and methods for annotating speech databases with various types of linguistic information are discussed. The decision tree paradigm is explored for pronunciation variation modelling using multiple linguistic context parameters derived from the annotation. Preliminary results suggest that decision tree induction is a suitable paradigm for the task.

Introduction

The pronunciation of words varies depending on the context in which they are uttered. A general model describing this variation can be useful e.g. for increasing the naturalness of synthetic speech of different speech rates and for simulating different speaking styles.

This paper describes some initial attempts at using the decision tree learning paradigm with multiple linguistic context parameters for creating models of pronunciation variation for central standard Swedish. The context parameters are derived from annotated speech data. Only pronunciation variation on the segment level is considered. Pronunciation in context is described in relation to a canonical reference transcription.

Background

General reduction phenomena have been described for Swedish e.g. by Gårding (1974), Bruce (1986) and Bannert and Czigler (1999). Jande (2003a; b) describes a reduction rule system building partly on these studies. This rule system was used for improving the naturalness of fast speech synthesis. Evaluations showed that reduced pronunciations were perceived as more natural than the default canonical transcriptions when the rate of the synthetic speech was above the synthesis default rate.

However, there were indications of word predictability (global word frequency) also influencing the perceived naturalness of the reduced word pronunciations. High frequency words showed a bias towards being preferred in their reduced form, irrespective of the speech rate. Low frequency words showed the opposite bias. This was not surprising, since word predictability has been shown in many studies for several languages to influence local speech rate and distinctness of pronunciation. Many other types of linguistic context also influence the pronunciation of words. Thus, including many types of linguistic information as context is necessary for creating a generally successful pronunciation variation model.

Annotation

For the purpose of studying the impact of e.g. variables influencing word predictability on the pronunciation of words in context, speech data is annotated with a variety of information potentially influencing segment level word realisation.

To a large extent, the annotation is supplied using automatic methods. Making use of automatic methods is of the essence, since manual annotation is very time consuming. This section gives a short description of the types of information provided and the tools and resources used for annotation.

Source Data

The data discussed in this paper is the annotation of the VaKoS spoken language database (Bannert and Czigler, 1999). This database consists of approximately 103 minutes of spontaneous speech from ten speakers of central standard Swedish. There is about ten minutes of spoken monologue from each speaker. The speech is segmented by hand on the word level and partly segmented on the phone level. There are also various other types of annotation.

The manually provided orthographic transcriptions and word boundaries are collected from the database together with information about prosodic boundaries, focal stress, hesitations, disfluencies (word fragments) and speaker gender. Automatic methods are used for providing a variety of other types of linguistic information serving as tentative predictors of the segmental realisation of words.

Pronunciation in context is modelled in relation to an automatically supplied canonical reference transcription and thus all annotation is aligned to canonical transcriptions, creating one data point per canonical segment.

Information

Information is provided at five levels of description, corresponding to linguistic units of different sizes. At the ***utterance level***, speaker gender information is provided. The annotation at the ***phrase level*** consists of phrase type tags and some different measures of phrase length and phrase prosodic weight.

The ***word level*** annotation consists of measures of word length, part of speech tags, word type information (function or content), estimations of global word frequencies weighted with collocation weights and the number of full form word and lexeme repetitions thus far in the discourse. Also supplied is information about the position of a word in the current phrase and in a collocation, information about focal stress, estimated word mean relative speech rate and information about adjacent hesitation sounds, word fragments and prosodic boundaries.

The annotation at the ***syllable level*** consists of information about syllable length, the position of the syllable in the word, the nucleus of the syllable, word stress, stress type and the estimated relative speech rate of the syllable.

At the ***segment level***, the annotation includes the identity of the canonical segment, a set of articulatory features describing the canonical segment and the position of the segment in the syllable (onset, nucleus or coda). There is also information about the position of a segment in a cluster and about the length of the current cluster. Finally, the identity of the detailed segment is included. The detailed segment identities are determined automatically and will need manual correction. However, the initial tests of decision tree inducers were conducted using the uncorrected transcriptions.

Annotation Resources

Canonical (signal independent) phonological transcriptions of the words in the database are produced by a system for automatic time-aligned phonetic transcription developed by Sjölander (2003), adapted to be able to use manually determined word boundaries.

