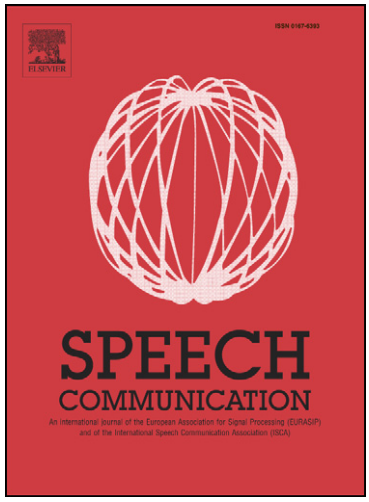
Spoken Language Annotation and Data-Driven Modelling of Phone-Level Pronunciation in Discourse Context

Per-Anders Jande



Spoken Language Annotation and Data-Driven Modelling of Phone-Level Pronunciation in Discourse Context

Per-Anders Jande

PII: S0167-6393(07)00142-2
DOI: 10.1016/j.specom.2007.07.004
Reference: SPECOM 1659

To appear in: *Speech Communication*

Received Date: 15 May 2006
Revised Date: 24 July 2007
Accepted Date: 25 July 2007

Please cite this article as: Jande, P-A., Spoken Language Annotation and Data-Driven Modelling of Phone-Level Pronunciation in Discourse Context, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.07.004

Spoken Language Annotation and Data-Driven Modelling of Phone-Level Pronunciation in Discourse Context

Per-Anders Jande *

Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

Abstract

A detailed description of the discourse context of a word can be used for predicting word pronunciation in discourse context and also enables studies of the interplay between various types of information on e.g. phone-level pronunciation. The work presented in this paper is aimed at modelling systematic variation in the phone-level realisation of words inherent to a language variety. A data-driven approach based on access to detailed discourse context descriptions is used. The discourse context descriptions are constructed through annotation of spoken language with a large variety of linguistic and related variables in multiple layers. Decision tree pronunciation models are induced from the annotation. The effects of using different types and different amounts of information for model induction are explored. Models generated in a tenfold cross validation experiment produce on average 8.2% errors on the phone level when they are trained on all available information. Models trained on phoneme level information only have an average phone error rate of 14.2%. This means that including information above the phoneme level in the context description can improve model performance by 42.2%.

Key words: Spoken language annotation, Pronunciation variation, Pronunciation modelling, Decision trees

1 Introduction

There is considerable variation in the spoken language performance of an individual speaker depending on the speaking situation. For example, word pronunciation depends heavily on speaking style (cf. e.g. Ostendorf et al., 1996; Van Bael et al., 2004) and speech rate (cf. e.g. Fosler-Lussier and Morgan, 1999; Zheng et al., 2000). The pronunciation of a certain word also depends on its local context, such as adjacent phonemes and the predictability of the word in its context (cf. e.g. Finke and Waibel, 1997; Duez, 1998; Fosler-Lussier and Morgan, 1999; Jurafsky et al., 2001). The variation in pronunciation is manifested on many levels. There is variation in prosodic features, such as speech rate, intonation, rhythm and accent. There is also variation in the phone-level realisation of words and in the fine-phonetic realisation of speech segments.

Although there is a certain degree of individual and random variation in the pronunciation of words in context, the variation due to context factors is largely systematic within a restricted, relatively homogeneous group of language users. This agreement on systematic variation strategies can be seen as a property of the language variety (e.g. dialect) spoken by the group.

* Tel: +46 8 79 9269; fax: +46 8 790 7854
Email address: per.anders.jande@gmail.com (Per-Anders Jande).

Modelling pronunciation variation in discourse context is interesting for the description of a language variety and can come to practical use e.g. for increasing the naturalness of synthetic speech. The influence of a large variety of variables on the pronunciation of words has been studied. However, the variables have mostly been studied in isolation or in connection with only a few other variables. A detailed description of a discourse, including a large variety of linguistic and related variables, would enable studies of the interplay between various information sources on e.g. phone-level pronunciation and allow data-driven creation of models for prediction of word pronunciation in context, e.g. for a speech synthesis application.

A strategy used for modelling phone-level pronunciation variation for speech synthesis is to model the pronunciation of a single speaker, typically the speaker whose voice is used in the concatenation database (cf. e.g. Miller, 1998; Bennett and Black, 2005). When the specific purpose is to get the most natural sounding speech synthesis, this is a good strategy.

However, the model created is adapted specifically to the target speaker and thus not general for any group of speakers. If the aim is to describe a *language variety* from a pronunciation variation point of view, it is necessary to study the behaviour of many speakers of the particular language variety. Statistics can then be used to single out common patterns from individual patterns. Such an approach has been employed by e.g. Werner et al. (2004a,b). They use a stochastic pronunciation net induced from a speech corpus including many speakers (thus being a more general pronunciation model) and a word duration model. The system first determines adequate word durations using the probability of each word in its context and then estimates the appropriate phone sequence given the specified durations, the transition probabilities from the word pronunciation nets and word transition probabilities.

A number of studies on pronunciation variation in Swedish on the phonological level have been carried out. For example, Gårding (1974) presents an empirically based rule system for transforming a phonemic representations describing a maximally detailed pronunciation into a representation corresponding to fast speech, focusing on consonant clusters at word boundaries. Bannert and Czigler (1999) report a study of variation in consonant clusters using a corpus of mainly spontaneous speech. Bruce (1986) discusses omissions of vowels and syllables in everyday speech pronunciation as compared to canonical pronunciation. According to Bruce (1986), omission phenomena are governed primarily by the syllable-bound rhythmical organisation of spoken language.

Inspired by the results from the above mentioned studies on pronunciation variation in Swedish, Jande (2003) constructed a tentative rule system for transforming canonical phonemic representations of words into representations corresponding to a fast speech rate. The rule system was used to create synthetic speech stimuli used in an assessment experiment. The results showed a significant increase in the preference bias for the reduced forms with increasing speech rate. The results also showed that the frequencies of the target words used in the synthetic stimuli affected the results: stimuli with high frequency targets were always preferred by a majority of the subjects in their reduced form (irrespective of speech rate), while stimuli with low frequency targets were always preferred in their canonical form. This trend concurred with earlier research showing that word predictability is a strong predictor of word pronunciation.

The aim of the project described in this paper is to model systematic discourse context-induced variation in phone-level pronunciation inherent in a language variety. The methods used for pronunciation modelling are data-driven. Spoken language is annotated with various kinds of linguistic and related information and machine learning is used to create pronunciation models from the annotation. The phoneme is the central unit in the approach employed and the annotation is aimed at describing the discourse context of a phoneme from high-level linguistic variables such as speaking style, down to the articulatory feature level. The work presented in this paper is the continuation of the work described in Jande (2005). More information has been added to the information and more data has been annotated. Models resulting from the annotation have been studied in more detail. The full analysis of the models can be found in Jande (2006b).

Models are created and evaluated using a cross-validation procedure, resulting in a measure of phone error rate (PER). The PER of the optimal set of models is compared to three baselines, 1) a phonemic transcript obtained through concatenating phonemic pronunciation representations from a lexicon, 2) the phonemic transcript processed through a sandhi rule system optimised for the data set and 3) the output of a pronunciation model trained

without access to prosodic information.

The pronunciation modelling project has two important specific goals. The first goal is to create models able to predict phone-level pronunciation in context with high accuracy. The second goal is to create models that can serve as linguistic descriptions of pronunciation variation. This second goal requires a method making it possible to create transparent models, revealing which variables are the most important for predicting pronunciation in context and how variables co-operate to make predictions.

This paper will describe the methods used for annotating spoken language with linguistic context information in multiple layers and the creation and evaluation of models induced from the annotation. The language variety annotated and modelled is central standard Swedish, the standard variety of Swedish spoken in the Stockholm area. The methods can, however, easily be adapted for modelling other language varieties and languages.

2 Annotation Method

A requirement of the data-driven approach taken to pronunciation modelling is, of course, data. In the current approach, the data consists of the annotation of spoken language, where the annotation is aimed at describing the discourse context of a phoneme from high-level linguistic variables such as speaking style, down to the articulatory feature level. It is important to have data that is accurate and also to have a sufficient amount of data. Mainly automatic methods are used for annotation, to make annotation fast in comparison to manual annotation and thus making it practically possible to obtain a sufficient amount of data. The price of using automatic methods is that the result may not always be as accurate as the result of manual annotation would have been. This section describes the speech data used for pronunciation modelling and the system and methods used for annotating the speech data.

The system used for annotation and the information provided is tailored for the current project. There are, of course, many other projects involving spontaneous speech data annotation for different purposes. For example, the Linguistic Data Consortium, LDC, has created annotated resources for evaluating metadata extraction systems. The LDC annotation project bears some resemblance to the annotation efforts described in this paper, although the purpose (and thus the nature of the annotation and methods used for annotation) is different. A specification of the LDC metadata annotation can be found in Strassel (2004).

2.1 Speech Data

The speech data used for pronunciation modelling consists of three speech databases: the VAKOS database, a RADIO INTERVIEW database and a RADIO NEWS database. The VAKOS database was originally constructed by Bannert and Czigler (1999) for a phonological study of variation in consonant clusters. The RADIO INTERVIEW database and the RADIO NEWS database consist of recordings originating from *Sveriges radio* (Swedish public service radio) and have previously used in the GROG project, aimed at modelling the structuring of speech in terms of prosodic boundaries and groupings, cf. Carlson et al. (2002). All speech data are digital studio recordings sampled at 16 kHz.

