Modelling Pronunciation in Discourse Context

Per-Anders Jande
Dept. of Speech, Music and Hearing/CTT, KTH
jande@speech.kth.se

Abstract

This paper describes a method for modelling phone-level pronunciation in discourse context. Spoken language is annotated with linguistic and related information in several layers. The annotation serves as a description of the discourse context and is used as training data for decision tree model induction. In a cross validation experiment, the decision tree pronunciation models are shown to produce a phone error rate of 8.1% when trained on all available data. This is an improvement by 60.2% compared to using a phoneme string compiled from lexicon transcriptions for estimating phone-level pronunciation and an improvement by 42.6% compared to using decision tree models trained on phoneme layer attributes only.

1 Introduction and background

The pronunciation of a word is dependent on the discourse context in which the word is uttered. The dimension of pronunciation variation under study in this paper is the phone dimension and only variation such as the presence or absence of phones and differences in phone identity are considered. The focus is on variation that can be seen as a property of the language variety rather than as individual variation or variation due to chance.

Creating models of phone-level pronunciation in discourse context requires a detailed description of the context of a phoneme. Since the discourse context is the entire linguistic and pragmatic context in which the word occurs, the description must include everything from high-level variables such as speaking style and over-all speech rate to low-level variables such as articulatory feature context.

Work on pronunciation variation in Swedish has been reported by several authors, e.g. Gårding (1974), Bruce (1986), Bannert & Czigler (1999), Jande (2003; 2005). There is an extensive corpus of research on the influence of various context variables on the pronunciation of words. Variables that have been found to influence the segmental realisation of words in context are foremost speech rate, word predictability (or word frequency) and speaking style, cf. e.g. Fosler-Lussier & Morgan (1999), Finke & Waibel (1997), Jurafsky et al. (2001) and Van Bael et al. (2004).

2 Method

In addition to the variables mentioned above, the influence of various other variables on the pronunciation of words has been studied, but these have mostly been studied in isolation or together with a small number of other variables. A general discourse context description for recorded speech data, including a large variety of linguistic and related variables, will enable data-driven studies of the interplay between various information sources on e.g. phone-level pronunciation. Machine learning methods can be used for such studies. A model of pronunciation variation created through machine learning can be useful in speech technology applications, e.g. for creating more dynamic and natural-sounding speech synthesis. It is possible to create models which can predict the pronunciation of words in context and which are simultaneously descriptive and to some degree explain the interplay between different types of variables involved in the predictions. The decision tree induction paradigm is a machine learning method that is suitable for training on variables of diverse types, as those that may be included in a general description of discourse context. The paradigm also creates transparent models. This paper describes the creation of pronunciation models using the decision tree paradigm.

2.1 Discourse context description

The speech databases annotated comprise ~170 minutes of elicited and scripted speech. Canonical phonemic word representations are collected from a pronunciation lexicon and the phoneme is used as the central unit in the pronunciation models. The annotation is aimed at giving a general description of the discourse context of a phoneme and is organised in six layers: 1) a discourse layer, 2) an utterance layer, 3) a phrase layer, 4) a word layer, 5) a syllable layer and 6) a phoneme layer. Each layer is segmented into a linguistically meaningful type of unit which can be aligned to the speech signal and the information included in the annotation is associated with a particular unit in a particular layer. For example, in the word layer, information about part of speech, word frequency, word length etc. is included. The information associated with the units in the phoneme layer is instead phoneme identity, articulatory features etc. For a more detailed description of the annotation, cf. Jande (2006).

2.2 Training data

Decision trees are induced from a set of training instances compiled from the annotation. The training instances are phoneme-sized and can be seen as a set of *context sensitive phonemes*. Each training instance includes a set of 516 attribute values and a phone realisation, which is used as the classification key. The features of the current unit at each layer of annotation are included as attributes in the training examples. Where applicable, information from the neighbouring units at each annotation layer is also included in the attribute sets.

The key phone realisations are generated by a hybrid automatic transcription system using statistical decoding and a posteriori correction rules. This means that there is a certain degree of error in the keys. When compared to a small gold standard transcription, the automatic transcription system was shown to produce a phone error rate (PER) of 15.5%. Classification is not always obvious at manual transcription, e.g. many cases of choosing between a full vowel symbol and a schwa. Defaulting to the system decision whenever a human transcriber is forced to make ad hoc decisions would increase the speed of manual checking and correction of automatically generated phonetic transcripts without lowering the transcript quality. If this strategy had been used at gold standard compilation, the estimation of the system accuracy would have been somewhat higher. The 15.5% PER is thus a pessimistic estimate of the transcription system performance.