A net describing tentative detailed (signal dependent) transcriptions is generated using a list of possible detailed realisations for each canonical segment. Segment HMMs and alignment tools developed by Sjölander (2003) are used for finding the detailed transcription with the optimal match to the signal. The detailed transcriptions are aligned to the canonical transcriptions using `null` symbols as placeholders for deleted segments.

Global word frequencies and collocation weights were estimated using the Göteborg Spoken Language Corpus (cf. e.g. Allwood, 1999), including roughly three million words of orthographically transcribed spoken language from different communicative situations.

The TnT tagger (Brants, 2000) trained on Swedish text (Megyesi, 2002) is used for part of speech tagging and the SPARK-0.6.1 parser (Aycock, 1998), with a context free grammar for Swedish written by Megyesi (2002) is used for chunking the transcribed and part of speech tagged orthographic transcriptions into phrase units.

Decision Tree Induction

For data-driven development of pronunciation variation models, machine learning methods of some type are necessary. For developmental purposes, it is preferred that the model can be represented on a human-understandable format. The decision tree induction paradigm is used to induce models on a tree format and the tree structures can be converted into human-readable rules.

A decision tree classifier is used to classify instances based on their sets of description parameters. In most cases, decision tree learning algorithms induce tree structures from data employing a best split first tactic. This means that the parameter used for splitting the data set at each node is the one that divides the set into the most separate groups (as determined e.g. by some entropy-based measure).

Decision Tree Learning Algorithms

Some freely available, open source decision tree learner implementations based on or similar to the C4.5 algorithm (Quinlan, 1993) were tested. The same training and test data (although on different formats for the different implementations) were used in each case. The C4.5 default splitting criterion (Gain Ratio) and settings for pruning (confidence level pruning with a confidence level of 0.25) were also used in each case. Each implementation offers its own optimisation options and the results re-

ported are not guaranteed to be optimal. However, in these initial tests, it was mainly the general suitability of the decision tree induction paradigm for the suggested task that was evaluated.

The implementations tested were Quinlan's *C4.5* decision tree learner, release 8[1], the Tilburg Memory-Based Learner (TiMBL) version 5.0[2], which is able to produce C4.5 type tree representations if the *IGTree* option is used, a "slightly improved" implementation of C4.5 called *J4.8* included in the University of Waikato Environment for Knowledge Analysis (Weka) machine learning toolkit for java version 3.4.1[3] and Christian Borgelt's reduction and decision tree implementation *Dtree*[4]. Some other implementations were also explored, but turned out not able to induce trees from the type of data at hand.

Training and Evaluation Data

The training data was compiled using the linguistic annotation provided for the VaKoS database. One parameter vector per canonical segment was composed, each vector containing 118 slots – 117 containing context attributes and one containing the class (detailed segment). The context attributes were the attributes of different linguistic units and attributes describing the sequential context of the units (i.e., the values to the left or to the right of the current unit at the current description level). Since not all decision tree implementations could handle continuous numerical values, all data was quantised so that the parameter vectors only contained discrete variables. This means that e.g. relative speech rate was described as *high*, *medium* or *low* in the parameter vectors.

The canonical transcriptions of the VaKoS data contained 55,760 segments and thus this was the number of parameter vectors created. The vector set was divided into 90% training data and 10% evaluation data using random sampling.

Results

Although not identical, the decision tree implementations all showed similar results in their ranking of parameters (split order) and in terms of prediction accuracy. As could be expected, attributes describing the phonological features of the canonical segment and the adjacent canonical segments were ranked the highest. Among the highest ranked attributes were also cluster length, the position in the cluster and cluster type. Other relatively high ranked attributes were hesitation context, syllable length, part of speech, disfluency context and syllable stress. Attributes that were generally ranked low were local speech rate estimates (perhaps due to quantisation effects), speaker gender and all phrase level attributes. The segment feature *sonority*, dividing vowels and consonants, was used for the fist split by all implementations.

The segment error rate was around 40%, ranging from 38.8% to 43.0%. The detailed segment classifications used in training had not been manually corrected and initial evaluations imply that the performance of the detailed transcription algorithm was not optimal. Thus, the error rate of the decision tree classifiers trained on the tentative classifications is only a rough estimate of the error rate for trees trained with manually corrected classifications. However, although manual correction of the detailed transcriptions will probably introduce some types of variability that cannot be produced by the detailed transcription algorithm, initial evaluations suggest that the correspondence between the canonical and the detailed transcriptions will actually be higher in the final training data. If this holds, the final data will be more consistent and the segment level attributes will be even better predictors of detailed pronunciation.