The VAKOS database is a set of elicited monologues; ten speakers talk about some suggested topic or topics to a recording assistant (who is silent). About ten minutes from each speaker is included in the database. The VAKOS database includes some manual annotation at different levels. The parts of the annotation re-used for the purpose of pronunciation modelling are the orthographic transcripts, the word-level segmentation, prosodic boundary annotation, focal accent annotation, and annotation of word fragments (interrupted words) and filled pauses.

The RADIO INTERVIEW database is a set of two 25-minute radio broadcast interviews, each including speech

mainly from three speakers, the interviewee and two interviewers. The interviewees are experienced public speakers (politicians) and are allowed to answer questions in length, rarely being interrupted. The RADIO NEWS database includes two radio news broadcasts, including speech from altogether three studio news announcers and eight reporters. Only studio environment recordings are included in the RADIO NEWS database. The radio broadcast databases include orthographic transcripts and manual annotation of prosodic boundaries originating from the GROG project. For one of the interviews, focally stressed words were also annotated in the GROG project. This information is re-used in the annotation for pronunciation modelling.

2.2 A Multi-Layer Annotation System

The annotation used for pronunciation modelling 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. The layers are segmented into units, which are linguistically meaningful and can be synchronised to the speech signal. The segmentation of each layer is strictly sequential, i.e., every part of the signal belongs to some unit at all layers and there is no overlap between units within a layer.

Durational boundaries are inherited from higher order layers to lower order layers, so that a discourse boundary is always also an utterance boundary, a phrase boundary, a word boundary, a syllable boundary and a phoneme boundary. The layers are thus hierarchically ordered so that a higher order unit serves as the parent of all lower order units within its segmental bounds. An arbitrary amount of information can be supplied for each unit in each layer. Figure 1 shows an excerpt of a sound file with some aligned example annotation.

The most important feature of this system of annotation is that information can be unambiguously inherited from units on higher layers by units on the layers below. Consequently, information connected to syllable, word, phrase, utterance and discourse layer units, respectively, as well as to the phoneme layer units, is accessible from the phoneme layer. This is important since the pronunciation models will use phoneme-sized units as input. Sequential context information, i.e., properties of the units adjacent to the current unit at the respective layers, is used at model induction together with information connected to the current units. Having the information stored in different layers enables easy access to the sequential context information.

Figure 1 HERE.

2.3 Segmentation

The annotation process begins by a segmentation of each annotation layer into its respective type of unit. The next step is to retrieve, calculate or estimate a set of features for each unit. With some minor exceptions, automatic methods are used for segmentation, however with manual supervision to improve accuracy at some intermediate stages.

In the current context, an utterance is defined as a discourse turn uttered by a single speaker. This means that a monologue discourse is treated as a single utterance. For dialogues, the corpus is manually segmented into utterances. During utterance segmentation, pauses between utterances are included in the utterance to the right. Overlapping speech *between utterances* is given the special utterance unit tag *<overlap>*, but no other information is associated with the unit.

An *<overlap>* utterance unit is extended to the nearest word boundary, so that a partially overlapped word is included in the *<overlap>* utterance unit in its full. Overlapping speech *within an utterance* (i.e., where the utterance has started before the overlap and continues through and after the overlap) is not annotated on the utterance layer

(it is on the word layer, however). The speech segments annotated as overlapping on the utterance layer are given <*overlap*> tags also on the word layer and a <*junk*> tag on the phoneme layer, but otherwise no information is included for lower order layers.

Automatic segmentation begins at the word level. Given an orthographic string, the corpus is segmented into word units using an automatic aligner (Sjölander, 2003; Sjölander and Heldner, 2004). Automatic speech recognition can be used to facilitate orthographic transcription. However, for the currently used databases, the orthographic string, including annotation of filled pauses and non-speech sounds, has been manually supplied. Special consideration has been taken to supplying an accurate word sequence, since the automatic alignment is highly dependent on the orthographic string. Further, manual correction of the word layer segmentation is performed, since all succeeding annotation depends on this segmentation. Manual supervision at this level is relatively fast and increases in the word layer segmentation accuracy give large improvements in the accuracy of successive annotation. Manual word layer segmentation followed the VAKOS database and one of the radio interviews.

The phrase layer is segmented using the SPARK parser (Aycock, 1998) with a context-free grammar for Swedish constructed by Megyesi (2002b,a) operating on a string of tags produced by the TNT part-of-speech and morphological tagger (Brants, 2000) trained by Megyesi (2002a). Only phrase chunk information is used and the phrases are aligned to the signal using the word boundaries. The parser was created for parsing written text, but it is robust and produces parses also for tagged orthographic transcripts of spoken language.

During phrase layer segmentation, only maximal phrases are considered. A noun phrase can include modifiers of different types, e.g. nouns, adjective phrases and prepositional phrases. The entire maximal projection of the noun phrase is counted as a single phrase and the identity and boundaries of any constituent phrases are ignored. Similarly, conjoined adjective phrases are counted as a single adjective phrase.

Some word units do not belong to any phrase chunk (mostly conjunctions). For phrase segmentation purposes, these words are given a *no phrase* tag and are treated as one-word phrases. Verb phrases are not included in the analysis. Verbs are instead parts of either a *verb cluster* or an *infinitive phrase*. A *verb cluster* is a single verb or a continuous sequence of verbs belonging to the same verb phrase (e.g. *'would have been'*) and an *infinitive phrase* is an infinite verb proceeded by an infinite particle. The infinitive phrase may contain adverb phrases and/or verb particles, e.g. *'to go out'*. The full set of phrase types produced by the parser can be seen in table A.1.

The phoneme layer is segmented word-by-word using the word boundaries and canonical phonemic representations as input to the automatic aligner. The phonemic representations are collected from the CENTLEX pronunciation lexicon (Jande, 2006a), if the word occurs in the lexicon. Words not occurring in the lexicon receive phonemic representations generated by a grapheme-to-phoneme conversion algorithm included in the RULSYS text-to-speech system (Carlson and Granström, 1976; Carlson et al., 1982). The speech databases contain some instances of interrupted words (i.e., parts of words). In the cases where these are not correctly handled by the grapheme-to-phoneme rules, the phoneme representations are corrected manually for consistency.

On the phoneme layer, the synchronisation of units to the signal is more abstract than on the higher order layers; not all phonemes in the canonical phonemic representations have overt correspondents in the speech signal, but nevertheless will have a duration in the annotation. As will be seen in Section 2.4, the abstract nature of the phoneme boundaries is exploited in phoneme duration-based measures.

For syllable boundary allocation, lists of phonotactically allowed onset and coda consonant sequences based on Sigurd (1965) and (Elert, 1970, pp. 89–90) are used to exclude impossible syllable boundaries. When it is allowed to place the syllable boundary at more than one location in a consonant sequence between two vowels, the coda of the first syllable is maximised if the vowel is a short stressed vowel, and the onset of the second syllable is maximised otherwise. Further, syllable boundaries are forced at word boundaries and at compound constituent boundaries (compound boundaries are included in the phonemic representations collected from the pronunciation lexicon). The syllable boundaries are synchronised to the signal using the phoneme boundaries.

Some units with special characteristics are introduced at the word layer to ensure that parts of the signal that are not

speech (or non-analysable speech) can be annotated. The special unit types are <*overlap*> (overlapping speech), <*pause*> (including pauses, inhalation and exhalation sounds), <*non_speech*> (including laughter, smacks, clicks, coughs and hawking sounds etc.) and <*filled pause*> (e.g. 'hesitation' sounds resembling /ə/, /eː/, /əm/, /ɜːm/ or /m/). The information supplied for normal word units is *not* included for these units. Within the boundaries of one of the special word layer units, a <*sil*> or a <*junk*> special phoneme unit is used and no additional annotation is supplied on the phoneme and syllable layers. The <sil> tag is used for pauses, including inhalations and exhalations, and the <junk> tag is used for parts of the signal that contain vocally produced, but phonetically unanalysable sound, e.g. overlapping speech, smacks, coughs and laughter.

2.4 Mean Phoneme Duration Measures

As previously discussed, speech rate is an important factor for the phonetic realisation of words. Speech rate can be defined as the number of phonemes per time unit, which is the inversion of mean phoneme duration. In the current speech annotation, several measures of mean phoneme duration are calculated, including measures of mean z-normalised phoneme duration. When normalised, duration values can be zero. Hence, converting the mean normalised phoneme duration to a speech rate measure may in these cases give rise to infinite speech rates being artifacts of the normalisation process. For this reason, the mean phoneme duration measures are used in the annotation rather than speech rate measures. *Mean phoneme duration* is measured globally, over the entire discourse, and locally over each utterance, phrase, word and syllable.

The mean phoneme duration measures are based on the automatic segmentation of the phoneme layer, conducted through automatic alignment of canonical phonemic representations of words to the speech signal. The *mean phoneme duration* is an abstract measure and coincides with the concrete measure *mean phone duration* when all phonemes in the phonemic representation are realised. The measure thus constitutes an estimate of what what the mean phone duration would be if all phonemes in the canonical pronunciation representation were realised over a unit of fixed duration.

