Inducing Probabilistic Syllable Classes Using Multivariate Clustering

Karin Müller, Bernd Möbius, and Detlef Prescher
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{karin.mueller|bernd.moebius|detlef.prescher}@ims.uni-stuttgart.de

Abstract

An approach to automatic detection of syllable structure is presented. We demonstrate a novel application of EM-based clustering to multivariate data, exemplified by the induction of 3- and 5-dimensional probabilistic syllable classes. The qualitative evaluation shows that the method yields phonologically meaningful syllable classes. We then propose a novel approach to grapheme-to-phoneme conversion and show that syllable structure represents valuable information for pronunciation systems.

1 Introduction

In this paper we present an approach to unsupervised learning and automatic detection of syllable structure. The primary goal of the paper is to demonstrate the application of EM-based clustering to multivariate data. The suitability of this approach is exemplified by the induction of 3- and 5-dimensional probabilistic syllable classes. A secondary goal is to outline a novel approach to the conversion of graphemes to phonemes (g2p) which uses a context-free grammar (cfg) to generate all sequences of phonemes corresponding to a given orthographic input word and then ranks the hypotheses according to the probabilistic information coded in the syllable classes.

Our approach builds on two resources. The first resource is a cfg for g2p conversion that was constructed manually by a linguistic expert (Müller, 2000). The grammar describes how words are composed of syllables and how syllables consist of parts that are conventionally called onset, nucleus and coda, which in turn are composed of phonemes, and corresponding graphemes. The second resource consists of a multivariate clustering algorithm that is used to reveal syllable structure hidden in unannotated training data. In a first step, we collect syllables by going through a large text corpus, looking up the words and their syllabifications in a pronunciation dictionary and counting the occurrence frequencies of the syllable types. Probabilistic syllable classes are then computed by applying maximum likelihood estimation from incomplete data via the EM algorithm. Two-dimensional EM-based clustering has been applied to tasks in syntax (Rooth et al., 1999), but so far this approach has not been used to derive models of higher dimensionality and, to the best of our knowledge, this is the first time that it is being applied to speech. Accordingly, we have trained 3- and 5-dimensional models for English and German syllable structure.

The obtained models of syllable structure were evaluated in three ways. Firstly, the 3-dimensional models were subjected to a pseudo-disambiguation task, the result of which shows that the onset is the most variable part of the syllable. Secondly, the resulting syllable classes were qualitatively evaluated from a phonological and phonotactic point of view. Thirdly, a 5-dimensional syllable model for German was tested in a g2p conversion task. The results compare well with the best currently available data-driven approaches to g2p conversion (e.g., (Damper et al., 1999)) and suggest that syllable struc-

class 0	0.212	NOP[I]	0.282	I	0.999	NOP[I]	0.460
		t	0.107			n	0.121
		l	0.074			N	0.096
		d	0.071			z	0.079
		b	0.065			t	0.042
		s	0.060			ts	0.013
						f	0.012

Figure 1: Class #0 of a 3-dimensional English model with 12 classes

class 46	0.007	NOP[E]	0.630	E	0.990	nt	0.602	INI	0.627	STR	0.596
		ts	0.256			t	0.128	FIN	0.331	USTR	0.403
		d	0.074			n	0.092	MED	0.040		
		n	0.001			pt	0.010				
						ks	0.004				

Figure 2: Class #46 of a 5-dimensional German model with 50 classes

ture represents valuable information for pronunciation systems. Such systems are critical components in text-to-speech (TTS) conversion systems, and they are also increasingly used to generate pronunciation variants in automatic speech recognition.

The rest of the paper is organized as follows. In Section 2 we introduce the multivariate clustering algorithm. In Section 3 we present four experiments based on 3- and 5-dimensional data for German and English. Section 4 is dedicated to evaluation and in Section 5 we discuss our results.

2 Multivariate Syllable Clustering

EM-based clustering has been derived and applied to syntax (Rooth et al., 1999). Unfortunately, this approach is not applicable to multivariate data with more than two dimensions. However, we consider syllables to consist of at least three dimensions corresponding to parts of the internal syllable structure: onset, nucleus and coda. We have also experimented with 5-dimensional models by adding two more dimensions: position of the syllable in the word and stress status. In our multivariate clustering approach, classes corresponding to syllables are viewed as hidden data in the context of maximum likelihood estimation from incomplete data via the EM algorithm. The two main tasks of EM-based clustering are (i) the induction of a smooth probability model on the data, and (ii) the automatic discovery of class structure in the data. Both aspects are considered in our application. We aim to derive a probability distribution $p(y)$ on syllables y from a large sample. The key idea is to view y as conditioned on an unobserved class $c \in C$, where the classes are given no prior interpretation. The probability of a syllable $y = (y_1, .., y_d) \in \mathcal{Y}_1 \times .. \times \mathcal{Y}_d, d \geq 3$, is defined as:

$$\begin{aligned} p(y) &= \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) p(y|c) \\ &= \sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c) \end{aligned}$$

