# Inducing Decision Tree Pronunciation Variation Models from Annotated Speech Data

*Per-Anders Jande*

Department of Speech, Music and Hearing
KTH, Stockholm, Sweden
jande@speech.kth.se

## Abstract

A model of pronunciation of words in discourse context has been induced from the annotation of a spoken language corpus. The information included in the annotation is a set of variables hypothesised to be important for the pronunciation of words in discourse context. The annotation is connected to segmentally defined units on tiers corresponding to linguistically relevant units: the discourse, the utterance, the phrase, the word, the syllable and the phoneme. The model is represented as a tree structure, making it transparent for analysis and easy to use in a speech synthesis system. Using phonemic canonical pronunciation representations to estimate the segmental string of the annotated data gives a 22.1% phone error rate. Decision tree pronunciation variation models generated in a tenfold cross validation procedure showed an average phone error rate of 9.9%. Using multiple context variables for modelling pronunciation variation could thus reduce the error rate by 55%, compared to a baseline using canonical pronunciation representations.

## 1. Introduction

The pronunciation of a word depends on the context in which the word is uttered. A model of pronunciation variation due to discourse context is interesting in a description of a language variety. Such a model can also be used to increase the naturalness of synthetic speech and to dynamically adapt synthetic speech to different areas of use and to different speaking styles.

As a first stage in a project aimed at creating a general model of pronunciation variation due to discourse context, models of pronunciation variation within a specific speaking style have been created using data-driven methods. This paper presents the methods used for model induction from data, properties of the model and initial model evaluation results.

The second stage of the project, currently in progress, is to include speaking style charachteristics as variables in the model. For this purpose, speech data from several different communicative situations (representing different speaking styles) are being annotated.

## 2. Background

Work on pronunciation variation in Swedish on the phonological level has been reported by several authors. Gårding [1] presents a rule system for transforming canonical phonemic representations of consonant clusters at word boundaries into representations corresponding to a pronunciation at fast speech rates and Bannert and Czigler [2] give an account of variations in consonant clusters using a corpus of recorded speech. Among other things, the authors report the frequencies of the different types of elision and assimilation processes they found in the corpus. Bruce [3] discusses omissions of vowels and syllables in everyday speech pronunciation as compared to canonical pronunciation.

Jande [4, 5] reports a test aimed at investigating whether the perceived naturalness of synthetic speech can be increased using phone-level pronunciation modelling. A tentative phonological rule system for transforming canonical phonemic representations of words into representations corresponding to a fast speech rate was evaluated using speech synthesis. A set of short sentences differing only in one word were used to generate stimuli. During the test, the subjects were to select the most natural sounding of a canonically pronounced synthesised sentence and a version of the same sentence phonologically adapted to a fast speech rate. Each variant pair was presented at several synthesis speech rates.

The results showed significant preference biases in favour of the reduced forms for speech rates higher than the synthesis default rate and a significant increase in the preference bias for the reduced forms with increasing speech rate. However, the different target words behaved differently. An examination of the frequency of occurrence of the target words in a large newspaper corpus revealed an interesting pattern. Sentences with high frequency target words were more prone to be judged more natural in their reduced forms, irrespective of the speech rate. Low frequency words were more prone to be judged more natural in their canonical forms, irrespective of the speech rate. Since many studies have shown word predictability (often estimated with global word frequency) to be an important variable for predicting the pronunciation of a word in context [6, 7], this result was expected. The results thus confirmed the notion that an adequate model of pronunciation variation due to discourse context must include more context variables than phonetic context.

## 3. Model Specification

Phonemes in a canonical phonemic pronunciation representation are the central units in the model. The canonical pronunciation of a word is collected from a pronunciation lexicon. A vector containing all available context information is connected to each canonical phoneme unit. The task of the model is to make a decision about the appropriate phone realisation given the context associated with each canonical phoneme. The model thus describes segment level phonetic variation only, i.e. the model only describes processes affecting the presence or absence of entire speech segments and processes affecting the phonetic identities of segments (deletion, insertion and substitution processes on the phonological level).

The model is represented as a tree structure and uses input which can be obtained in a speech synthesis context. The tree structure also makes the model transparent for analysis.

