Integrating Linguistic Information from Multiple Sources in Lexicon Development and Spoken Language Annotation

Per Anders Jande

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

Abstract

In this paper, two related spoken language-oriented projects are presented. Both projects deal with integrating linguistic information from multiple sources. The first project described is the development of a multi-purpose central lexicon database including phonemic representations. Special emphasis is put on central availability and facilitating incremental development. The second project described is a spoken langue annotation project aimed at creating data for data-driven pronunciation modelling. The annotation is designed to form a general description of discourse context, including variables from the discourse level down to the articulatory feature level. A multi-layer annotation scheme for spoken language is described and the information included in the annotation is presented. Models of pronunciation variation induced from the annotation are evaluated in a tenfold cross validation experiment. On average, the models produce 8.1% errors on the phone level. Models trained on phoneme level information only produce an average error of 14.2%. This means that including information above the phoneme level in the context description can improve model performance by 42.6%.

1. Introduction

Studies of spoken language commonly involve various types of linguistic information. For example, in data-driven modelling of various spoken language phenomena, it is often necessary to annotate spoken language data with information on the phoneme and/or phone level as well as information on the word level, such as part of speech and morphology. At the development of speech synthesis systems and automatic speech recognition systems, pronunciation lexica are important.

In this paper, two spoken language-oriented projects are presented. A common denominator of the projects is that they deal with integrating linguistic information from multiple sources. The first project discussed is the development of a multi-purpose central lexicon database, which is used for the annotation of spoken language and in various other contexts. The second project, which is the main focus of this paper, is a data-driven approach to modelling phone-level pronunciation variation, involving the annotation of spoken language with various kinds of linguistic information in multiple layers.

2. A Central Lexicon Database

A multi-purpose central lexicon database called CENTLEX is being developed at the department of Speech, Music and Hearing (TMH) and the Centre for Speech Technology (CTT) at KTH. The lexicon is based on lexical resources of different types and on different formats, developed for various research projects at TMH/CTT over the years. The information is stored in a relational database with separate tables for different types of information.

2.1. Information Included in the Lexicon

CENTLEX is a full-form lexicon, with each entry minimally containing an orthographic word form and a grammatical analysis (part of speech and morphology). An entry can also have an arbitrary number of phonemic representations, ordered by their probability of use. Each phonemic representation can be enriched with information about the intended context of the representation (e.g. *reduced form* or *foreign language*). Such information is added e.g. for proper names, since orthographically identical names may be pronounced differently depending on the native language environment of the person bearing the name. An entry also contains information about the probability of the particular grammatical analysis given the orthographic word (estimated from a large automatically tagged text corpus). Presently, the database contains about 410,000 entries with 330,000 unique orthographic word forms.

2.2. Availability

One of the main ideas with the CENTLEX database is that all lexical data used in projects at TMH and within CTT is stored centrally, so that the data is immediately and easily available for all researchers at the department and for all partners involved in the Centre. Lexicon-related work conducted in different projects can be easily integrated with the central lexical resource, and the results immediately available for all users. Standards for mapping between the CENTLEX format and several commonly used formats have been developed to facilitate information sharing.

An interface to the database on the TMH internal web makes it possible to search the lexicon and to check out purpose-specific lexica with the set of information requested on several different output formats. Selected users also have the possibility to edit the lexicon via the web interface, to stimulate continuous lexicon expansion and improvement of existing data. The web interface is not suited for large-scale changes of the database, so a stand-alone annotation/correction tool has been developed for lexicon development on a larger scale. This tool stores information on a CENTLEX import format, so that it can be easily incorporated with the database.

The lexicon is thus incrementally built and the latest version is always available at a central location. Some of the information first included in the database has been automatically generated and the initial information merger was done with automatic methods. The data thus has to be checked with respect to quality, which is done continuously. Sub-

sequently added information is, however, mostly information which is manually obtained or checked. Each lexicon entry is annotated with information about whether it has been manually checked/corrected, by whom and when, to separate information of different quality.