Mean phone duration cannot be used for prediction, since the phone string is the variable to be predicted by the pronunciation model. The exact number of phones is thus not known in advance when the model is used. The abstract nature of the mean phoneme duration measure is likely to make it a strong predictor of phone level pronunciation; high speech rate is generally a good predictor of phonological assimilation and reduction processes and mean phoneme duration emphasises sections of the speech signal with high speech rates more than measures corresponding to words or phones per time unit. For the mean phoneme duration measure to be usable in the absence of a speech signal, a prosodic model estimating the durations of syllables (and hence, of units on higher order layers) is necessary.

Different phonemes have different inherent lengths and additionally, central standard Swedish has phonologically long and phonologically short vowels. Neither inherent length nor phonological length/complementary length have anything to do with speech rate. A one-syllable word with a phonologically long vowel may have a longer duration than a word with a phonologically short vowel. However, this does not reflect a difference in speech rate between the words. If mean phoneme duration is calculated over larger units such as the phrase or the utterance, differences due to inherent length and phonological length will to a large degree even out. However, when speech rate is calculated locally over words and syllables, they mostly will not. For this reason, measures based on normalised phoneme duration are included in the annotation alongside measures based on absolute phoneme duration. During normalisation, the duration of each phoneme token is related to the mean duration of the particular phoneme type using the normal transformation. Phonologically long phonemes (including consonants) are separated from phonologically short phonemes, and vowels serving as nuclei in stressed syllables are separated from their phonologically identical counterparts in unstressed syllables.

A variant of the mean phoneme duration measure included in the annotation is the *mean vowel duration*. For this measure, all segments except vowels are ignored under the assumption that perceived speech rate may be better modelled by vowel duration alone than by general segment duration. The mean phoneme duration measures and

the mean vowel duration measures are calculated both from duration on a linear scale and from duration on a logarithmic scale. Since small differences in speech rate probably have larger effects on phone-level pronunciation when the speech rates compared are high than when the speech rates are low, the relative size of small differences in duration is increased through transferring the phoneme durations to the logarithmic scale ($\log_e$).

To sum up, there are measures based on all phonemes and on vowels only; there are measures based on absolute duration and on normalised duration; and, finally, there are measures calculated on a linear time scale and on a logarithmic time scale. All combinations of variants are calculated, resulting in a total of eight *mean phoneme duration* measures.

2.5 Pitch Dynamics and Pitch Range Estimation

Pitch movement is correlated with emphasis; much pitch movement over a particular unit makes the unit stand out from its surrounding and signals that the unit is emphasised. Emphasis is also correlated with segmental pronunciation, in such a way that the pronunciation tends to be more similar to the canonical pronunciation for emphasised words than for non-emphasised words. This means that there is a correlation between pitch dynamics and phone-level pronunciation. The measures described below are included in the annotation to make use of this correlation.

The ESPS pitch extraction algorithm incorporated in the SNACK Sound Toolkit (Sjölander, 2004; Sjölander and Beskow, 2000) is used to extract the pitch contour from the speech data with a sampling frequency of 100 Hz. Using the extracted pitch contours, the *pitch range* and four measures of *pitch dynamics* ('liveliness') are calculated over each utterance, phrase and word unit.

Pitch range is defined as the difference between the largest pitch maximum and the smallest pitch minimum contained within a unit. The first and the last voiced sample of the unit over which the pitch based measures are measured are counted as extreme values. *Pitch dynamics* measures are based on the absolute distance of maximum and minimum points or plateaus from a base frequency. Two base frequencies are used: 1) the median pitch over the unit and 2) a base frequency estimating liveliness variation as perceived by human listeners (Traunmüller and Eriksson, 1995a; Jande, 2006b).

The absolute distances of maximum and minimum points or plateaus from the respective base frequencies are summed up over a unit, and based on these sums, two different pitch dynamics measures are calculated for each base frequency. First, the sums are divided by the number of minimum or maximum points or plateaus contained by the unit, to obtain a measure of pitch dynamics differentiating between units with pitch extremes with *large* average deviations from the base frequency and units with pitch extremes with *small* average deviations from the base frequency. Second, the sums are divided by the number of (non-zero) pitch samples contained within the unit, resulting in a measure differentiating between units with *fast* average pitch movement and units with *slow* average pitch movement.

Equal differences in pitch measured in Hz are not perceptually equivalent across different pitch levels. Hence, three scales constructed to mirror the response of the human auditory system (psychoacoustic scales) are used for measuring pitch in addition to the linear Hz frequency scale. The three psychoacoustic scales used are the MEL scale (Stevens and Volkman, 1940), the equivalent rectangular bandwidth (ERB) scale (More and Glasberg, 1983; Hermes and Gestel, 1991) and the semitone scale, shown to give the best results in terms of perceptual equivalence by e.g. Traunmüller and Eriksson (1995b) and Nolan (2003). This results in a total of four different variants of the *pitch range* measure and of each of the four *pitch dynamics* measures.

2.6 Word Predictability and Related Measures

The predictability of a word has been shown to be important for the realisation of the word (cf. e.g. Fosler-Lussier and Morgan, 1999; Jurafsky et al., 2001). Many variables influence the predictability of a word in context. Measures related to word predictability included in the annotation described here are *collocation frequency*, *word repetitions*, *lexeme repetitions*, *the position of the word in a phrase*, *part of speech*, *the position of the word in a frequent collocation* and *global word frequency*. A special measure termed *word predictability* is also included in the annotation.

The word predictability statistic is the word probability in trigram context with back-up smoothing using bigram and unigram probabilities. The trigram weight is set to 0.6, the bigram weight is set to 0.3 and the unigram weight is set to 0.1. Unigram, bigram and trigram probabilities were collected from a formatted version of the Göteborg Spoken Language Corpus, GSLC (Allwood et al., 2000). GSLC contains orthographic transcripts of spoken language from a variety of communicative situations. After formatting, excluding some types of non-word units and converting transcripts to standard orthography, the size of the corpus is approximately 1.3 million words. Probabilities are calculated utterance by utterance by introducing two utterance boundary symbols in between each two consecutive utterances before calculating trigram statistics and one utterance boundary symbol before calculating the bigram statistics. Simple full-form word probabilities were used for the unigram probability.

The estimated *global word probability* is sometimes used as a rough estimator of word predictability (e.g. in Fosler-Lussier and Morgan, 1999). Since an estimate of global word probability from GSLC is available (the unigram probability), it is included in the annotation. The position of a word in its phrase or in a collocation affects the predictability of the word, and the positions of a word in the phrase and in a collocation, respectively, are included in the annotation as three-way classifications: *initial*, *medial* or *final*, where *initial* is the default value used for one-word phrases. Collocations are in the current context defined as trigrams occurring at least four times in GSLC or bigrams occurring at least three times.

Two measures of the *number of word repetitions* are included in the annotation, the number repetitions of the full-form word thus far in the discourse and the number of repetitions of the lexeme thus far in the discourse. PCKIMMO, the SIL implementation of Koskeniemis's two-level morphology system (SIL International, 1995) with lexica and rules for Swedish compiled by Ridings (2002) is used for finding the lemma form of each word. The combination of the lemma form and the part of speech is used to define a lexeme.

2.7 Automatic Phonetic Transcription

Phonetic identity is the variable to be estimated by the pronunciation models and hence, the phonetic annotation is used as the key in model training. Manual phonetic annotation is a time-consuming and thus expensive task. A system for automatic phonetic transcription has been built to facilitate the current annotation. The automatic transcription system is a hybrid phonetic decoder using statistical decoding and a set of a posteriori correction rules. The task of the system is to supply the context-dependent realisation of each phoneme in the canonical pronunciation representation collected from a lexicon. The realisation can be ∅ (*'no realisation'*). The phone label set is the same as the phoneme label set and includes 23 vowel symbols and 23 consonant symbols (cf. Table 1). There is also a place filler ∅ label in the phone label set that occupies a phoneme position with no realisation in the phonetic string.

TABLE 1 HERE.

2.7.1 Statistical Decoding

Finite state transition networks representing the possible realisations of a word are built using an empirically compiled context-insensitive list of possible realisations for each phoneme (cf. Table 1). Statistical decoding is conducted in a word-by-word manner, forcing phoneme boundaries at the manually annotated word boundaries. The part of the speech signal corresponding to a specific word is sampled and parameterised to form a sequence of observations using the SNACK sound toolkit (Sjölander and Beskow, 2000). Viterbi decoding is used to find the path through the network with the highest probability of having produced the observation sequence and the corresponding phone sequence (aligned to the signal) is the output of the statistical decoder. In a post processing step, the phone string is aligned to the phoneme string using phoneme position indices and $\emptyset$ *'null'* place filler phones.

2.7.2 A Posteriori Correction Rules

The tentative phone string resulting from the statistical decoding process can be viewed as the result of a set of phonological transformation rules operating on the canonical phoneme string. A set of a posteriori rules inverting some of these phonological rules under certain conditions has been developed to correct some systematic errors made by the statistical decoder. The a posteriori correction rule set also includes some phonological rules. Figure 2 shows an example of a correction rule that compensates for the fact that the statistical decoder is overly prone on realising a phonemic dental as a retroflex.

FIGURE 2 HERE.

Both the phonological rules and the inverted phonological rules can use phoneme context (including word stress annotation) and tentative phone context. They can also use estimated phoneme and tentative phone duration as context. Some special rules for high frequency function words even use the orthography as context. A rule may be duration-independent or duration-dependent. A duration-independent rule is applied regardless of the estimated phoneme duration and phone duration and a duration-dependent rule is only applied when the estimated durational context is appropriate. By separating duration-independent and duration-dependent processes, the a posteriori correction rules are able to utilise the information from the statistical decoding maximally to improve the phonetic transcripts.