# 4. Method

The general method for creating the pronunciation variation model is the data-driven paradigm. Speech data is annotated with variables hypothesised to be important for the pronunciation of words in discourse context and the annotation is used for creating models using decision tree induction.

## 4.1. Speech Data

The language variety modelled in the current effort is central standard Swedish. The speech data used was not recorded specifically for this project, but collected from various sources. The speech corpus includes data recorded or made available for research within the fields of phonetics, phonology and speech technology in different earlier research projects.

The speech data used in the first stage of the project – the creation of a speaking style specific pronunciation variation model – is a corpus of elicited monologue, the VAKOS corpus [2]. The corpus was originally recorded and annotated for the study of variation in consonant clusters in central standard Swedish. It consists of ∼103 minutes of speech from ten native speakers of central Standard Swedish.

## 4.2. Annotation Structure

All annotation is connected to some duration-based linguistic unit at one of six hierarchically ordered tiers. The tiers correspond to 1) the discourse, 2) the utterance, 3) the phrase, 4) the word, 5) the syllable and 6) the phoneme. Each tier is strictly sequentially segmented into its respective type of units. Some non-word units can be introduced at the word tier to ensure all parts of the speech signal belongs to some unit at all levels of annotation. Non-word units can be e.g. <*pause*>, <*inhalation*> or <*cough*>. These units are all part of some phrase, utterance and discourse. No annotation is connected to the non-word units on the syllable and phoneme tiers.

A boundary on a higher tier is always also a boundary on a lower tier. An utterance boundary is thus also always a phrase boundary, a word boundary, a syllable boundary and a phoneme boundary. Information can thus be unambiguously inherited from units on higher tiers to units on the tiers below. A unit can pass on its information to all the units within its boundaries on the tiers below. Information connected to syllable, word, phrase, utterance and discourse tier units, as well as to the phoneme tier units, is thus accessible from the phoneme tier. This is important since the pronunciation variation model uses phoneme-sized units as input and the information from all tiers thus must be connected to the phoneme unit at model induction.

Sequential context information, i.e., properties of the units adjacent to the current unit at the respective tiers is also used at model induction. Having the information stored at different tiers enables easy access to the sequential context information.

## 4.3. Segmentation

Each annotation tier is segmented into its corresponding units, beginning at the word tier. Based on the word tier segmentation and information derived from the word units, the tiers below and above the word tier are segmented. The phoneme tier is segmented word-by-word using the orthographic annotation, a canonical pronunciation lexicon and an HMM phoneme aligner [8]. The phonemes are clustered into syllables with forced syllable boundaries at word boundaries and the syllable tier is segmented using this clustering and the durational boundaries from the phoneme segmentation. Utterance boundaries are located manually with support from the word tier segmentation. For monologues, the discourse and utterance units coincide, i.e., the entire discourse is considered to be a single utterance. The phrase tier is segmented utterance-by-utterance using the output of a part of speech tagger [9, 10] and a parser [11, 12, 10].

## 4.4. Information Included in the Annotation

The discourse level annotation includes variables defining speaking style characteristics and some different measures of global speech rate. The utterance tier annotation includes the variable *speaker sex* and a number of measures of the mean *speech rate* over the utterance unit. There is also an *utterance type* variable only relevant for dialogue data. The phrase tier annotation includes the variables *phrase type*, *phrase length* (word, syllable and phoneme counts), *prosodic weight* (stress count, focal stress count), and measures of local *speech rate* over the phrase unit and of *pitch dynamism* and *pitch range*.

The variables included in the word tier annotation are *word length* (syllable and phoneme counts), *part of speech*, *morphology* (number, definiteness, case, pronoun form, tense/aspect, mood, voice and degree), *word type* (content word or function word), *word repetitions* (full-form and lexeme), *word predictability* (estimation based on trigram, bigram and unigram statistics from an orthographically transcribed spoken language corpus), *global word probability* (unigram probability), *the position of the word in the phrase*, *focal stress*, *distance to preceding and succeeding foci* (in number of words), *pause context*, *filled pause context*, *interrupted word context* and *prosodic boundary context* and different measures of *speech rate* over the word unit and of *pitch dynamism* and *pitch range*.