2.3. Applications

Thus far, the CENTLEX database has been used as a lexicon in an experimental speech synthesis system (used in various research-oriented applications at the department of Speech, Music and Hearing at KTH) and in a large vocabulary speech recognition system. CENTLEX has also been used for training grapheme-to-phoneme conversion rules for commercial speech synthesis and as a lexicon for commercial speech synthesis applications. It has further been used as a reference in the development of a system for production of talking books with synthetic speech for visually impaired and dyslectic university students. Finally, CENTLEX has been used for annotation in research projects aimed at context-sensitive prosody prediction and phone-level pronunciation prediction.

3. Pronunciation Variation Modelling

Although there is a certain degree of individual and random variation in the pronunciation of words in context, the variation 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. The aim in the pronunciation variation modelling project described here is to model this systematic variation inherent to a language variety, with the focus on variation in phone level realisation. The target language variety used in the work presented in this paper is central standard Swedish.

3.1. Annotating Spoken Language Data

The methods used for pronunciation variation modelling are data-driven. Spoken language is annotated with various kinds of linguistic and related information, which is used by machine learning algorithms to create pronunciation models. The phoneme is the central unit in the approach 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. This multi-variable linguistic context description is then used to predict the context-sensitive realisation of the phoneme.

The results reported in this paper are based on recent additions to the annotated data. The effect of making information on different linguistic levels available as predictors of phone level pronunciation is investigated and the predictive power of specific linguistic variables is discussed.

3.2. Background

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

The influence of various other variables on the pronunciation of words has also been studied, but these have mostly been studied in isolation. When more variables are taken into account, the number of variables simultaneously under study is in most cases limited to less than a handful. Describing the discourse context more generally, including a large variety of linguistic and related variables, enables studies of the interplay between various information sources on e.g. phone-level pronunciation.

Machine learning methods can be used for such studies. A model of pronunciation variation created through machine learning can be useful in speech technology applications, e.g. for creating more dynamic and natural-sounding speech synthesis. In addition to models which can predict the pronunciation of words in context, it is possible to create models which are descriptive and to some degree explains the interplay between different types of variables involved in the predictions.

3.3. Speech Data

The speech data used for pronunciation variation modelling is the VAKOS database, originally constructed by Bannert and Czigler (1999) for a phonological study of variation in consonant clusters, a RADIO INTERVIEW database and a RADIO NEWS database, with recordings originating from *Sveriges radio* (Swedish public service radio).

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 also includes some manual annotation at different levels. 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 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.

3.4. A Multi-Layer Annotation System

The annotation used for pronunciation variation 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

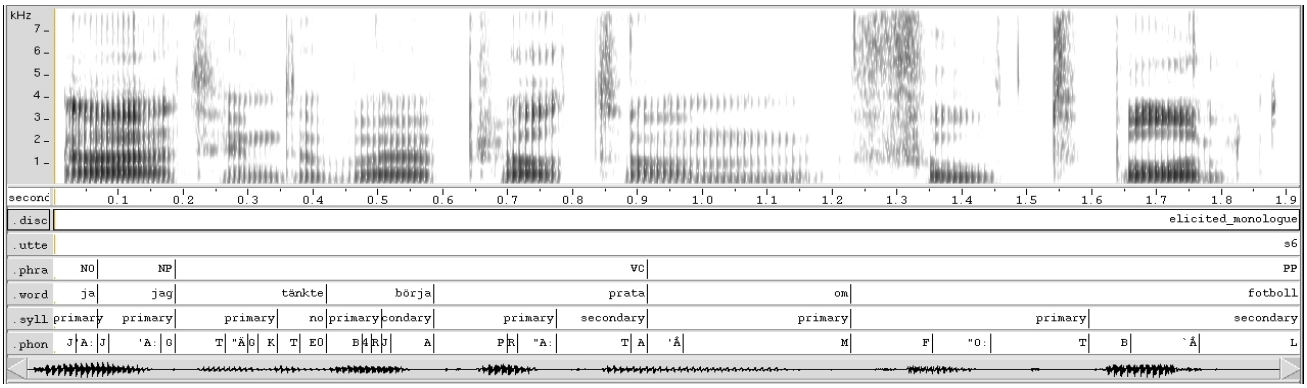

Figure 1: Annotation layers with example annotation aligned to the speech signal

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. A unit can thus pass on its information to all the units within its bounds in the lower order layers. 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 variation 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 unit. Having the information stored in different layers enables easy access to the sequential context information.