2.7.3 Transcription System Evaluation

The automatic transcription system has been evaluated against a small manually transcribed gold standard, including the first minute of speech from five randomly selected speakers from the VAKOS database. The transcription system produced an overall phone error rate (PER) of 15.5%, which is an error reduction by 40.4% compared to using the phoneme string for estimating the phone string. For details on the evaluation procedure, cf. Jande (2006b).

Since manual transcription is restricted by a relatively small set of phone symbols, some decisions about phone identity are not obvious, most notably many cases of choosing between a full vowel 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 transcript checking and correction considerably without lowering the quality of the resulting transcript. It is worth noting that if this strategy had been used for compiling the gold standard transcript, the PER would have been somewhat lower. The 15.5% PER is thus a slight under-estimation of the system performance. For the work reported in this paper, pronunciation models are trained on the transcripts produced by the automatic transcription system. If these transcripts are manually corrected and the pronunciation models trained on the corrected versions, the pronunciation models produced are likely to be more accurate than those presented in this paper.

3 Information Included in the Annotation

Values for a set of variables hypothesised to be important for predicting the realisation of a phoneme in its discourse context is attached to each unit at each layer of annotation. This section briefly describes the information attached to the units at each respective layer.

3.1 The Discourse Layer

A set of 'inverted speech rate' measures based on the global *mean phoneme duration* is attached to discourse layer units, calculated as explained in Section 2.4. The discourse layer information also includes four speaking style-related variables: *number of discourse participants*, *degree of formality*, *degree of spontaneity* and *type of interaction*. Table A.1 summarises the discourse layer annotation.

Degree of spontaneity is a five-way variable, taking the values *spontaneous*, *elicited*, *scripted*, *acted* and *read*. Spontaneous speech is, in this context, defined as completely free and uncontrolled, while elicited speech is somehow evoked, e.g. by an interviewer asking questions or a subject being asked to talk about some specific topic. Elicited speech is, however, not based on some written or spoken script. Scripted speech may be a subject retelling a written or spoken text, however not being forced to exactly follow the script. Acted speech is speech closely following a written script, although with acted emotion. Finally, read speech is the result of reading a written text aloud in a 'neutral' fashion.

3.2 The Utterance Layer

In the utterance layer, mostly speaker attributes are annotated. Table A.1 gives a summary of the utterance layer annotation. *Speaker pitch register* is a binary variable that differentiates speakers with a high pitch register (90–600 Hz) from speakers with a low pitch register (60–300 Hz). This variable may interplay with measures based on pitch movement. A coarse four-way division into *utterance types* is used to take the influence of discourse structure into account in dialogue data. Utterances are classified as belonging to one of the four types *statement*, *question/request response*, *answer/response* or *feedback*. For monologues, the default utterance type is *statement*. A set of *mean phoneme duration* measures over the utterance and sets of *pitch range* and *pitch dynamics* 'speech liveliness' (cf. section 2.5) measures are also included in the utterance layer annotation.

Speaker age, dialect and social factors all influence spoken language performance. However, the speakers used for the current pronunciation modelling project are all part of a very coherent group from the perspectives of dialect, sociolect and age. The speakers are all university-educated adults below the age of retirement and they are all speaking the central standard variety of Swedish. Pronunciation variation due to dialectal, social and age factors are thus not modelled in the current effort.

3.3 The Phrase Layer

The complete list of variables included in the phrase layer annotation and their possible values are shown in Table A.1. Two measures associated with the *prosodic weight* of each phrase are calculated: the number of *stressed syllables* and the number of *focally stressed words* included in the phrase.

3.4 The Word Layer

Since the word is the principal conveyor of meaning in language and the principal syntactic unit, there is a large variety of variables that can be included in the word layer. The complete list of variables included in the word layer annotation and their possible values are shown in Table A.1. *Part of speech* and morphological information from the tagger is included in the annotation. *Morphology* is included as a set of tags corresponding to different morphological dimensions. Based on the part of speech tags, a division of words into *word types* (content words vs. function words) is made. A similar variable denoted *function word* has the entire closed set of function words and a generic 'content word' representation as its possible values. There are pronunciation variation strategies specific to certain function words and the *function word* variable should be a strong predictor of this behaviour.

The distance to the preceding and to the succeeding focally stressed word can be important factors in predicting the pronunciation of the current word and these distances (in number of words) are therefore included in the word layer annotation. The presence of a filled pause immediately succeeding or preceding the current word may also be of importance for the pronunciation of the current word. Information about the presence of a filled pause in these two positions is thus included in the annotation. Interrupted (not articulatorily completed) words and other types of "disfluencies" have been shown to have an effect on adjacent words (e.g. Eklund, 1999). For this reason, the presence of interrupted words immediately succeeding or preceding the current word is included in the annotation.

Prosodic boundaries are important for grouping coherent subunits in the speech signal. For listeners, this grouping facilitates parsing the sound stream. Manual prosodic boundary annotation has been supplied for the databases used. In the annotation, prosodic boundaries can be of two types, *strong* and *weak*. The adjacent prosodic boundary variables can thus take the values *strong*, *weak* and *no*. Information about the presence of pauses adjacent to the current word and about the duration of adjacent pauses may be important for predicting the realisation of the word. Two *adjacent pause duration* measures are included in the annotation, *absolute duration* and *normalised duration*, relating the pause duration to the mean duration of all pauses in the database and hence, to the speaking style.

Focal stress may be an important variable for predicting word realisation, since placing stress on a word is to make it more salient; to make it stand out from the surrounding sound stream. For the VAKOS database and one of the radio interviews, manual *focal stress* annotation is available. This information is included in the annotation. It would be possible to use automatic focal stress detection built on e.g. overall intensity and spectral emphasis (Fant et al., 2001; Heldner, 2003) to facilitate annotation when manual annotation is not available. However, no attempt has been made to build or use an automatic focal stress detector for the annotation reported here. Hence, for the remaining speech data, the value of the *focal stress* variable is set to *unknown*.

3.5 The Syllable Layer

The variables included in the syllable layer annotation are presented in Table A.1. Information about the stress and accent of the current syllable is derived from the phonemic representations. Swedish has two different types of word stress, *accent I* and *accent II*. In central standard Swedish, *accent I* has a single stressed syllable while *accent II* has a primary and a secondary stress. There is also a special compound accent similar to *accent II*, with primary stress on the first compound constituent and a secondary stress on the last compound constituent. The *stress* annotation is a simple division between stressed and unstressed syllables, while the *stress type* annotation takes the word accent into account, thus making the *stress type* classification a division into finer stress type classes.

The distances to the nearest preceding stressed syllable and to the nearest preceding syllable with *primary stress* (measured in number of syllables) are included in the syllable layer annotation. The distances to succeeding stresses are also included. The word stress positions are fixed for Swedish words. In the canonical pronunciation representations used for the stress annotation, every word has at least one stressed syllable. The realisation of this word stress is relative to e.g. the stress context. The degree of prominence of a specific syllable in a specific word thus varies with the context. For function words, there is mostly no overt realisation of stress in continuous speech. The

idea behind including the distances to previous and succeeding stresses is that this will give a picture of word stress with higher resolution than the stress of the current syllable alone can give. The initial and final syllables of a word are generally less prone to syllable reduction than medial syllables, which makes the *position of the syllable in the word* an important variable to include in the annotation. The position is annotated as a three-way variable, where each syllable is categorised as either *initial*, *medial* or *final*. The value used for monosyllabic words is *initial*.

3.6 The Phoneme Layer

The variables in the phoneme layer annotation and their values are shown in Table A.1. A set of articulatory features describing the canonical phoneme is associated with each phoneme unit. Five feature dimensions, shared by consonants and vowels, are used. The *sonorant* and *phonological length* dimensions have values shared by consonants and vowels, while all other feature dimensions have separate sets of values for consonants and vowels, respectively.

The *position of the phoneme in the syllable* has been shown to be an important factor for predicting the realisation of the phoneme (cf. e.g. Duez, 1998). Thus, information about in which part of the syllable (*onset*, *nucleus* or *coda*) the phoneme is located is included in the annotation. For a consonant phoneme, the length of the cluster in which it appears and its position in the cluster may be important for its realisation. Hence, information about these variables is included in the phoneme layer annotation. Only consonants adhering to the same syllable as the current phoneme are counted as parts of the current cluster. That is, cluster boundaries are forced at syllable boundaries.

4 Creating Pronunciation Models

Using the annotation from the speech databases, pronunciation models can be created with different types of machine learning methods. If a model is to be used for descriptive purposes, it must be transparent, i.e., it must contain information such that the model can be represented in a format interpretable by a human familiar with linguistic theory. A machine learning paradigm that creates transparent models and is suitable for the type of data at hand is the *decision tree induction* paradigm. A decision tree inducer commonly needs no ad hoc knowledge and can induce models directly from training data. It is thus easy to use once you have the data. For these reasons, the decision tree paradigm has been selected for creating the models reported in this paper. It has not been tested whether the decision tree paradigm necessarily produces the best models. Other machine learning paradigms may be able to create more accurate models or models which meet certain application-specific demands.

4.1 Decision Tree Induction

A decision tree induction algorithm builds a tree level by level using training instances through splitting the set of instances using the optimal attribute for a given sub-set of instances according to some criterion (generally based on entropy minimisation). For creating decision tree pronunciation models, training instances are compiled from the structured 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 the 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. For example, the values of the part-of-speech and morphology variables of the words at positions $n\pm4$ are included, n being the position of the current word in the word layer annotation. The values of the variables of the phonemes at positions $m\pm4$, m being the position of the current phoneme in the phoneme layer annotation, are also

included. For most other variables, a context range of ± 2 is used. Training instances are created for each unit in the phoneme layer annotation, except for the special units <sil> and <junk>. These units are, however, used in the phoneme context attributes.

The task of a finished decision tree model is to take instances in the same format as the training instances and make a decision about the appropriate phone realisation (which may be $\emptyset$) of each instance. The model will thus describe phone level pronunciation only. The relation between a phoneme and its phone realisation can be seen as a phonological process. From a phonological point of view, the models describe processes affecting the presence or absence of phones and processes affecting the broad-phonetic phone identities. However, processes that do not change the broad-phonetic identities of phones, e.g. nasalisation and devoicing of certain phonemes in Swedish, are not handled by the models.

Training data generally contain some degree of noise and a decision tree may be biased toward the particular noise in the data used for inducing the tree (over-trained). However, once a tree is constructed, it 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.

4.2 Decision Tree Inducer and Optimisation

The DTREE program suite (Borgelt, 2004) was used for inducing the pronunciation models presented in this paper. The DTREE inducer can use both attributes with categorical values and attributes with continuous values. A categorical attribute has a finite number of unordered values. For categorical attributes, the tree branches into n branches, where n is the number of values for the attribute occurring in the training data set. For continuous values, the inducer finds a single optimal cut-off point and performs binary branching at this point.

The best classification performance was obtained when selecting attributes with a measure referred to as *symmetric information gain ratio* (Lopez de Mantaras, 1991), allowing the inducer to group discrete values to obtain the optimal number of nodes at each level, and using the default values for all other optimisation options. This was thus the setting used for inducing the final models.

5 Model Evaluation

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 left out partition is then used for evaluation.

A separate tenfold cross validation evaluation was performed for data from each of the three databases (VAKOS, RADIO INTERVIEW and RADIO NEWS) and for the collapsed data set. The prosodic attributes (variables based on pitch and duration measures calculated from the signal) 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 in which the decision tree inducer did not have access to the prosodic information was performed. As a baseline, an evaluation of trees induced from phoneme layer information only was also performed for each data set. Thus, twelve different tenfold cross validation experiments were performed. The models trained with access to different numbers of attributes were trained on the same samples, to make the resulting models comparable. The attribute set including all information is denoted *attribute set A*, the set with prosodic attributes excluded is denoted *attribute set B* and the set with only phoneme layer attributes is denoted *attribute set C*.

Each tree created for the cross validation experiment was pruned and the optimal tree, either pruned or unpruned,

was selected to be used in the evaluation. Although referred to as *unpruned*, the original trees had been subjected to *basic pruning*. In performing this *basic pruning*, the model is pruned only if this does not change the output of the model on the training data.

5.1 Baselines

The results of the pronunciation models can be compared to the results from estimating the phone string with the phoneme string. The phoneme string is the simplest baseline used. However, since there may be assimilation processes always occurring at word boundaries when words are put together, the phonemic representations for isolated words collected from a lexicon may not be a fair baseline. To explore this possibility, some phonological sandhi rules (word boundary rules) were constructed to adapt the phonemic representations for isolated words to their phonemic context.

In the sandhi rule system, three place assimilation rules were included: a *recursive rightward retroflex assimilation rule* (the [+retroflex] feature of a consonant to the left of a word boundary will recursively spread rightwards to [+dental] consonants to the right of the word boundary), a *leftward bilabial assimilation rule* (the [+bilabial] feature of a consonant to the right of the word boundary will spread to an /n/ to the left of the word boundary, changing it to an /m/) and a *leftward velar assimilation rule* (the [+velar] feature of a consonant to the right of the word boundary will spread to an /n/ to the left of the word boundary, changing it to an /ŋ/). Also included in the sandhi rule system were *a leftward voice assimilation rule* (the [+/-voice] feature of a plosive consonant to the right of the word boundary will spread to a plosive with the same place of articulation to the left of the word boundary) and *a double consonant elision rule* (a consonant to the right of the word boundary will be deleted if the same consonant occurs to the left of the word boundary).

The rules are applied in a strict order, but each rule can be set to either *on* or *off*, so that the effects of all combinations of rules resulting from either applying or not applying each rule can be explored. The combination of rules giving the adapted phoneme strings with the highest prediction accuracy (over the entire data set) is used as the second baseline. When exploring the combinations of rules, specific rules are used rather than rules on the general format presented above. For example, the voice assimilation rule can affect three pairs of plosives, /p/-/b/, /t/-/d/ and /k/-/g/ and both the [+voice] feature and the [-voice] feature can spread leftwards. Thus, the voice assimilation rule is split into six rules, which can then be applied (or not applied) separately.

As mentioned, a third baseline used is the result of pronunciation models trained with access only to attributes originating from the phoneme layer annotation. This baseline can be used to show the effect of including variables above the phoneme layer when modelling pronunciation in discourse context.

5.2 Phone Error Rates

Table 2 summarises the results from the cross validation experiments. On average, we get a phone error rate (PER) of 8.2% when training on 90% of the collapsed data set and allowing the decision tree inducer to use all available information (type A tree). Using the phoneme string to estimate phone realisations gives a PER of 20.4%, which means that phone errors can be reduced by 60.0% by using an average pronunciation variation model instead of a phoneme string collected directly from a lexicon.

TABLE 2 HERE.

Applying phonological sandhi rules to adapt the phonemic representations for isolated words to their context did not give rise to any large changes in the PER produced by the phoneme string. All combinations of applying or

not applying each rule in the rule set described in Section 5.1 was tested. The combination of rules giving the largest decrease of PER compared to using the original phoneme string lowered the PER only 0.6 per cent units from 20.4% to 19.8% (the change is, however, statistically significant, $p<0.01$, using the McNemar test). The phonological sandhi rule set giving rise to a reduction of PER is shown in Figure 3.

FIGURE 3 HERE.

As can be seen from Table 2, we get a reduction of PER from 14.2% to 8.2% when switching from a model trained on phoneme level information only (type C tree) to a type A tree. This is an improvement with 42.2%, as can be seen in Table 3.

TABLE 3 HERE.

5.3 Data Size and Speaking Style

It is likely that the data presented in Table 2 reflects the fact that both the amount and the type of training data affects the performance of the models induced. Neither models trained on the VAKOS database nor models trained on the RADIO NEWS database have the lowest PER, although the VAKOS database has the largest number of training instances and the RADIO NEWS database has the most formal, strict type of speech. Instead, the models trained on the RADIO INTERVIEW database show the lowest PER (type A trees). The RADIO INTERVIEW database has the advantages of having relatively formal speech in comparison with the VAKOS database, relatively few speakers and many more training instances than the RADIO NEWS database.

Further, we can see from Table 3 that models trained on the VAKOS database are more dependent on prosodic information and generally on information from layers above the phoneme, while the models trained on the RADIO NEWS database are less dependent on this type of information.

5.4 Attribute Ranking

Table 4 shows the 24 top ranking attributes over the ten optimal type A trees trained on the collapsed data set. The layer from which the attribute originates is used as a prefix in the attribute names. Attributes can refer to the current unit or to units at ± 4 positions from the current unit on the specific annotation layer. Duration measures can be based on the duration of all *phonemes* or on the duration of *vowels* only; they can be based on *normalised* or *absolute* phoneme duration; and they can be based on duration on a *log* scale.

TABLE 4 HERE.

The ranking in the first column of Table 4 is based on the position of the attribute in the ten type A trees. For this measure, the attribute governing the largest number of subtrees (leaves excluded) will get the highest rank (1). The second column weights the subtree count with the number of classifications involving the attribute (over the training data). For this measure, an attribute involved in many classifications can climb in rank even if it does not appear in the absolute top of the tree (near the root). The *phoneme identity* attribute appears in the top node of all trees. This means that it governs all subtrees and is involved in all classifications made by the trees. Hence, *phoneme identity* ends up at the top irrespective of ranking method.

For the trees trained on all available information from all databases, variables from all layers of annotation are used. In fact, from 516 available attributes, as many as 449 were used at least once in the ten trees. However, the phoneme and word layer attributes are the attributes most commonly used in the higher levels of the trees. The top ranking utterance layer attribute shows up at rank 55 using the first ranking method and at rank 43 using the second ranking method. For the first method, the attribute is a phoneme-based duration measure and for the second method, the attribute is a vowel-based duration measure. The top discourse layer attribute is also a vowel-based duration measure and shows up at rank 27 for both ranking methods.

The *word frequency* and *word predictability* attributes both get relatively low ranks (*word frequency* is ranked 67 and 94 by the respective ranking methods and *word predictability* is ranked 91 and 133). The relatively weak predictive strength of these variables may be due to the fact that they are obscured by the *function word* variables, who get high ranks and, to a certain degree, contain information overlapping with the *word frequency* and *word predictability* variables. Also, the *word frequency* and *word predictability* measures are estimated from a corpus of transcribed speech, relatively small in comparison to standard text corpora. These measures would probably be improved if supplemented with data from text corpora.