The syllable tier annotation includes the variables *stress*, *accent*, *distance to preceding and succeeding stressed syllable* (in number of syllables), *syllable length* (phoneme count), *syllable nucleus*, *the position of the syllable in the word* and measures of *speech rate* over the syllable unit. On the phoneme level, the annotation provided includes the *canonical phoneme* and a set of *articulatory features* describing the canonical phoneme, *the position of the phoneme in the syllable* and *in a consonant cluster*, *consonant cluster length* (phoneme count) and – used as the key at model induction – the realised *phone*. The phone alphabet includes the same symbols as the phoneme alphabet and an additional place filler `null` symbol for phonemes without any realisation in the speech signal (signalling a phonological deletion process).

## 4.5. Annotation Methods

Automatic methods (with some minor exceptions) are used for annotation. For the VAKOS corpus [2], manual word level segmentation and orthographic transcripts were supplied. For speech data where this information is not available, an automatic speech recogniser and an alignment system [8] are utilised. Manual correction of the orthographic string and the segmentation is relatively fast and improves segmentation at all tiers of annotation.

A hybrid system using statistical decoding and a set of correction rules is used for phonetic transcription [13]. The performance of the autotranscription system was evaluated using a manually transcribed subset of the corpus as an evaluation gold

standard. This evaluation showed the system to have a phone error rate of 14.37%. The impact of the noise in the transcription key on the final model is presently unknown. Manual correction of phonetic transcripts will be done for a larger subset of the speech data to evaluate the impact of the phone key errors on the final model performance. A decision on whether it is worth to manually correct the reminder of the speech data will be based on this evaluation.

Speech rate is estimated by inverse segment duration. Segments are estimated by the canonical phonemes and segment boundaries by the automatically obtained alignment of the phoneme string to the signal. Speech rate estimates based on all segments and estimates based on vowel segments only were calculated. Duration normalised for inherent phoneme length and for speaker, respectively, is used as well as non-normalised duration. Both duration on a linear scale and on a logarithmic scale is used. All combinations of strategies are included in the annotation, resulting in 16 different speech rate measures for each unit.

### 4.6. Model Induction

Since the current models are specific to elicited monologue and the discourse and utterance tiers coincide for monologues, only the utterance tier speech rate estimates were used for model induction. Also, using a single speaking style meant that the speaking style oriented discourse tier variables were the same for all discourses and these variables were therefore also not used. Thus, no discourse tier variables were available as context for the current models.

The freely available DTREE program suite [14] was used for decision tree induction. An evaluation of available attribute selection measures and optimisation options revealed that using symmetric information gain [15] as the measure for selecting attributes and allowing the decision tree inducer to form subsets of symbolic attributes gave optimal performance.

A tenfold cross validation procedure was used for model evaluation. The training data was thus divided into ten equally sized partitions using random sampling. Ten different decision trees were induced, each with one of the partitions left out from training. The left out partition was then used for evaluation. Each model in the cross validation setting was trained on 52,929 samples (each sample corresponding to a canonical phoneme in the annotation of the speech data) and evaluated on 5,881 samples. The annotation variables with sequential context (were applicable and relevant) gave rise to a 323 attribute vector for each training sample. Both trees with basic pruning performed at induction and trees with additional confidence level pruning were created. The optimal tree for each data set was selected to represent the set at tenfold cross validation.

## 5. Results and Discussion

The average phone error rate (PER) across the optimal trees was 9.9%. The tree with the lowest error rate had a PER of 9.0% and the tree with the highest error rate had a PER of 10.5%. A baseline was calculated using the canonical phonemes to estimate phone realisations, giving a PER of 22.1%. This means phone errors are reduced by 55% by the pronunciation variation model. The PER of each tree is shown in Figure 1.

The decisions to be made by the trees can be seen as transformations applied to phoneme strings. The transformation process can be to *keep* the phoneme symbol in the phone string, to *substitute* the phoneme symbol for the (correct) phone symbol or symbols or to *delete* the phoneme symbol. The training data contained 77.9% phoneme-key phone pairs corresponding to *keep* processes, 12.7% pairs corresponding to *substitute* processes and 9.4% pairs corresponding to *delete* processes. From the total of each kind of process, 94.9% of the *keep* processes, 74.7% of the *substitute* processes and 71.4% of the *delete* processes were correctly executed by the models.