3.5. Segmentation

With some minor exceptions, automatic methods are used for segmentation, however with manual supervision to improve accuracy at some intermediate stages. The annotation process begins with segmenting 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. An utterance is in this context 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.

Automatic segmentation begins at the word level. Given an orthographic string, the corpus is segmented into word units using an automatic aligner Sjölander (2003). Manual correction of the word layer segmentation is performed, since all succeeding annotation depends on this segmentation and increases in the segmentation accuracy on this level gives large improvements in the accuracy of successive annotation. Manual word layer segmentation was already included in the VAKOS database.

The phrase layer is segmented with the help of a shallow parser (Megyesi, 2002) using a string of tags produced by a part of speech and morphological tagger. 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.

The phoneme layer is segmented word-by-word using the word boundaries and phonemic representations from the CENTLEX database as input to an automatic aligner (Sjölander, 2003). The phonemes are clustered into syllables with forced syllable boundaries at word boundaries and the syllable layer is segmented using this clustering and the durational boundaries from the phoneme level segmentation.

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*>. 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*> (for pauses) or a <*junk*> special phoneme unit is used as a place filler at the phoneme layer, but no additional annotation is supplied on lower order layers. Every special word layer unit is, however, included in a phrase unit, an utterance unit and in the discourse unit.

3.6. Adding Information to the Units

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. The following sections will briefly describe the information attached to the units at each layer.

3.6.1. The Discourse Layer

A set of 'inverted speech rate' measures based on the global *mean phoneme duration* is attached to discourse layer units. Phoneme durations are estimated from the automatic alignment of the phonemic word representations to the signal. 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*.

3.6.2. The Utterance Layer

In the utterance layer, mostly speaker attributes are annotated. *Speaker pitch register* is a binary variable that differentiates speakers with a high pitch register from speakers with a low pitch register. This variable may interplay with measures based on pitch movement. A set of *mean phoneme duration* measures over the utterance and sets of *pitch range* and *pitch dynamics* ('speech liveliness') measures are also included in the utterance layer annotation.

3.6.3. The Phrase Layer

An attribute called *phrase type* corresponds to the type of the current phrase according to the shallow parser used for phrase chunking. Also included in the phrase layer annotation is a set of *phrase length* measures: the number of *words*, *syllables* and *phonemes*, respectively, contained by each phase unit. Further, two measures associated with the *prosodic weight* of a phrase are calculated: the number of *stressed syllables* and the number of *focally stressed words* contained by the phrase (focal stress annotation was manually provided for a subset of the speech data). Finally, *pitch dynamics*, *pitch range* and *mean phoneme duration* measures are calculated over each phrase unit.

3.6.4. The Word Layer

The word is generally conceived of as the most central linguistic unit, in that it is the principal conveyor of meaning in language and the principal syntactic unit. There is thus a large variety of features that can be attached to the word units. To begin with, *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 predictability of a word has been shown to be important for the realisation of the word, cf. e.g. Fosler-Lussier and Morgan (1999) and Jurafsky et al. (2001b). Many variables influence the predictability of a word in context. Measures related to word predictability included in the word layer annotation are *word repetitions* and *lexeme repetitions* (the number of times the full-form word and the lexeme, respectively, has been repeated thus far in the discourse), *the position of the word in a phrase*, *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. This measure is an estimation based on a weighted combination of unigram, bigram and trigram probabilities collected from the Göteborg Spoken Language Corpus (Allwood et al., 2002). The *part of speech* variable already mentioned also affects the predictability of a word in context, since there are syntactic constraints governing language production.