Note that conditioning of y_i on each other is solely made through the classes c via the independence assumption $p(y|c) = \prod_{i=1}^{d} p(y_i|c)$. This assumption makes clustering feasible in the first place; later on (in Section 4.1) we will experimentally determine the number $|C|$ of classes such that the assumption is optimally met. The EM algorithm (Dempster et al., 1977) is directed at maximizing the incomplete data log-likelihood $L = \sum_y \tilde{p}(y) \ln p(y)$ as a function of the probability distribution p for a given empirical probability distribution $\tilde{p}$. Our application is an instance of the EM-algorithm for context-free models (Baum et al., 1970), from which simple re-estimation formulae can be derived. Let $f(y)$ the frequency of syllable y, and $|f| = \sum_{y \in \mathcal{Y}} f(y)$ the total frequency of the sample (i.e. $\tilde{p}(y) = \frac{f(y)}{|f|}$), and $f_c(y) = f(y) p(c|y)$ the estimated frequency of y annotated with c. Parameter updates $\hat{p}(c), \hat{p}(y_i|c)$ can thus be computed by $(c \in C, y_i \in \mathcal{Y}_i, i = 1, .., d)$:

$$\hat{p}(c) = \frac{\sum_{y \in Y} f_c(y)}{|f|}, \text{ and}$$

class 0 0.071	D 0.745 NOP[@] 0.166	@ 1	NOP[@] 0.877 m 0.0792	ONE 0.999	STR 1
class 1 0.049	NOP[I] 0.914 h 0.071 b 0.010	I 1	n 0.387 z 0.360 t 0.180 f 0.042 ts 0.02	ONE 0.916 INI 0.069	STR 1
class 3 0.040	t 0.206 s 0.106 d 0.104 NOP[I] 0.101 n 0.052	I 0.993	N 0.466 d 0.167 z 0.152 Nz 0.012	FIN 0.997	USTR 0.999
class 4 0.037	t 0.211 v 0.115 D 0.102 d 0.095 NOP[@] 0.072	@ 0.978 O: 0.009	r* 0.597 z 0.115 d 0.057 l 0.054 n 0.045	FIN 0.996 MED 0.003	USTR 0.999
class 10 0.028	S 0.257 m 0.227 d 0.063 t 0.059 NOP[@] 0.007	@ 0.926 I 0.031 I@ 0.015 E 0.005	n 0.388 nt 0.191 nz 0.088 l 0.066 nts 0.049 ns 0.048	FIN 0.999	USTR 0.999
class 14 0.026	m 0.116 p 0.108 k 0.090 g 0.088 t 0.080 pl 0.052 st 0.051	eI 0.426 A: 0.165 E 0.140 O: 0.110	t 0.162 s 0.131 n 0.088 d 0.079 k 0.079 nd 0.052 ts 0.037	ONE 0.696 FIN 0.276	STR 0.984
class 17 0.023	NOP[@] 0.973	@ 1	NOP[@] 0.325 r* 0.317	ONE 0.944 INI 0.050	STR 1

Figure 3: Classes #0, #1, #3, #4, #10, #14, #17 of the 5-dimensional English model

$$\hat{p}(y_i|c) = \frac{\sum_{y \in \mathcal{Y}_1 \times .. \times \mathcal{Y}_{i-1} \times \{y_i\} \times \mathcal{Y}_{i+1} \times .. \times \mathcal{Y}_d} f_c(y)}{\sum_{y \in \mathcal{Y}} f_c(y)}$$

As shown by Baum et al. (1970), every such maximization step increases the log-likelihood function L, and a sequence of re-estimates eventually converges to a (local) maximum.

3 Experiments

A sample of syllables serves as input to the multivariate clustering algorithm. The German data were extracted from the Stuttgarter Zeitung (STZ), a newspaper corpus of about 31 million words. The English data came from the British National Corpus (BNC), a collection of written and spoken language containing about 100 million words. For both languages, syllables were collected by going through the corpus, looking up the words and their syllabifications in a pronunciation dictionary (Baayen et al., 1993)[1] and counting the occurrence frequencies of the syllable types[2].

[1]We slightly modified the English pronunciation lexicon to obtain non-empty nuclei, e.g. /idealism/ [aI][dI@][lIzm,] was modified to [aI][dI@][lI][z@m] (SAMPA transcription).

[2]Subsequent experiments on syllable types (Müller et al., 2000) have shown that frequency counts represent valuable information for our clustering task.

In two experiments, we induced 3-dimensional models based on syllable onset, nucleus, and coda. We collected 9327 distinct German syllables and 13,598 distinct English syllables. The number of syllable classes was systematically varied in iterated training runs and ranged from 1 to 200.

Figure 1 shows a selected segment of class #0 from a 3-dimensional English model with 12 classes. The first column displays the class index 0 and the class probability $p(0)$. The most probable onsets and their probabilities are listed in descending order in the second column, as are nucleus and coda in the third and fourth columns, respectively. Empty onsets and codas were labeled "NOP[nucleus]". Class #0 contains the highly frequent function words *in, is, it, its* as well as the suffixes *-ing ,-ting, -ling*. Notice that these function words and suffixes appear to be separated in the 5-dimensional model (classes #1 and #3 in Figure 3).

In two further experiments, we induced 5-dimensional models, augmented by the additional parameters of position of the syllable in the word and stress status. Syllable position has four values: monosyllabic (ONE), initial (INI), medial (MED), and final (FIN). Stress

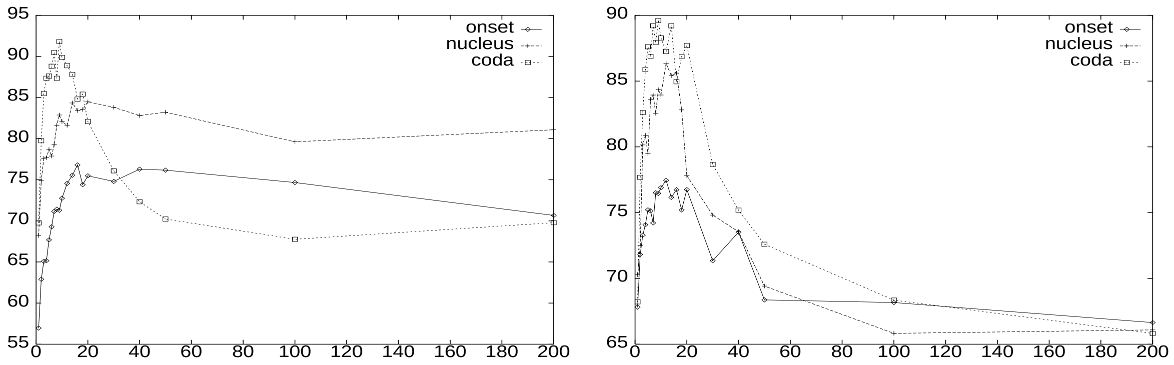

Figure 4: Evaluation on pseudo-disambiguation task for English (left) and German (right)

has two values: stressed (STR) and unstressed (USTR). We collected 16,595 distinct German syllables and 24,365 distinct English syllables. The number of syllable classes ranged from 1 to 200. Figure 2 illustrates (part of) class #46 from a 5-dimensional German model with 50 classes. Syllable position and stress are displayed in the last two columns.

4 Evaluation

In the following sections, (i) the 3-dimensional models are subjected to a pseudo-disambiguation task (4.1); (ii) the syllable classes are qualitatively evaluated (4.2); and (iii) the 5-dimensional syllable model for German is tested in a g2p task (4.3).

4.1 Pseudo-Disambiguation

We evaluated our 3-dimensional clustering models on a pseudo-disambiguation task similar to the one described by Rooth et al. (1999), but specified to onset, nucleus, and coda ambiguity. The first task is to judge which of two onsets on and on' is more likely to appear in the context of a given nucleus n and a given coda cod. For this purpose, we constructed an evaluation corpus of 3000 syllables (on, n, cod) selected from the original data. Then, randomly chosen onsets on' were attached to all syllables in the evaluation corpus, with the resulting syllables (on', n, cod) appearing neither in the training nor in the evaluation corpus. Furthermore, the elements on, n, cod, and on' were required to be part of the training corpus.

Clustering models were parameterized in (up to 10) starting values of EM-training, in the number of classes of the model (up to 200), resulting in a sequence of 10×20 models. Accuracy was calculated as the number of times the model decided $p(on, n, cod) \geq p(on', n, cod)$ for all choices made. Two similar tasks were designed for nucleus and coda.

Results for the best starting values are shown in Figure 4. Models of 12 classes show the highest accuracy rates. For German we reached accuracy rates of 88-90% (nucleus and coda) and 77% (onset). For English we achieved accuracy rates of 92% (coda), 84% (nucleus), and 76% (onset). The results of the pseudo-disambiguation agree with intuition: in both languages (i) the onset is the most variable part of the syllable, as it is easy to find minimal pairs that vary in the onset, (ii) it is easier to predict the coda and nucleus, as their choice is more restricted.

4.2 Qualitative Evaluation

The following discussion is restricted to the 5-dimensional syllable models, as the quality of the output increased when more dimensions were added. We can look at the results from different angles. For instance, we can verify if any of the classes are mainly representatives of a syllable class pertinent to a particular nucleus (as it is the case with the 3-dimensional models). Another interesting aspect is whether there are syllable classes that represent parts of lexical content words, as opposed to high-frequency function words. Finally, some syllable classes may correspond to productive affixes.

class 4 0.032	NOP[aI] 0.624 z 0.163 k 0.043 v 0.029 fR 0.021 m 0.016	aI 1	NOP[aI] 0.689 n 0.303 nst 0.002 ns 0.001	INI 0.755 ONE 0.226	STR 0.999
class 7 0.029	NOP[I] 0.730 z 0.259	I 1	n 0.533 x 0.204 st 0.150 nt 0.067 ns 0.007 m 0.003	ONE 0.867 INI 0.128	STR 0.915 USTR 0.084
class 26 0.017	f 0.573 NOP[E] 0.351 ts 0.009 h 0.006	E 0.987 o: 0.007 O 0.001	R 0.983	INI 0.906 MED 0.093	USTR 0.994
class 34 0.011	l 0.408 t 0.175 d 0.133	I 0.905	x 0.690 xt 0.108 k 0.047	FIN 0.936 MED 0.063	USTR 0.999
class 40 0.009	b 0.144 R 0.128 t 0.119 v 0.095 ts 0.090 gl 0.022	aI 0.999	NOP[aI] 0.706 n 0.103 x 0.077 ts 0.057 s 0.016 l 0.015	MED 0.876 FIN 0.119	USTR 0.596 STR 0.403

Figure 5: Classes #4, #7, #26, #34, #40 of the 5-dimensional German model

German. The majority of syllable classes obtained for German is dominated by one particular nucleus per syllable class. In 24 out of 50 classes the probability of the dominant nucleus is greater than 99%, and in 9 cases it is indeed 100%. The only syllable nuclei that do not dominate any class are the front rounded vowels /y:, Y, 2:, 9/, the front vowel /E:/ and the diphthong /OY/, all of which are among the least frequently occurring nuclei in the lexicon of German. Figure 5 depicts the classes that will be discussed now.

Almost one third (28%) of the 50 classes are representatives of high-frequency function words. For example, class #7 is dominated by the function words *in, ich, ist, im, sind, sich*, all of which contain the short vowel /I/.

Another 32% of the 50 classes represents syllables that are most likely to occur in initial, medial and final positions in the open word classes of the lexicon, i.e. nouns, adjectives, and verbs. Class #4 covers several lexical entries involving the diphthong /aI/ mostly in stressed word-initial syllables. Class #40 provides complimentary information, as it also includes syllables containing /aI/, but here mostly in word-medial position.

We also observe syllable classes that represent productive prefixes (e.g., *ver-, er-, zer-, vor-, her-* in class #26) and suffixes (e.g., *-lich, -ig* in class #34). Finally, there are two syllable classes (not displayed) that cover the most common inflectional suffixes involving the vowel /@/ (schwa).

Class numbers are informative insofar as the classes are ranked by decreasing probability. Lower-ranked classes tend (i) not to be dominated by one nucleus; (ii) to contain vowels with relatively low frequency of occurrence; and (iii) to yield less clear patterns in terms of word class or stress or position. For illustration, class #46 (Figure 2) represents the syllable *ent* [Ent], both as a prefix (INI) and as a suffix (FIN), the former being unstressed (as in *Entwurf* "design") and the latter stressed (as in *Dirigent* "conductor").

English. In 24 out of the 50 syllable classes obtained for English one dominant nucleus per syllable class is observed. In all of these cases the probability of the nucleus is larger than 99% and in 7 classes the nucleus probability is 100%. Besides several diphthongs only the relatively infrequent vowels /V/, /A:/ and /3:/ do not dominate any class. Figure 3 shows the classes that are described as follows.