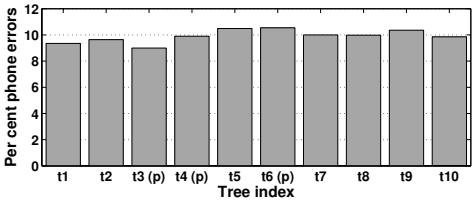


Figure 1: *Decision tree performance (p = pruned)*

The decision tree paradigm ensures that the types of errors that can be generated by a model for a certain phoneme is constricted to the variation in realisation present in the training data. The most critical type of error that can be made is probably erroneous phoneme deletion, especially erroneous vowel deletion. The results from the tenfold cross validation showed that 21.7% of the deletions performed by the models were erroneous (i.e., the processes described by the key were not *delete* processes). However, only 2.7% were erroneous vowel deletions.

### 5.1. Model Complexity

Trees with only basic pruning gave the best performance in seven cases out of ten. Both the tree with the lowest and the tree with the highest PER were pruned trees. Although pruning gave a decrease in performance given the evaluation data in seven cases out of ten, the decrease was only significant in two cases ($p<0.1$) using the McNemar test. However, the pruning gave rise to considerable reduction of model complexity.

For example, one tree used 220 attributes and had 3177 nodes and 26 levels prior to confidence interval pruning. After pruning, the tree used 56 attributes and had 385 nodes on 8 levels. The model performance decreased from a PER of 10.0% to a PER of 11.3% (an increase in phone errors with 13.7%). This example is the least complex tree after pruning and also the tree with the largest decrease in performance caused by pruning. The tree with the best performance used 227 attributes and had 3095 nodes and 24 levels prior to confidence interval pruning. After pruning, the tree used 70 attributes and had 437 nodes on 9 levels. In this case, the pruning *decreased* the PER from 9.9% to 9.0% (a 9.6% phone error reduction). This was the only case where the decrease in PER caused by pruning was significant ($p<0.01$).

There can be considerable merits with a less complex model in a synthesis system, since less input variables will have to be supplied to the model. Whether the gain in model simplicity is worth the possible reduction in model performance is a question that will have to be considered at integration of the pronunciation variation model into a synthesis system.

### 5.2. Annotation Used by the Models

The first attribute used to split the data set at decision tree induction was the canonical phoneme identity in all cases. This is not surprising, since the baseline results show that 78.9% of the phonemes in the canonical representation should be realised by the phoneme's canonical realisation when modelling elicited

monologues. For spontaneous dialogues, this rate is probably considerably lower, but it is unlikely that the canonical phoneme will not be the main predictor for phoneme realisation in any speaking style. The right phonemic context was the most frequently used second branching attribute.

None of the utterance tier attributes were used in any of the pruned models. This was not surprising, since speaker sex was not expected to be critical and the mean segment duration over the utterance was very similar across speakers. From the phrase, word, syllable and phoneme tiers, many different types of attributes were used.

In the tree with the best performance, the phrase tier variables *phrase type*, *phrase length* (segment count), *phrase pitch range* in Hz (for the current and the succeeding phrase) and *phrase pitch dynamic* measures (in Hz and Mel, for the current and the succeeding phrase) were used. Further, a variety of *speech rate* measures over the phrase unit were used. Most combinations of strategies for calculating the speech rate (see section 4.5) were included.

From the word tier annotation, *word length* (segment count, for the current and the succeeding word), number of *lexeme repetitions*, *part of speech* (and part of speech sequential context), *word probability*, *global word frequency* (for the current and the succeeding word), right *hesitation sound context*, a variety of *speech rate* measures over the word unit, *pause context* and *adjacent pause length* were the variables used in best model.

Used syllable tier variables were *stress* and *accent* (for the current and the succeeding syllable), *syllable length* (segment count), *syllable nucleus* and several *speech rate* measures over the syllable unit. The *canonical phoneme* and its sequential context and the *position of the phoneme in the syllable* (onset, nucleus, coda) were phoneme tier attributes used in the model. Also, the phoneme feature attributes *intrinsic length*, *place of articulation/vowel height* and *voice/rounding* were used.