The distances to the preceding and the succeeding focally stressed word can be important factors in predicting the pronunciation of the current word and these distances (measured in number of words) are therefore included in the word layer annotation. Information about the presence of a *pause*, a *filled pause* or an *interrupted word* adjacent to the current word is also included. 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.

Word length is measured as the number of syllables and as the number of phonemes, respectively, contained by the word. Finally, some measures of *pitch dynamics*, *pitch range* and *mean phoneme duration* over each word unit are included in the word layer annotation.

3.6.5. The Syllable Layer

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 *accent* annotation takes the word accent into account, thus making the *accent* classification a division into finer stress type classes.

Further, 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. *Syllable length* is measured in number of phonemes. 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. Lastly, a set of *mean phoneme duration* measures over the syllable are calculated.

3.6.6. The Phoneme Layer

The *phoneme identities* included in the phoneme layer annotation are represented by the phoneme symbols from CENTLEX. A set of *articulatory features* describing the phoneme is associated with each phoneme unit. The *position of the phoneme in the syllable* may be important for predicting the realisation of the phoneme. Hence, information about in which part of the syllable (*onset*, *nucleus* or *coda*) the phoneme is located is included in the annotation. A *consonant cluster length* variable takes as its value the length (phoneme count) of the consonant cluster of which the current phoneme is a part. This measure defaults to 0 for vowels.

The *phone* is the context-dependent realisation of the phoneme. Phonetic identity is the variable to be estimated by the pronunciation variation models and consequently, the phone is used as the key in model training. The phones are supplied by a hybrid automatic transcription system, using statistical decoding and a set of a posteriori correction rules.

A place filler ∅ symbol is used to signal that there is no realisation of a particular phoneme in the phonetic string.
The SNACK sound toolkit (Sjölander and Beskow, 2000) is used for building and decoding statistical models representing the possible realisations of a word. Models are built using an empirically compiled context-insensitive list of possible realisations (tentative phones) for each phoneme and a set of HMM monophone models. The speech signal is parameterised to form a sequence of observations. The path trough the statistical model most closely matching this observation sequence (using Viterbi decoding) can be represented as a string of phones, and this string is the output of the statistical decoder.
Evaluated against a small manually transcribed gold standard, statistical decoding alone was shown to give higher phone error rates (PER) than estimating the phonetic transcript with the phoneme string. However, due to the systematic nature of the errors made by the statistical decoder, a set of correction rules that significantly lowered the error rate could be compiled. The final hybrid transcription system produces an average of 15.5% errors on the phone level when compared to an enlarged gold standard transcription. This means that the PER is reduced by 40.4% compared to using the phoneme string for estimating the phone realisation.
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 symbol and a schwa. Defaulting to the system decision whenever a human transcriber is forced to make ad hoc decisions would increase the speed of manual transcript checking and correction considerably without lowering the transcription quality. 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. Manual correction of the automatically obtained transcripts will most likely result in more accurate pronunciation variation models.

4. Creating Pronunciation Variation Models

Using the annotation from the speech databases, pronunciation variation models can be created with different types of machine learning methods. If the 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 on 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 very 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 is not claimed that 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

Decision trees are induced from a set of training instances compiled from the structured annotation. The training instances are phoneme-sized and can be seen as a set of *context sensitive phonemes* with their respective phone realisations. 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. The algorithm used for inducing the pronunciation variation models is that included in the DTREE program suite (Borgelt, 2004).
Decision tree induction is non-iterative and trees are built level by level, which makes the learning procedure fast. However, the optimal tree is not guaranteed. At each new level created during the tree induction procedure, the set of training instances is split into subsets according to the values of one of the attributes. The attribute selected is the attribute that best meets a given criterion, generally based on entropy minimisation. In the current case, a measure referred to as *symmetric information gain ratio* (Lopez de Mantaras, 1991) is used. The inducer is set to allow grouping of discrete values to obtain the optimal number of nodes at each level.

4.1.1. Pruning

Since training data generally contain some degree of noise, a decision tree may be biased toward the particular noise in the training data (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. At pruning, a sub-tree of a particular node is replaced by a leaf with the most common class of the leaves governed by the sub-tree, when some criterion is met.

4.2. 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 information cannot be fully exploited in its current form in e.g. a speech synthesis context. Thus, it was interesting topic investigate the influence of the prosodic information (variables based on f_0, duration, focal stress and prosodic boundary information) on model results. To investigate this, an experiment where the decision tree inducer did not have access to the prosodic information was performed for each of the four data sets. As a baseline, an evaluation of trees induced from phoneme layer information only was also performed for each data set. The same

Table 1: Mean and standard deviation of phone error rate (PER) for each data set

Database	**All**			**VAKOS**			**RADIO INTERVIEW**			**RADIO NEWS**		
# training instances	93,996			52,263			31,779			9,936		
# evaluation instances	10,444			5,807			3,531			1,104		
Trained on attributes	all	nopro*	pho†	all	nopro*	pho†	all	nopro*	pho†	all	nopro*	pho†
$\bar{x}_{\mathrm{PER}}$ **(per cent)**	8.14	13.08	14.19	9.07	14.90	15.60	8.94	12.32	13.74	9.34	10.57	11.70
σ_{PER} **(per cent)**	0.15	0.25	0.23	0.39	0.49	0.53	0.42	0.30	0.54	1.23	1.23	1.34

**no prosodic attributes, †phoneme level attributes only*

Table 2: Error reduction as a result of making more information available for the decision tree inducer

Database	**All**		**VAKOS**		**RADIO INT.**		**RADIO NEWS**	
Tree types	pho†>all‡	nopro*>all‡	pho†>all‡	nopro*>all‡	pho†>all‡	nopro*>all‡	pho†>all‡	nopro*>all‡
Error reduction (per cent)	42.64	37.77	41.86	39.12	34.93	27.43	20.20	11.65

*‡trained with access to all attributes, *trained access only to non-prosodic attributes, †trained with access only to phoneme level attributes*

randomisation was used under all conditions.
Each tree was pruned under a range of pruning criteria and the tree with the optimal performance on the evaluation data was selected to be used in the evaluation. The pruning criteria used all yielded the same pruned tree and the optimal tree could thus either be the *pruned* tree or the original, *unpruned* tree. The *symmetric information gain ratio* attribute selection measure created trees, which were near the optimal before pruning. Hence, the effect of pruning on model performance was small. In most cases, pruning affected model performance (on the test data) negatively and, on average, pruning gave rise to a *decrease* in model performance with 0.6%. The unpruned trees were actually subjected to *basic pruning*, at which the trees were pruned to the extent that no deterioration of accuracy on the training data occurred. Thus, following "Occam's razor", if there were several trees giving the same result, the simplest of these trees was selected.

5. Results and Discussion

Table 1 summarises the results from the cross validation experiments. On average, we get a phone error rate of 8.1% when training on 90% of the collapsed data set and allowing the decision tree inducer to use all available information.

5.1. Phone Error Rates

Using the phoneme string to estimate phone realisations gives a PER of 20.4%, which means that phone errors can be reduced by 60.2% by using an average pronunciation variation model in stead of using a phoneme string collected directly from a lexicon. Applying phonological sandhi rules to adapt the phonemic representations for isolated words to their context decreased the PER for the phoneme string only to 20.3%. The error reduction resulting from using the pronunciation variation model is thus significant. Further, as can be seen from Table 2, we get a reduction of PER by 42.6% when switching from a classifier trained on phoneme level information only to a classifier trained on all available information.

5.2. Data Size and Speaking Style

It is likely that the PERs presented in Table 1 reflect the fact that both the amount and the type of training data affects the performance of the models induced. If all attributes are used, 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 type of speech. Instead, the models trained on the RADIO INTERVIEW database show the lowest PER. The RADIO INTERVIEW database has the advantages of having relatively formal speech compared to the VAKOS database, relatively few speakers and many more training instances than the RADIO NEWS database.
Further, we can see from Table 2 that models trained on the VAKOS database are more dependent on prosodic information and generally on information on layers above the phoneme, while the models trained on the RADIO NEWS database are less dependent on this type of information.