The particular decision tree inducer implementations all had their pros and cons. Two algorithms had problems with insufficient memory at tree induction and at tree-to-rule conversion, respectively. Possible solutions to these problems have to be investigated. Prediction accuracy will be the first consideration when it comes to choosing an implementation. Which algorithm or algorithms to use will be clear when the optimisation options of the different algorithms are explored. However, all in all, the decision tree paradigm seems to be useful for the type of pronunciation variation modelling suggested.

Conclusions

Spoken language data has been annotated with various types of linguistic information, mostly with automatic means. The information has been used to create training data for decision tree induction. Some different decision tree learners have been tested to evaluate the suitability of the decision tree induction paradigm for pronunciation variation modelling using multiple linguistic parameters. The results suggest that decision trees are suitable for the task.

Future Work

This paper presents work in progress. As the result of the exploration of resources and methods for annotation, one database has been fully annotated. For the final pronunciation variation model, more speech data reflecting different speaking styles will be included. Databases available and partially annotated include human-computer dialogues, human-human dialogues, monologues and read aloud texts. With more varied speech data, a discourse annotation level with speaking style classifications will be included. Speaker age group will be included at the utterance level (when available) as well as the utterance mean relative speech rate.

Much of the information provided with automatic methods depends on the presence of manually determined word boundaries. Such boundaries are not available for most databases. However, orthographic transcriptions are available. This means that an automatic alignment system (e.g. Sjölander, 2003) can be used and the output manually corrected. Information about prosodic boundaries and focal stress is available only for some of the speech databases. Supplying this information for all speech data will require some manual work, although the work can probably be facilitated through some degree of automation.

Initial evaluations of the detailed transcriptions suggest that the error rate of the detailed transcription algorithm can be reduced by restricting the list of possible realisations for some segments to only the most common ones. The detailed transcription algorithm will be optimised and the output manually corrected. Also, more machine learning paradigms will be evaluated, starting with other rule induction methods.

Qualitative evaluations of the decision tree classifications will have to be conducted and good evaluation measures developed. Different types of errors should be weighted for their gravity, using some (context dependent) phonetic distance measure. Some decision tree inducers allow different severity weights for different classification errors. This kind of error measure could thus also be used for model induction.

Finally, it would be interesting to evaluate the model using synthetic speech. In a synthesis implementation, the parameters will have to be either supplied by the user or estimated from the input. Redundancy and co-variation between parameters will have to be investigated in order to make the best use of the information that can be made available in a synthesis context.

Notes

1. *www.cse.unsw.edu.au/~quinlan/*
2. *ilk.kub.nl/software.html*
3. *www.cs.waikato.ac.nz/ml/weka/*
4. *fuzzy.cs.uni-magdeburg.de/~borgelt/doc/dtree/dtree.html*

Acknowledgements

Many thanks to Kåre Sjölander and Bea Megyesi for their help with the annotation and to Robert Bannert and Peter Czigler for making their VaKoS database available. The research reported in this paper was carried out at the Centre for Speech Technology (CTT) at KTH.

References

Allwood, J. (1999) The Swedish spoken language corpus at Göteborg University. Proc Fonetik 1999.

Aycock, J. (1998) Compiling little languages in Python. Proc 7th International Python Conference.

Bannert, R. and Czigler, P. E. (1999) Variations in consonant clusters in standard Swedish. Phonum 7, Umeå University.

Brants, T. (2000) TnT – A statistical part-of-speech tagger. Proc 6th ANLP.

Bruce, G. (1986) Elliptical phonology. Papers from the Ninth Scandinavian Conference on Linguistics, 86–95.

Gårding, E. (1974) Sandhiregler för svenska konsonanter. Svenskans beskrivning 8, 97–106.

Jande, P-A (2003a) Evaluating rules for phonological reduction in Swedish. Proc Fonetik 2003, 149–152.

Jande, P-A (2003b) Phonological reduction in Swedish. Proc 15th ICPhS, 2557–2560.

Megyesi, B. (2002) Data-driven syntactic analysis – Methods and applications for Swedish. Ph. D. Thesis. KTH, Stockholm.

Sjölander, K. (2003) An HMM-based system for automatic segmentation and alignment of speech. Proc Fonetik 2003, 193–196.

Quinlan, J. R. (1993) C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann.