A MACHINE LEARNING APPROACH TO SWEDISH WORD
PRONUNCIATION

Hanna Lindgren and Jessica Granberg
Infovox, Telia Promotor AB

ABSTRACT

This study focuses on word pronunciation in Text-to-Speech systems for Swedish. The purpose is to investigate whether machine learning techniques match knowledge-based systems in Swedish word pronunciation. The experiments show a maximum grapheme accuracy of just over 97%, and word accuracies from 67.0% for word pronunciation excluding stress assignment, which compares favourably to existing knowledge-based state-of-the-art systems.

1. INTRODUCTION

Word pronunciation is one topic out of many in the Natural Language Processing (NLP) area. Traditionally, many word pronunciation and other NLP systems use knowledge-based methods. The disadvantages with such methods are the language-dependency and the time-consuming work required to construct rules. With a data-oriented machine learning approach, it is possible to create a word pronunciation system for any language faster and with less language-specific and linguistic knowledge.

The study presented in this article is an attempt to use machine learning techniques in Swedish word pronunciation to investigate if they match the performance of existing knowledge-based methods. Two experiments were performed using the TiMBL machine learning system. The main experiment did not include stress or lexical tone, i.e. it treated pure grapheme-to-phoneme conversion (GP). A preliminary study in grapheme-to-phoneme conversion including stress and lexical tone (GPS) was also performed.

First the lexicon was transformed into the required TiMBL input format, which included grapheme-to-phoneme alignment of the lexicon entries. A training-testing procedure was conducted with the algorithms in TiMBL and the evaluation of the results involved comparison to a commercial knowledge-based word pronunciation system and a comparison of performance between the algorithms in TiMBL and the grapheme contexts used. Earlier experiments with similar set-ups and algorithms were also referenced.

2. THE ALGORITHMS

The machine learning algorithms used in this study are available in the software package TiMBL, Tilburg Memory-Based Learner [5]. Two main machine learning approaches are implemented in TiMBL: the memory-based IB1 and the decision tree-based algorithm IGTree. In IB1, Weighted Overlap (WO) or Modified Value Difference Metric (MVDM) are used for estimating the distance between patterns in the training data. WO defines the distance as the number of feature-value pairs with distinct values. In MVDM, the distance between the patterns is the sum of the differences between the target probabilities for each feature-value pair. The construction of decision trees in IGTree is based on Information Gain or Gain Ratio weighting, which are measures of the relative importance of each feature. Gain Ratio is a modification of Information Gain that has been developed to normalise the weighting for patterns with different numbers of feature-values [5], [7]. The weighting information is also used in IB1 to adjust the similarity between instances in the training set.

3. EXPERIMENT SET-UP

The experiments in this study are based on data from the Swedish Pronunciation Lexicon, Swelex. compiled by Per Hedelin and Anders Jonsson at the Dept. of Information Theory, Chalmers University of Technology. Swelex consists of just over 110,000 words with transcriptions including four categories for stress and lexical tone.

TiMBL demands an input format consisting of one instance per line. In word pronunciation, it is natural that each instance (line) consists of a grapheme surrounded by some left and right neighbour graphemes (context), and the target phoneme. Since the lexicon was word-based, it had to be reconstructed.

Large contexts seem likely to give higher accuracies than small, since graphemes that do not immediately surround the focus grapheme often influence the target pronunciation, although the improvements with contexts larger than 3X3 (i.e. three graphemes to the left and three to the right) are modest [9]. Due to the increase in processing time and computational expense caused by large data-files, tests with larger contexts were not carried out in this experiment.

Since all features used in this experiment were graphemes, the number of possible feature-values was considered constant. Therefore, Gain Ratio and Information Gain weighting should give the same results (Section 2) and the experiments were run only with Gain Ratio. Both IB1 and IGTree were used. The IB1 experiments were run with WO and MVDM.

For a comparison, the words in Swelex were transcribed with the knowledge-based system Rulsys [3], which is the word pronunciation module in the Text-to-Speech (TTS) synthesizer Infovox’ provided by Telia Promotor AB. Just as the Swelex transcriptions, the Rulsys transcriptions contained stress/tone information, but only two categories as opposed to the four in Swelex. Due to these differences, a direct comparison between the two systems was not possible, and one of the tests was run with stress and lexical tone removed both from Rulsys and Swelex.

1 URL: http://www.infovox.se
The results were obtained by running a 10-fold cross-validation (10-FC) test [11]. The training data from the lexicon consisted of approximately one million instances and the average duration of the training-testing sequences was about three hours with IGTree and thirty hours with IB1 on a SUN Sparc Ultra-1.

The TiMBL results were on a grapheme level and the results from Rulsys were word-based. Therefore, the accuracies were not comparable without transforming the results from the former into a word-based form. It seemed reasonable to assume that a poor grapheme-based result would also be a poor word-based result. Hence, only the results from the best performing TiMBL experiments were transformed into a word-based form.

4. RESULTS

4.1. Grapheme Accuracy

For all the tests performed with TiMBL, the grapheme accuracies were as shown in Table 1. The best accuracies were 97.21% for GP and 94.01% for GPS. Overall, MVDM performed slightly better than or equal to WO.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Context</th>
<th>Similarity</th>
<th>Accuracies (%)</th>
<th>GP</th>
<th>GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGTree</td>
<td>2X2</td>
<td>–</td>
<td>96.30</td>
<td>91.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3X3</td>
<td>–</td>
<td>96.93</td>
<td>93.19</td>
<td></td>
</tr>
<tr>
<td>IB1</td>
<td>2X2</td>
<td>WO</td>
<td>96.43*</td>
<td>91.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3X3</td>
<td>WO</td>
<td>97.21</td>
<td>93.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3X3</td>
<td>MVDM</td>
<td>97.19</td>
<td>94.01</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Grapheme-phoneme accuracies from TiMBL. Best results in bold type.

*Only 10% of the material has been run, i.e. step one of the 10-fold cross-validation

Table 2 shows the Gain Ratio weighting for the tests with the highest accuracies. In both cases, feature number four (the focus grapheme) was considered the most important.

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Weighting</th>
<th>GP</th>
<th>WO</th>
<th>GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left context</td>
<td>1</td>
<td>0.0387</td>
<td>0.0699</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0706</td>
<td>0.0943</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.1765</td>
<td>0.1882</td>
<td>3</td>
</tr>
<tr>
<td>Focus</td>
<td>4</td>
<td>0.8840</td>
<td>0.8854</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.1974</td>
<td>0.2096</td>
<td>2</td>
</tr>
<tr>
<td>Right context</td>
<td>6</td>
<td>0.0799</td>
<td>0.1094</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.0386</td>
<td>0.0680</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Feature weights from the tests with highest accuracies.

4.2. Word Accuracy

For both Rulsys and the best-performing TiMBL tests, the word accuracies are shown in Table 3. It is indicated that TiMBL manages word pronunciation better than Rulsys.

In these experiments there is a distinction between the /e/ and /ê/ phonemes (IPA). This is not an evident distinction to make, since there are many dialects where these two are pronounced in the same way. However, most dialects do make the distinction, which is why they generally count as two phonemes [6]. A great amount of the errors from Rulsys seemed to be that Rulsys suggested /ê/ when the key transcription was /e/ or vice versa. Therefore, in Table 4, the accuracies where the both phonemes were unified are compared to the results in Table 3. Not surprisingly, the accuracy for the Rulsys transcriptions increased more than for the TiMBL tests.

<table>
<thead>
<tr>
<th></th>
<th>Accuracies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GP (Rulsys)</td>
</tr>
<tr>
<td>/e = ê</td>
<td>46.3%</td>
</tr>
<tr>
<td>/e ≠ ê</td>
<td>50.1%</td>
</tr>
</tbody>
</table>

Table 3: Word-based accuracies.

In Swedish, vowel length can alter the meaning of a word and therefore it is important that a Swedish word pronunciation system manages vowel length. Preliminary experiments with TiMBL and Rulsys indicated some difficulties in determining vowel length. Table 5 shows the difference between length-included and length-ignored accuracies.

Table 4: Word-based accuracies.

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Weighting</th>
<th>GP</th>
<th>WO</th>
<th>GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left context</td>
<td>1</td>
<td>0.0387</td>
<td>0.0699</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0706</td>
<td>0.0943</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.1765</td>
<td>0.1882</td>
<td>3</td>
</tr>
<tr>
<td>Focus</td>
<td>4</td>
<td>0.8840</td>
<td>0.8854</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.1974</td>
<td>0.2096</td>
<td>2</td>
</tr>
<tr>
<td>Right context</td>
<td>6</td>
<td>0.0799</td>
<td>0.1094</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.0386</td>
<td>0.0680</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 5: Word-based accuracies.

5. DISCUSSION

5.1. Machine Learning versus Knowledge-Based Word Pronunciation

In the word pronunciation experiments in this study, the machine learning algorithms generally performed better than the
knowledge-based Rulsys. Even the stress/tone-included experiments showed a better performance than Rulsys (Table 3).

When certain error categories were ignored, such as when /e/ and /e/ were mapped onto the same phoneme (Table 4), the Rulsys accuracies improved more than the results from TiMBL. A commercial TTS system with only one phoneme for /e/ and /e/ is in fact a realistic idea, since many users, even those who differ /e/ from /e/ in their own speech, may not notice the generalisation. As shown in Table 5, the vowel length ignorance increased the word accuracy for Rulsys from 46.3% to 59.8%, which was a greater increase than for the tests performed with TiMBL.

With both one phoneme for /e/ and /e/ and ignored vowel length (Table 5), Rulsys came up with a result that better matched the GP in TiMBL. Only with the vowel length excluded, Rulsys performed better than the GP in TiMBL.

However, according to the tests in this study, it is still a fact that TiMBL generally performed better than Rulsys. With vowel length included and /e/ and /e/ unified, the word-based results were 50.1%, 67.3% and 56.6% for Rulsys GP, TiMBL GP and GPS respectively (Table 4).