5.3. Attribute Ranking

Table 3 shows the 18 top ranking attributes over the ten optimal trees trained on all information from all databases. The layer from which the attribute is collected is used as a prefix in attribute names. Attributes can refer to the current unit or to units at ± 4 positions from the current unit at 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.
The ranking in the first column of Table 3 is based on the position of the attribute in the ten trees. For this measure, the attribute governing the largest number of sub-trees (leaves excluded) will get the highest rank (1). The second column weights the sub-tree 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. The *phoneme identity* attribute appears in the top node of all trees. This means that it governs all sub-trees and is involved in all classifications made by the trees.

5.4. Attributes Used by the Models

From Table 4, it can be seen that variables from all layers of annotation are used by the trees trained on all available information from all databases. In fact, from 516 available attributes, as many as 470 were used at least once in the

Table 3: The 18 top ranking attributes for trees trained on all information from all databases

Rank	Rank based on # sub-trees	Rank based on # sub-trees · # classifications
1	phoneme_identity	phoneme_identity
2	phoneme_identity+1	phoneme_identity+1
3	word_function_word-1	word_duration_phonemes_absolute
4	word_duration_phonemes_absolute	word_function_word+1
5	word_function_word+1	word_function_word
6	phoneme_identity+4	word_function_word-1
7	phoneme_identity-2	phoneme_identity-1
8	word_function_word	phoneme_identity+2
9	phoneme_identity-1	phoneme_identity-3
10	phoneme_identity-4	phoneme_identity+4
11	phoneme_identity+2	word_duration_vowels_absolute
12	phoneme_identity-3	phoneme_identity-2
13	phoneme_identity+3	phoneme_identity+3
14	word_duration_vowels_absolute	phoneme_identity-4
15	syllable_accent	syllable_accent
16	syllable_nucleus	phrase_duration_phonemes_absolute
17	word_duration_vowels_normalised	word_duration_vowels_normalised
18	word_duration_vowels_log_absolute	syllable_nucleus

Table 4: Probability of variables from each annotation layer at the top twelve tree levels

Level	P(Phoneme)	P(Syllable)	P(Word)	P(Phrase)	P(Utterance)	P(Discourse)	$\sum$
1	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
2	0.4101	0.0791	**0.4820**	0.0144	0.0144	0.0000	1.0000
3	**0.4113**	0.0493	0.3941	0.1404	0.0025	0.0025	1.0000
4	0.4052	0.0507	**0.4298**	0.0897	0.0145	0.0101	1.0000
5	0.3728	0.0310	**0.4281**	0.1294	0.0310	0.0077	1.0000
6	**0.3936**	0.0330	0.3729	0.1460	0.0348	0.0198	1.0000
7	**0.3952**	0.0316	0.3416	016.27	0.0383	0.0306	1.0000
8	**0.4338**	0.0408	0.3168	013.47	0.0397	0.0342	1.0000
9	**0.4140**	0.0440	0.3299	014.10	0.0543	0.0168	1.0000
10	**0.4180**	0.0384	0.3250	014.77	0.0561	0.0148	1.0000
11	**0.3958**	0.0545	0.3189	015.22	0.0529	0.0256	1.0000
12	**0.4096**	0.0422	0.3293	014.46	0.0562	0.0181	1.0000

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 is a vowel-based duration measure showing up at place 50 using the first ranking strategy and on place 46 using the second ranking strategy. The top discourse layer attribute is also a vowel-based duration measure and shows up at place 31 and 35, respectively.

The *word frequency* and *word predictability* attributes both get ranks around 110. The relatively weak predictive strength of these variables may be due to the fact that they are obscured by the *function word* variable, which gets high ranks. Further, 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 may be improved with text data.