A large variety of the duration and pitch based measures, respectively, are represented among the variables used by the optimal trees. The first measure based on pitch shows up at place 44 using the first ranking strategy and on place 47 using the second ranking strategy. The highest ranking pitch-based attributes are two different pitch dynamics measures calculated over the phrase. Most of the duration measures seem to be nearly equivalent in terms of predictive power, with vowel-based measures working somewhat better over durationally larger units than over smaller units. Since higher order layer units are large in terms of duration, it is not possible to make exact predictions from these units only and attributes from these levels mostly end up in the lower levels of the decision trees, as a result of the 'greedy' induction algorithm used

5.5 Gold Standard Evaluation

Although it is hard to speculate about how the model performance would be affected by more accurate training data, the transcripts generated by the current models can be evaluated against actual target transcripts. When evaluated against the small gold standard consisting of five minutes of manually transcribed speech from the VAKOS database, the optimal type A trees trained on all data produced a PER of 17.7%, which means that the deterioration in performance when using the model instead of the automatic transcription system is only 12.3% and that the improvement arising from using the model instead of the phoneme string is 31.2%.

5.6 Error Analysis

Table 5 shows the most frequent phone classification errors made by the trees as the share of the total number of errors. It can be seen that errors mostly go both ways. For example, there are equally many erroneous [r] elisions and [r] insertions. We can also see that the choice between a [ə] and a full vowel is a large source of errors.

The errors in choosing between a [ə] and a full vowel are probably not only actual errors, but also artefacts of free variation. That is, a [ə] and a full vowel may be equally correct in many cases. If the model is used in a speech synthesis setting, such deviation from the key transcript due to free variation would neither affect the intelligibility nor the perceived naturalness of the resulting speech. In cases where the classification is actually erroneous, the error would probably not affect intelligibility in any critical way. A more serious type of error is erroneous vowel elision. Erroneous consonant elisions may also in many contexts affect the naturalness and/or intelligibility. Out of the total number of errors produced by the ten optimal models trained on all data, as many as 18.6% were erroneous elisions. However, only 1.6% were erroneous vowel elisions.

TABLE 5 HERE.

5.7 Model Complexity

The ranking of attributes is closer to optimal when using symmetric information gain ratio than when using other selection measures given the type of training data used and thus trees are generally smaller after basic pruning when symmetric information gain ratio is used. Symmetric information gain ratio thus gives both better predictions and less complex models than using e.g. information gain ratio for selecting attributes. For this reason, the effect of pruning on model performance was small for the decision trees evaluated. In most cases, pruning affected model performance (on the test data) positively.

Six pruned trees performed better than their unpruned counterparts. On average over the ten type A trees trained on all data, pruning decreased the PER only by 0.5%, but decreased the average number of attributes used by the models by 82.0% (from 302.8 to 54.5). The model complexity thus dropped significantly as a result of pruning: the average number of nodes decreased by 89.6% (from 4151.9 to 433.1) and the average number of tree levels decreased by 62.2% (from 32.3 to 12.2). Using the McNemar test, the difference in PER between pruned and unpruned models was shown *not* to be significant.

A pruned model is much simpler than an unpruned model and thus requires less input attributes to be obtained. Although the McNemar test showed that there is no gain in predictability associated with pruning, it also showed that there is no loss of predictability associated with pruning. Hence, a pruned model would be the choice in an application. However, it should be noted that although a pruned model uses less attributes than an unpruned model, there are still attributes from all annotation layers used in the pruned models.

6 Conclusions

In this paper, a project aimed at modelling discourse context-specific phone-level pronunciation has been presented. A data-driven approach has been taken for this task and the work involved annotating spoken language with linguistic and related information on levels ranging from the discourse down to articulatory features. Annotation was structured 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. The layers were segmented into their specific unit types and linguistic information was attached to each unit at each layer.

The resulting annotation was used for machine learning to create models describing variation in phoneme realisation. Using the phoneme as the primary unit, a set of training instances, essentially being context-sensitive phonemes, were created. Each instance contained information about the current phoneme, and about the current unit in all higher annotation layers. The instance also contained information about the sequential context of the current unit in each layer.

In a tenfold cross validation experiment, it was shown that including information from multiple layers can improve model performance, most notably for spontaneous speech, where the predictive power of phonological and grammatical information is relatively low. Attributes from all layers of annotation were used in the models with the highest prediction accuracy. The optimal models produced an average phone error rate of 8.2%, which is an improvement of 60.0% compared to using the phoneme string for estimating phone-level realisation. A comparison between models trained only on phone layer attributes and models trained on attributes from all layers showed that the prediction accuracy of pronunciation models could be improved by 42.2% by including information above the phoneme level.

Acknowledgements

The research reported here 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 and supported by the Swedish National Graduate School of Language Technology (GSLT).

I would like to thank my supervisor Rolf Carlson for valuable comments and ideas. Thanks also to everyone who have made tools and resources used for the work described in this paper available. Special thanks Robert Bannert and Peter Czigler for their VAKOS database, to the GROG project participants for the Radio speech data and annotation, to the department of Linguistics at Göteborg University for access to the Göteborg Spoken Language Corpus, to Kåre Sjölander for his excellent free software and for access to his aligner and phoneme models, and to Beáta Megyesi for part-of-speech tagging and parsing software. Finally, I would like to thank my reviewer for useful comments.

References

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E., Ottesjö, C., 2000. The spoken language corpus at the linguistics department, Göteborg University. Forum: Qualitative Social Research 1 (3).

Aycock, J., November 1998. Compiling little languages in Python. In: Proc. 7th Internat. Python Conf. Houston, Texas.

Bannert, R., Czigler, P., 1999. Variations in consonant clusters in standard Swedish. Phonum 7, Reports in Phonetics. Umeå: Umeå University.

Bennett, C., Black, A., March 2005. Prediction of pronunciation variations for speech synthesis: A data-driven approach. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing. Philadelphia, Pennsylvania, pp. 297–300.

Borgelt, C., 2004. Dtree. http://fuzzy.cs.uni-magdeburg.de/∼borgelt/dtree.html.

Brants, T., April 2000. TnT – A statistical part-of-speech tagger. In: Proc. 6th Applied Natural Language Processing Conf. Seattle, Washington, pp. 224–231.

Bruce, G., 1986. Elliptical phonology. In: Papers from the Ninth Scandinavian Conf. on Linguistics. Stockholm, Sweden, pp. 86–95.

Carlson, R., Granström, B., April 1976. A text-to-speech system based entirely on rules. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing. Philadelphia, Pennsylvania, pp. 686–688.

Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E., Swerts, M., May 2002. Boundaries and groupings – The structuring of speech in different communicative situations: A description of the GROG project. In: Proc. Fonetik. Stockholm, Sweden, pp. 65–68.

Carlson, R., Granström, B., Hunnicutt, S., May 1982. A multi-language text-to-speech module. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing. Vol. 3. Paris, France, pp. 1604–1607.

Duez, D., September 1998. Consonant sequences in spontaneous French speech. In: Proc. ESCA Sound Patterns of Spontaneous Speech Workshop (SPoSS-98). La Baume-les-Aix, France, pp. 63–68.

Eklund, R., July 1999. A comparative study of disfluencies in four Swedish travel dialogue corpora. In: Proc. ICPhS Disfluency in Spontaneous Speech Workshop (DiSS). San Francisco, California, pp. 3–6.

Elert, C.-C., 1970. Ljud och ord i svenskan (Sounds and Words in the Swedish Language). Stockholm: Almqvist & Wiksell.

Fant, G., Kruckenberg, A., Liljencrants, J., Botinis, A., September 2001. Prominence correlates. A study of Swedish. In: Proc. Eurospeech. Aalborg, Denmark, pp. 657–660.

Finke, M., Waibel, A., September 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: Proc. Eurospeech. Rhodes, Greece, pp. 2379–2382.

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

Gårding, E., 1974. Sandhiregler för svenska konsonanter (Sandhi rules for Swedish consonants). In: Svenskans beskrivning 8. Lund, Sweden, pp. 97–106.

Heldner, M., 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish. Journal of Phonetics 31 (1), 39–62.

Hermes, D., Gestel, J., 1991. The frequency scale of speech intonation. J. Acoust. Soc. Amer. 90, 97–102.

Jande, P.-A., August 2003. Phonological reduction in Swedish. In: Proc. Internat. Congress of Phonetic Sciences. Barcelona, Catalonia, pp. 2557–2560.

Jande, P.-A., September 2005. Inducing decision tree pronunciation variation models from annotated speech data. In: Proc. Interspeech. Lisboa, Portugal, pp. 1945–1948.

Jande, P.-A., May 2006a. Integrating linguistic information from multiple sources in lexicon development and spoken language annotation. In: Proc. LREC workshop on merging and layering linguistic information. Genoa, Italy, pp. 1–8.

Jande, P.-A., December 2006b. Modelling phone-level pronunciation in discourse context. Ph.D. thesis, KTH, School of Computer Science and Communication, Department of Speech, Music and Hearing, Stockholm, Sweden.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W., May 2001. The effect of language model probability on pronunciation reduction. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing. Vol. 2. Salt Lake City, Utah, pp. 2118–2121.

Lopez de Mantaras, R., 1991. A distance-based attribute selection measure for decision tree induction. Machine Learning 6 (1), 81–92.