2.3 Decision tree model induction

Decision tree induction is non-iterative and trees are built level by level, which makes the learning procedure fast. However, the optimal tree is not guaranteed. At each new level created during the tree induction procedure, the set of training instances is split into subsets according to the values of one of the attributes. The attribute selected is the attribute that best meets a given criterion, generally based on entropy minimisation. Since training data mostly contain some degree of noise, a decision tree may be biased toward the noise in the training data (over-trained). However, a tree can be pruned to make it more generally applicable. The idea behind pruning is that the most common patterns are kept in the model, while less common patterns, with high probability of being due to noise in the training data, are deleted.

3 Model performance

A tenfold cross validation procedure was used for model evaluation. Under this procedure, the data is divided into ten equally sized partitions using random sampling. Ten different decision trees are induced, each with one of the partitions left out during training. The partition not used for training is then used for evaluation. A pruned and an unpruned version of each tree were created and the version with the highest prediction accuracy on the evaluation data was used for calculating the average prediction accuracy. The annotation contains some prosodic information (variables based on pitch and duration measures calculated from the signal), which cannot be fully exploited in e.g. a speech synthesis context. Thus, it was interesting to investigate the influence of the prosodic information on model performance. For this purpose, a tenfold cross validation experiment where the decision tree inducer did not have access to the prosodic information was performed. As a baseline, an experiment with trees induced from phoneme layer information only was also performed.

3.1 Results

Attributes from all layers of annotation were used in the models with the highest prediction accuracy. The topmost node of all trees was *phoneme identity* and other high ranking attributes were phoneme context, *mean phoneme duration* measured over the word and over the phrase, and *function word*, a variable separating between a generic content word representation and the closed set of function words. The trees produced an average phone error rate of 8.1%, which is an improvement by 60.2% compared to using a phoneme string compiled from a pronunciation lexicon for estimating the phone-level realisation.

The average PER of models trained on phoneme layer attributes only was 14.2%, which means that the prediction accuracy was improved by 42.6% by adding attributes for units above the phoneme layer to the training instances. A comparison between the models trained on all attributes and the models trained without access to prosodic information showed that the prosodic information gave a decrease in PER from 13.1 to 8.1% and thus increased model performance by 37.8%.

The phonetic transcript generated by the models trained on all attributes was also evaluated against actual target transcripts, i.e., the gold standard used to evaluate the automatic transcription system. In this evaluation, the models produced a PER of 16.9%, which means that the deterioration in performance when using an average decision tree model instead of the automatic transcription system is only 8.5% and that the improvement using the model instead of a phoneme string is 34.9%.

4 Model transparency

Figure 1 shows a pruned decision tree trained on all available data. The tree uses 58 of the 516 available attributes in 423 nodes on 12 levels. The transparency of the decision tree representation becomes apparent from the magnification of the leftmost sub-tree under the top node, shown in the lower part of Figure 1.

The top node of the tree is *phoneme identity* and the magnified branch is the branch representing phoneme identity /v/. It can be seen that there are two possible realisations of the phoneme /v/, [v] and `null` (no realisation) and it is easy to understand the conditions under which the respective realisations are used. If the mean phoneme duration over the word is less than 35.1 ms, the /v/ is never realised. If the mean phoneme duration is between 31.5 and 38.2 ms, the current word is decisive. If the word is one of the function words *vad*, *vi*, *vara*, *vid*, or *av*, the /v/ is not realised. If the word is any content word or the function word *blev*, the /v/ is realised as [v]. Finally, if the mean phoneme duration over the word is more than 38.2 ms, the /v/ is realised (as [v]) unless the phoneme to the right is also a /v/.