A large variety of the duration and pitch based measures are represented among the variables used by the optimal trees (the first measure based on pitch shows up at place 42 using the fist ranking strategy and on place 55 using the second ranking strategy). Most of the duration measures seem to be nearly equivalent in terms of predictive power, with vowel-based measures working somewhat better over larger units. Units on higher order layers are both larger in terms of duration and conceptually more abstract than units on lower order layers. Because of this, it is not possible to make exact predictions from higher order layer units only and attributes from these levels end up in the lower levels of the decision trees, as a result of the 'greedy' induction algorithm used.

5.5. Effects of Noise

The erroneous classifications possible for a phoneme are limited to the set of realisations for the phoneme found in the training data. Both training and evaluation data contain up to 15.5% errors on the phone level, as previously discussed. Since the phone string is generated by an automatic transcription system with a priori restrictions on the possible realisations of each phoneme, the range of variation is probably less than it would have been if the transcripts had been produced by a human. It is not immediately obvious whether this translates into lower phone error rates for the pronunciation variation models than would have been the case if the phones in the training and evaluation data had been supplied by a human transcriber.

5.6. Gold Standard Evaluation

Although it is hard to speculate about how the model performance would be affected by more accurate training data, the transcriptions generated by the current models can be evaluated against actual target transcriptions. When evaluated against the small gold standard consisting of five minutes of manually transcribes speech from the VAKOS database, the models produce a PER of 16.9%, which means that the deterioration in performance when using the

model instead of the automatic transcription system is only 8.5% and that the improvement using the model instead of the phoneme string is 34.9%

6. Conclusions

In this paper, two related spoken language-oriented projects have been described, each dealing with the issue of integrating linguistic information from multiple sources. First, the work with developing a multi-purpose central lexicon database including phonemic representations was described. The central ideas behind this project are central availability and incremental development. Tools for facilitating continuous and simultaneous lexicon development have been created.

Second, a project aimed at modelling phone-level pronunciation in discourse context was presented. A data-driven approach was taken for this task and the work involved annotating spoken language with linguistic and related information ranging from the discourse level down to articulatory feature level. 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 level.

The resulting annotation was used for machine learning of 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 annotation layers above. Instances also contained information about the sequential context of the current unit in each layer.

In the evaluation of models created from the multi-layer linguistic annotation, emphasis was put on the effects of adding information of different types to the training data in addition to phoneme-level variables. It was shown that including information from multiple layers improves 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 and as many as 470 out of 516 available attributes were actually used by at least one of the models (optimally pruned decision trees) in a tenfold cross validation experiment. The optimal models produced an average phone error rate of 8.1%, which is an improvement with 60.2% 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 could be improved by 42.6% by adding attributes for units above the phoneme layer.

The classification keys used at model training were generated by an automatic transcription system with access to the speech signal. Evaluated against gold standard transcriptions, the models produced a phone error rate of 16.9%. This means that the deterioration in performance when using the model instead of the automatc transcription system is only 8.5% and that the improvement using the model instead of a phoneme string from a lexicon is 34.9%.

Acknowledgements

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

7. References

J. Allwood, L. Grönqvist, E. Ahlsén, and M. Gunnarsson. 2002. Göteborgskorpusen för talspråk (The Göteborg spoken language corpus). In *Nydanske Sprogstudier 30*, pages 39–58. København: Akademisk Forlag.

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

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

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

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

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

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

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

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

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

D. Jurafsky, A. Bell, M. L. Gregory, and W. D. Raymond. 2001b. The effect of language model probability on pronunciation reduction. In *Proceedings of ICASSP*, volume 2, pages 2118–2121.

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

B. Megyesi. 2002. Shallow parsing with PoS taggers and linguistic features. *Journal of Machine Learning Research*, 2:639–668.

K. Sjölander and J. Beskow. 2000. WaveSurfer – a public domain speech tool. In *Proceedings of ICSLP*, pages 464–467.

K. Sjölander. 2003. An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik*, pages 93–96.

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