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

Megyesi, B., 2002b. Shallow parsing with PoS taggers and linguistic features. J. Machine Learning Research 2, 639–668.

Miller, C., November 1998. Individuation of postlexical phonology for speech synthesis. In: Proc. ESCA/COCOSDA 3rd Internat. Workshop on Speech Synthesis. Jenolan Caves, Australia, pp. 133–136.

More, B., Glasberg, B., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Amer. 74, 750–753.

Nolan, F., August 2003. Intonational equivalence: An experimental evaluation of pitch scales. In: Proc. Internat. Congress of Phonetic Sciences. Barcelona, Catalonia, pp. 771–774.

Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfeld, T., October 1996. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In: Proc. Internat. Conf. on Spoken Language Processing. Philadelphia, Pensylvania, pp. 1039–1042.

Ridings, D., 2002. Swedish resources for language engineering. http://folk.uio.no/danielr/swedish.html.

Sigurd, B., 1965. Phonotactic structures in Swedish. Ph.D. thesis, Lund University.

SIL International, 1995. PCKIMMO. http://www.sil.org/pckimmo.

Sjölander, K., June 2003. An HMM-based system for automatic segmentation and alignment of speech. In: Proc. Fonetik. Umeå, Sweden, pp. 93–96.

Sjölander, K., 2004. The Snack sound toolkit. http://www.speech.kth.se/snack.

Sjölander, K., Beskow, J., October 2000. WaveSurfer – a public domain speech tool. In: Proc. Internat. Conf. on Spoken Language Processing. Vol. IV. Beijing, China, pp. 464–467.

Sjölander, K., Heldner, M., May 2004. Word level precision of the NALIGN automatic segmantation system. In: Proc. Fonetik. Stockholm, Sweden, pp. 116–119.

Stevens, S., Volkman, J., 1940. The relation of pitch to frequency: A revised scale. American Journal of Psychology 53, 329–353.

Strassel, S., February 2004. Simple metadata annotation specification version 6.2. Linguistic Data Consortium, http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf.

Traunmüller, H., Eriksson, A., 1995a. The frequency range of the voice fundamental in the speech of male and female adults. http://www.ling.su.se/staff/hartmut/ f0_m&f.pdf.

Traunmüller, H., Eriksson, A., 1995b. The perceptual evaluation of F0-excursions in speech as evidenced in liveliness estimations. J. Acoust. Soc. Amer. 97 (3), 1905–1915.

Van Bael, C., van den Heuvel, H., Strik, H., October 2004. Investigating speech style specific pronunciation variation in large spoken language corpora. In: Proc. Internat. Conf. on Spoken Language Processing. Jeju Island, Korea, pp. 586–589.

Werner, S., Wolff, M., Eichner, M., Hoffman, R., May 2004a. Modeling pronunciation variation for spontaneous speech synthesis. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing. Vol. 1. Montreal, Canada, pp. 673–676.

Werner, S., Wolff, M., Eichner, M., Hoffman, R., 2004b. Toward spontaneous speech synthesis – Utilizing language model information in TTS. IEEE Transactions on Speech and Audio Processing 12 (4), 436–445.

Zheng, J., Franco, H., Stolcke, A., May 2000. Rate-dependent acoustic modeling for large vocabulary conversational speech recognition. In: Proc. NIST Speech Transcription Workshop. College Park, Maryland, pp. 57–60.

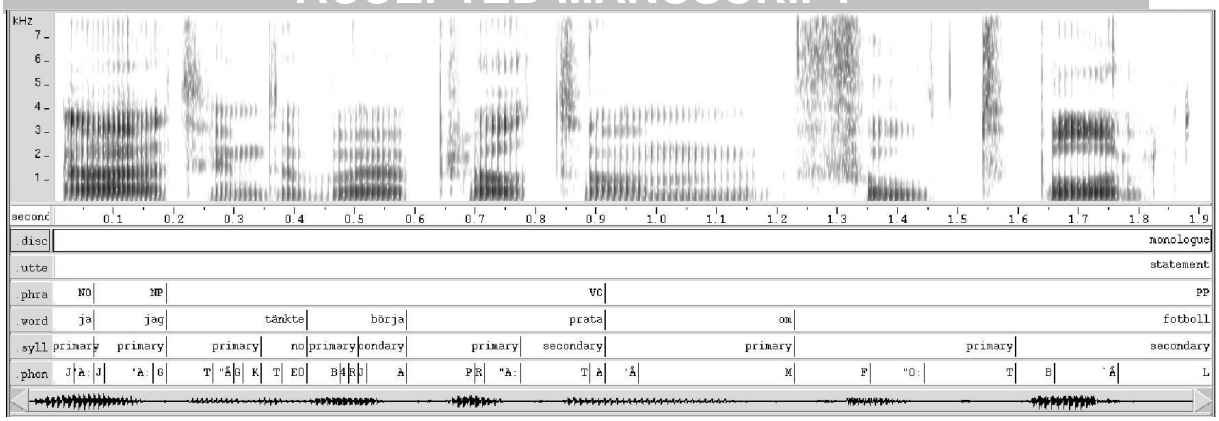


Figure 1. Annotation layers with example annotation aligned to the speech signal.

Table 1
Sets of possible realisations of phonemes.

Consonant	Realisations	Unstressed vowel	Realisations	Stressed vowel	Realisations
p	p	ə	ə		
t	∅, ʈ, t	a	ə, a	a	a
k	∅, k	ɑː	ə, a, ɑː	ɑː	a, ɑː
b	b	e	ə, e	e	e
d	∅, r, ɖ, d	eː	e, eː	eː	e, eː
ɡ	∅, ɡ	ɪ	ə, ɪ	ɪ	ɪ
f	∅, f	iː	ɪ, iː	iː	ɪ, iː
v	v	ʊ	ə, ʊ	ʊ	ʊ
s	s, ʂ	uː	ə, ʊ, uː	uː	ʊ, uː
ɧ	ɧ	ɵ	∅, ɵ	ɵ	ɵ
ç	ç	ʉ̟ː	ɵ, ʉ̟ː	ʉ̟ː	ɵ, ʉ̟ː
h	∅, h	ʏ	ʏ	ʏ	ʏ
m	m	yː	ə, ʏ, yː	yː	ʏ, yː
n	ŋ, ɳ, m, n	ɔ	∅, ə, ɔ	ɔ	ɔ
ŋ	ŋ	oː	ə, ɔ, oː	oː	ɔ, oː
l	∅, ɭ, l	ɛ	∅, ə, ɛ	ɛ	ɛ
ɟ	∅, ɟ	ɛː	ə, ɛ, ɛː	ɛː	ɛ, ɛː
r	∅, r	æ	ə, æ	æ	æ
ʈ	ʈ	æː	ə, æ, æː	æː	æ, æː
ɖ	ɖ	œ	∅, ə, œ	œ	œ
ɭ	∅, ɭ	øː	œ, øː	øː	œ, øː
ɳ	∅, ɳ	œ̟	∅, ə, œ̟	œ̟	œ̟
ʂ	∅, ʂ	œ̟ː	ə, œ̟, œ̟ː	œ̟ː	œ̟, œ̟ː

invert rule [+dental] → [+retroflex]
if [+retroflex] _ $\#_w$ [-retroflex] ∧ ¬[r]

Figure 2. Example of inversion rule. The rule stipulates that a [+retroflex] resulting from the phonological rule [+dental] → [+retroflex] should be cancelled (i.e. that the inverse rule should be applied) if the resulting [+retroflex] is not followed by a retroflex or an [r] in the output from the statistical decoder (and previously applied correction rules).

$C_\alpha \rightarrow \emptyset$ / _ $\#_w C_\alpha$
n → ŋ / _ $\#_w$k
n → m / _ $\#_w$[+bilabial]
g → k / k$\#_w$ _
k → g / g$\#_w$ _
$C_\alpha \rightarrow \emptyset$ / _ $\#_w C_\alpha$

Figure 3. The set of phonological sandhi rules giving rise to a reduction of phone error rate. In these rules, $\#_w$ denotes a word boundary and C_α denotes a specific consonant. The double consonant elision rule is applied first and then re-applied when all other rules have been applied.

Table 2
Mean and standard deviation of phone error rate (PER) for sets of decision trees. Each mean and standard deviation is based on the ten optimal trees resulting from one of the twelve tenfold cross validation experiments. Attribute set C contains only attributes from the phoneme layer, set B contains all attributes except prosodic ones and set A contains all available attributes.

Database	All			VaKoS			Radio Interview			Radio News		
# training instances	93 996			52 263			31 779			9936		
# evaluation instances	10 444			5807			3531			1104		
Attributes	set A	set B	set C	set A	set B	set C	set A	set B	set C	set A	set B	set C
$\bar{x}_{PER}$ (per cent)	8.17	13.18	14.15	9.08	14.99	15.53	8.91	12.43	13.48	9.24	10.70	11.53
σ_{PER} (per cent)	0.25	0.36	0.38	0.31	0.33	0.49	0.34	0.70	0.54	0.72	0.95	0.93

Table 3
Error reduction (per cent) as a consequence of using trees trained on all attributes compared to using trees trained on subsets of attributes. Type C trees are trained with access only to phoneme level attributes, type B trees are trained with access only to non-prosodic attributes and type A trees are trained with access to all attributes.