### 5.3. Further Evaluation

Using the decision trees to predict the segmental realisation of phonemes in an actual database means that each decision made by a tree can be either correct or incorrect. In a speech synthesis setting, however, several alternative pronunciations can be equally natural sounding. Listening experiments with speech synthesised using pronunciation variation modelling are planned to evaluate the type of model described in this paper from a perceived naturalness point of view.

When used in a synthesis setting, a pronunciation variation model is applied in between two successive passes through a diphone synthesiser front-end. The segment-level realisation depends on the prosodic model of the synthesiser front-end. The prosodic input to the tree is calculated from parameters generated by the prosodic model on the basis of a canonical pronunciation representation. Prosodic input based on speech data and input based on model-generated values are comparable, since the speech rate measures are based on canonical phoneme string and only mean speech rate measures over units larger than the phoneme are used. Pitch-based measures are calculated from information about pitch minima and pitch maxima only.

## 6. Conclusions

Models of pronunciation variation have been created using data-driven methods. Variables hypothesised to be important for the pronunciation of words in discourse context were automatically annotated for spoken language corpora and decision trees were induced from the annotation. The task of the trees was to make decisions about the phonetic realisation of phonemes given a set of context variables. The trees gave an average phone error rate of 9.9% when evaluated on the type of data on which they were trained. This meant an error reduction of 55% compared to a baseline using canonical pronunciation representations.

## 7. Acknowledgements

The research reported in this paper was carried out at the Centre for Speech Technology (CTT), a competence centre at KTH, supported by VINNOVA (the Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. The work was conducted within the frame of a graduate programme supported by the Swedish Graduate School of Language Technology, GSLT. Many thanks to all the people who have contributed with tools and resources used for the work reported in this paper.

## 8. References

[1] E. Gårding, "Sandhiregler för svenska konsonanter (sandhi rules for Swedish consonants)," in *Svenskans beskrivning 8*, 1974, pp. 97–106.

[2] R. Bannert and P. Czigler, *Variations in consonant clusters in standard Swedish*, ser. Phonum 7, Reports in Phonetics. Umeå: Umeå University, 1999.

[3] G. Bruce, "Elliptical phonology," in *Papers from the Scandinavian Conference on Linguistics*, 1986, pp. 86–95.

[4] P.-A. Jande, "Evaluating rules for phonological reduction in Swedish," in *Proc Fonetik*, 2003, pp. 149–152.

[5] ——, "Phonological reduction in Swedish," in *Proc ICPhS*, 2003, pp. 2557–2560.

[6] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, no. 2–4, pp. 137–158, 1999.

[7] D. Jurafsky, A. Bell, M. Gregory, and W. Raymond, "The effect of language model probability on pronunciation reduction," in *Proc ICASSP*, 2001, pp. 2118–2121.

[8] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proc Fonetik*, 2003, pp. 93–96.

[9] T. Brants, "TnT – a statistical part-of-speech tagger," in *Proc Applied Natural Language Processing Conference (ANLP)*, 2000.

[10] B. Megyesi, "Data-driven syntactic analysis – methods and applications for swedish," Ph.D. dissertation, KTH, Stockholm, 2002.

[11] J. Aycock, "Compiling little languages in python," in *Proc International Python Conference*, 1998.

[12] B. Megyesi, "Shallow parsing with pos taggers and linguistic features," *Journal of Machine Learning Research*, vol. 2, pp. 639–668, 2002.

[13] P.-A. Jande, "Annotating speech data for pronunciation variation modelling," in *Proc Fonetik*, 2005, pp. 25–28.

[14] C. Borgelt, "A decision tree plug-in for dataengine," in *Proc European Congress on Intelligent Techniques and Soft Computing (EUFIT)*, vol. 2, 1998, pp. 1299–1303.

[15] R. Lopez de Mantaras, "A distance-based attribute selection measure for decision tree induction," *Machine Learning*, vol. 6, no. 1, pp. 81–92, 1991.