### 5.2. Algorithm Evaluation

The results from the experiments with different contexts (Table 1) were expected. It is likely that more input information generally leads to higher output accuracy and the tests with 3X3 context did result in higher accuracies than those with 2X2. In word pronunciation, there may be cases where the third graphemes to the left and right of the focus grapheme obviously affect the pronunciation of the focus. With 2X2, these effects will not be detected. One such case is självmordstanke [sjɛlvmaːdːɑŋkaː] (suicidal thought). The experimental set-up in this study would not be able to determine the correct realisation of the <s> (the retroflex [!] without knowledge of the <e> which is the third grapheme to the left.

In many cases IB1 performs better than IGTree [5]. This proved to be the case also in this study. The reason may be that the decision tree structure in IGTree overlooks the relative importance of the features while IB1 stores it during the learning phase. Therefore, when two features have very similar weighting, as for features number five and three (Table 2), IGTree favours the most important of the features, while IB1 still recognises the similarity in importance between the features. Still, in a real TTS situation, IGTree may be preferable since the classification is much faster than IB1 [5].

The reason why MVDM generally performed better than WO is that the former classifies the relative distance between feature-values. With WO, the distance between <e> and <ä> is the same as the distance between e.g. <e> and <p>. However, due to the phonetic resemblance it is a reasonable assumption to see <e> and <ä> as more similar than <e> and <p>. From a sufficiently large training data set, MVDM will find that the two former phonemes more often lead to the same target phoneme than the latter.

In the GPS experiments, large context seemed more important than in GP. As shown in Table 2, the feature weightings in the GP experiments decreased faster than in GPS when the distance to the focus grapheme was increased. The GPS experiments also resulted in a greater difference between 2X2 and 3X3 context. This implies that in stress/tone-included word pronunciation, there is a greater need of extended context information.

### 5.3. Similar Experiments

At the Institute of Language Technology, Dept. of Computational Linguistics, Tilburg University, many machine learning grapheme-to-phoneme experiments have been performed, a few of them also for stress prediction. In e.g. [1], [2], [4], [8] and [10] machine learning experiments using algorithms similar to those in TiMBL are presented.

These experiments treat word pronunciation for e.g. Dutch and English. None of them are for Swedish. There may be a language-dependent difference in the way phonemes can be predicted. This is indicated in [4]. The results from the experiments mentioned above may not be fully comparable with the results from this experiment but they seem to compete well with earlier experiments with similar set-ups. For the GP, the maximum grapheme accuracy in this study was 97.21% (Table 1). The IGTree tests scored 96.93%. In the experiments for Dutch and English referred to above, the accuracies are between 93% and 98%.

The stress prediction experiments mentioned above show much better accuracies than those in this study. However, most of them are for languages with no lexical tone, and therefore use less stress/tone categories than the four in this study.

### 6. CONCLUSION

Both the memory-based learning algorithm IB1 and the decision tree IGTree manage GP. The accuracies compete well with similar machine learning experiments for other languages than Swedish and in this study they exceed the results obtained for the knowledge-based Swedish word pronunciation system in the TTS synthesiser Infovox. An implementation of a machine learning GP module may improve text-to-speech synthesisers that use knowledge-based GP systems.

### 7. FURTHER RESEARCH

Due to the human factor, Swelex, just as many lexicons of that size, is inconsistent and contains noise such as incorrect transcriptions. If the errors were corrected, the results might improve. Removing some of the foreign words from Swelex could also make it easier for TiMBL, since the pronunciation of a grapheme in a given context differ depending on the origin of a word. However, none of these suggestions are easy to perform.

Apart from further modification of the lexicon, there are several ways of improving word pronunciation in a machine learning system. If the foreign words were removed from the lexicon they should not be ignored in constructing a TTS module for Swedish. Instead, they may be used to create foreign-word converters e.g. for words of English origin that are common in Swedish.

For maximum accuracy in stress/tone-included word pronunciation, aligning graphemes to phonemes and stress/tone at the same time is not necessarily the ultimate line of action, since the similarity between e.g. the target classifications ‘a’ and ‘a’ in this study (same phoneme but different lexical tone on a suprasegmental level) is completely overlooked. A cascaded
approach, where the first part manages the word pronunciation and the second part the stress/tone prediction, may increase the accuracies. Just as the word pronunciation experiments performed in this study, part one would be grapheme-based. The stress/tone prediction could be phoneme-based (phonemes mapped to a stress/tone label form the input information) or syllable-based, where the input consists of syllables, and each syllable corresponds to a stress/tone category. A cascaded approach to GPS may also increase the accuracy for vowel length and related phenomena.

8. ACKNOWLEDGEMENTS

This study was conducted during an internship at Infovox, Telia Promotor AB and the authors would like to thank all the members of the Infovox group, especially Bengt Waernulf for giving us the opportunity to perform this study. We would also like to express our gratitude to Harald Berthelsen for his excellent supervision and never-ending patience and Nikolaj Lindberg and Robert Eklund for exhaustive comments down to phoneme level on this paper. Special thanks to Karin Lundahl for encouragement and support.

9. REFERENCES