Database	type B vs. type A	type C vs. type A
All	37.97	42.23
VaKoS	39.42	41.50
Radio Interview	28.33	33.93
Radio News	13.63	19.87

Table 4
The 24 top ranking attributes for trees trained on all information from all databases (type A trees), using two different ranking methods.

	Subtree rank	Subtree · classification rank
1	phoneme_identity	phoneme_identity
2	phoneme_identity+1	phoneme_identity+1
3	word_duration_phonemes_absolute	word_duration_phonemes_absolute
4	word_function_word-1	word_function_word
5	word_function_word+1	word_function_word+1
6	phoneme_identity+4	word_function_word-1
7	phoneme_identity-2	phoneme_identity-1
8	word_function_word	word_duration_vowels_absolute
9	phoneme_identity-1	phoneme_identity+2
10	phoneme_identity+2	phoneme_identity-3
11	phoneme_identity-4	phoneme_identity+4
12	phoneme_identity+3	phoneme_identity+3
13	phoneme_identity-3	phoneme_identity-2
14	word_duration_vowels_absolute	phoneme_identity-4
15	syllable_stress_type	syllable_stress_type
16	syllable_nucleus	phrase_duration_phonemes_absolute
17	word_duration_vowels_normalised	word_duration_vowels_normalised
18	word_duration_vowels_log_absolute	syllable_nucleus
19	syllable_position_in_word	phoneme_feature_py+1
20	phoneme_feature_py+1	syllable_position_in_word
21	phrase_duration_phonemes_log_absolute	word_duration_vowels_log_absolute
22	phrase_duration_phonemes_absolute	word_duration_phonemes_log_normalised
23	phrase_duration_phonemes_log_normalised	phrase_duration_phonemes_log_absolute
24	syllable_duration_vowels_absolute	phrase_duration_phonemes_log_normalised

Table 5
The most frequently occurring phone classification errors.

	$\text{Phone}_{\text{key}}$	$\text{Phone}_{\text{model}}$	Occurrences	Share of total
1	∅	r	566	6.63%
2	r	∅	504	5.90%
3	r	d	466	5.46%
4	ə	ɪ	433	5.07%
5	e	ə	389	4.56%
6	ə	a	375	4.39%
7	a	ə	320	3.75%
8	∅	d	294	3.44%
9	∅	n	260	3.05%
10	∅	g	226	2.65%
11	g	∅	178	2.09%
12	d	∅	173	2.03%
13	ɔ	oː	144	1.69%
14	n	∅	140	1.64%
15	∅	ɔ	131	1.53%
16	ə	e	129	1.51%
17	d	r	125	1.46%
	e	eː		
	ɪ	ə		
	∅	f		
18	s̩	s	121	1.42%
	∅	j		
19	ɪ	iː	114	1.34%
20	oː	ɔ	112	1.31%
21	∅	ɵ	95	1.11%
22	ɛː	æ	91	1.07%
23	∅	h	86	1.01%
24	ə	ɔ	84	0.98%

A Annotation

Table A.1
Annotation included in the six annotation layers.

Variable	Values	***Discourse layer***
Number of discourse participants	*monologue*, *two-part dialogue*, *multi-part dialogue*	
Type of interaction	*human-directed*, *computer-directed*	
Degree of formality	*formal*, *informal*	
Degree of spontaneity	*spontaneous*, *elicited*, *scripted*, *acted*, *read*	
Mean phoneme duration	Several continuous measures, $\mathbb{R}$	
Variable	**Values**	***Utterance layer***
Speaker pitch register	*high*, *low*	
Utterance type	*statement*, *question/request response*, *answer/response*, *feedback*	
Pitch dynamics	Several continuous measures, $\mathbb{R}$	
Pitch range	Several continuous measures, $\mathbb{R}$	
Mean phoneme duration	Several continuous measures, $\mathbb{R}$	
Variable	**Values**	***Phrase layer***
Phrase type	*adverb phrase*, *adjective phrase*, *noun phrase*, *prepositional phrase*, *verb cluster*, *infinitive phrase*, *numeral expression*, *no phrase*	
Phrase length (words)	Continuous, $\mathbb{Z}$	
Phrase length (syllables)	Continuous, $\mathbb{Z}$	
Phrase length (phonemes)	Continuous, $\mathbb{Z}$	
Phrase length label	*long*, *medium*, *short*	
Prosodic weight (stresses)	Continuous, $\mathbb{Z}$	
Prosodic weight (foci)	Continuous, $\mathbb{Z}$	
Pitch dynamics	Several continuous measures, $\mathbb{R}$	
Pitch range	Several continuous measures, $\mathbb{R}$	
Mean phoneme duration	Several continuous measures, $\mathbb{R}$	
Variable	**Values**	***Word layer***
Part of Speech	*adverb*, *determiner*, *wh-adverb*, *wh-determiner*, *wh-pronoun*, *possessive wh-pronoun*, *infinitival* marker, *interjection*, *adjective*, *conjunction*, *noun*, *participle*, *verb particle*, *proper name*, *pronoun*, *preposition*, *possessive pronoun*, *cardinal number*, *ordinal number*, *subjunction*, *foreign word*, *verb*	
Morphology (gender)	*common*, *neutre*, *masculine*, *unspecified*, *no value*	
Morphology (number)	*singular*, *plural*, *unspecified*, *no value*	
Morphology (definiteness)	*indefinite*, *definite*, *unspecified*, *no value*	
Morphology (case)	*nominative*, *genitive*, *no value*	
Morphology (pronoun form)	*subject*, *object*, *unspecified*, *no value*	
Morphology (tense/aspect)	*present*, *preterite*, *infinitive*, *imperative*, *supinum*, *perfect*, *no value*	
Morphology (mood)	*conjunctive*, *no value*	
Morphology (voice)	*active*, *passive/s-form*, *no value*	
Morphology (degree)	*positive*, *comparative*, *superlative*, *no value*	
Word type	*content word*, *function word*	
Function word	*content word*, set of function words	
Word predictability	Continuous, $\mathbb{R}$	
Global word probability	Continuous, $\mathbb{R}$	
Position in phrase	*initial*, *medial*, *final*	
Position in collocation	*initial*, *medial*, *final*	
Word repetitions (full-form)	Continuous, $\mathbb{Z}$	
Word repetitions (lexeme)	Continuous, $\mathbb{Z}$	

→

Variable	Values	***Word layer cont.***
Left-adjacent filled pause	*yes*, *no*	
Right-adjacent filled pause	*yes*, *no*	
Left-adjacent interrupted word	*yes*, *no*	
Right-adjacent interrupted word	*yes*, *no*	
Left-adjacent prosodic boundary	*strong*, *weak*, *no*	
Right-adjacent prosodic boundary	*strong*, *weak*, *no*	
Left-adjacent pause	*yes*, *no*	
Right-adjacent pause	*yes*, *no*	
Left-adjacent pause duration	Two continuous measures, $\mathbb{Z}$	
Right-adjacent pause duration	Two continuous measures, $\mathbb{Z}$	
Word length (syllables)	Continuous, $\mathbb{Z}$	
Word length (phonemes)	Continuous, $\mathbb{Z}$	
Word length label	*long*, *medium*, *short*	
Focal stress	*focally stressed*, *not focally stressed*, *unknown*	
Distance to previous focus (words)	Continuous, $\mathbb{Z}$	
Distance to next focus (words)	Continuous, $\mathbb{Z}$	
Pitch dynamics	Several continuous measures, $\mathbb{R}$	
Pitch range	Several continuous measures, $\mathbb{R}$	
Mean phoneme duration	Several continuous measures, $\mathbb{R}$	

Variable	Values	***Syllable layer***
Stress	*stressed*, *unstressed*	
Stress type	*no stress*, *(primary) stress in accent I word*, *primary stress in accent II word or compound*, *secondary stress in accent II word*, *secondary stress in compound*	
Dist. to prev. stress (syllables)	Continuous, $\mathbb{Z}$	
Dist. to prev. prim. stress (syllables)	Continuous, $\mathbb{Z}$	
Dist. to next stress (syllables)	Continuous, $\mathbb{Z}$	
Dist. to next prim. stress (syllables)	Continuous, $\mathbb{Z}$	
Syllable length (phonemes)	Continuous, $\mathbb{Z}$	
Syllable nucleus	Vowel symbols (cf. Table 1)	
Position in the word	*initial*, *medial*, *final*	
Mean phoneme duration	Several Continuous measures, $\mathbb{R}$	

Variable	Values	***Phoneme layer***
Phoneme identity	Phoneme set (cf. Table 1), <sil>, <junk>	
Sonorant	*yes*, *no*	
Phonological length	*long*, *short*	
Manner of articulation/frontedness	*stop*, *fricative*, *nasal*, *approximant*, *lateral approximant*, *front*, *central*, *back*	
Place of articulation/openness	*bilabial*, *labiodental*, *alveolar*, *dental*, *retroflex*, *palatal*, *velar*, *glottal*, *close*, *close-mid*, *mid*, *open-mid*, *open*	
Voicing/lip rounding	*voiced*, *unvoiced*, *rounded*, *unrounded*	
Position in syllable	*onset*, *nucleus*, *coda*	
Consonant cluster length	Continuous, $\mathbb{Z}$	
Position in cluster	Continuous, $\mathbb{Z}$	
Phone identity	Phoneme set (cf. Table 1), $\emptyset$, <sil>, <junk>	