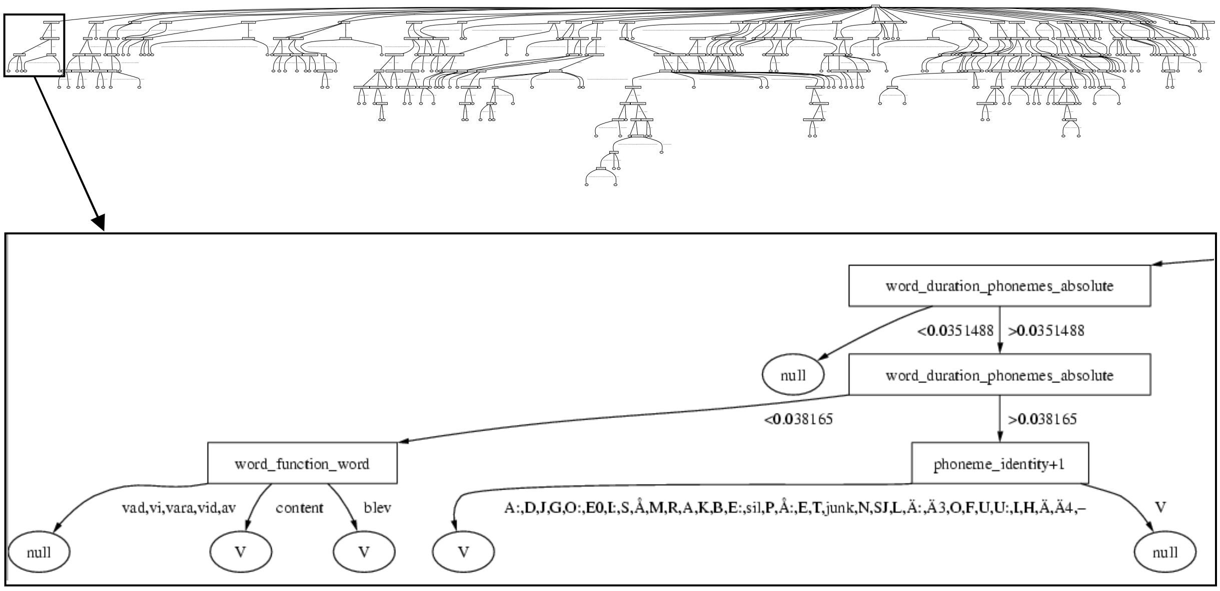
word_duration_phonemes_absolute
<0.0351488
>0.0351488
null
word_duration_phonemes_absolute
<0.038165
>0.038165
word_function_word
phoneme_identity+1
vad,vi,vara,vid,av
content
blev
A:,D,J,G,O:,E0,I:,S,Å,M,R,A,K,B,E:,sil,P,Å:,E,T,junk,N,SJ,L,Ä:,Ä3,O,F,U,U:,I,H,Ä,Ä4,–
V
null
V
V
V
null

Figure 1. The upper part of the figure shows the pruned version of a decision tree and the lower part of the figure shows a magnification of a part of the tree.

Acknowledgements

The research reported in this paper is carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (the Swedish Agency for Innovation Systems), KTH and participating companies and organisations. The work was supported by the Swedish Graduate School of Language Technology, GSLT.

References

Bannert, R. & P.E. Czigler, 1999. Variations in consonant clusters in standard Swedish. *Phonum 7*. Umeå: Umeå University.

Bruce, G., 1986. Elliptical phonology. *Papers from the Scandinavian Conference on Linguistics*, 86-95.

Finke, M. & A. Waibel, 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. *Proceedings of Eurospeech*, 2379-2382.

Fosler-Lussier, E. & N. Morgan, 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication 29(2-4)*, 137-158.

Gårding, E., 1974. Sandhiregler för svenska konsonanter (Sandhi rules for Swedish consonants). *Svenskans beskrivning 8*, 97-106.

Jande, P.A., 2003. Phonological reduction in Swedish. *Proceedings of ICPhS*, 2557-2560.

Jande, P.A., 2005. Inducing decision tree pronunciation variation models from annotated speech data. *Proceedings of Interspeech*, 1945-1948.

Jande, P.A., 2006. Integrating linguistic information from multiple sources in lexicon development and spoken language annotation. *Proceedings of LREC workshop on merging and layering linguistic information* (accepted).

Jurafsky, D., A. Bell, M. Gregory & W. Raymond, 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 229-254.

Van Bael, C.P.J., H. van den Heuvel & H. Strik, 2004. Investigating speech style specific pronunciation variation in large spoken language corpora. *Proceedings of ICSLP*, 586-589.