High-frequency function words are represented by 10 syllable classes. For example, class #0 and #17 and are dominated by the determiners *the* and *a*, respectively, and class #1 contains function words that involve the short vowel /I/, such as *in, is, it, his, if, its*.

Productive word-forming suffixes are found in class #3 (*-ing*), and common inflectional suffixes in class #4 (*-er, -es, -ed*). Class #10

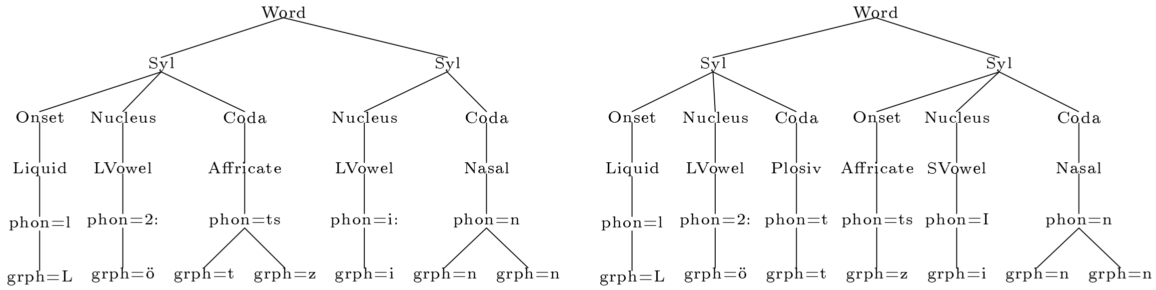

Figure 6: An incorrect (left) and a correct (right) cfg analysis of *Lötzinn*

is particularly interesting in that it represents a comparably large number of common suffixes, such as *-tion, -ment, -al, -ant, -ent, -ence* and others.

The majority of syllable classes, viz. 31 out of 50, contains syllables that are likely to be found in initial, medial and final positions in the open word classes of the lexicon. For example, class #14 represents mostly stressed syllables involving the vowels /eI, A:, e:, O:/ and others, in a variety of syllable positions in nouns, adjectives or verbs.

4.3 Evaluation by g2p Conversion

In this section, we present a novel method of g2p conversion (i) using a cfg to produce all possible phonemic correspondences of a given grapheme string, (ii) applying a probabilistic syllable model to rank the pronunciation hypotheses, and (iii) predicting pronunciation by choosing the most probable analysis. We used a cfg for generating transcriptions, because grammars are expressive and writing grammar-rules is easy and intuitive. Our grammar describes how words are composed of syllables and syllables branch into onset, nucleus and coda. These syllable parts are re-written by the grammar as sequences of natural phone classes, e.g. stops, fricatives, nasals, liquids, as well as long and short vowels, and diphthongs. The phone classes are then re-interpreted as the individual phonemes that they are made up of. Finally, for each phoneme all possible graphemic correspondences are listed.

Figure 6 illustrates two analyses (out of 100) of the German word *Lötzinn (tin solder)*. The phoneme strings (represented by non-terminals named "phon=...") and the syllable boundaries (represented by the non-terminal "Syl") can be extracted from these analyses. Figure 6 depicts both an incorrect analysis [l2:ts][i:n] and its correct counterpart [l2:t][tsIn]. The next step is to rank these transcriptions by assigning probabilities to them. The key idea is to take the product of the syllable probabilities. Using the 5-dimensional[3] German syllable model yields a probability of $7.5 \cdot 10^{-7} \cdot 3.1 \cdot 10^{-7} = 2.3 \cdot 10^{-13}$ for the incorrect analysis and a probability of $1.5 \cdot 10^{-7} \cdot 6.5 \cdot 10^{-6} = 9.8 \cdot 10^{-13}$ for the correct one. Thus we achieve the desired result of assigning the higher probability to the correct transcription.

We evaluated our g2p system on a test set of 1835 unseen words. The ambiguity expressed as the average number of analyses per word was 289. The test set was constructed by collecting 295,102 words from the German Celex dictionary (Baayen et al., 1993) that were *not seen* in the STZ corpus. From this set we manually eliminated (i) foreign words, (ii) acronyms, (iii) proper names, (iv) verbs, and (v) words with more than three syllables. The resulting test set is available on the World Wide Web[4].

Figure 7 shows the performance of four g2p systems. The second and fourth columns show the accuracy of two baseline systems: g2p conversion using the 3- and 5-dimensional empirical distributions (Section 2), respectively. The third and fifth columns show the word

[3]Position can be derived from the cfg analyses, stress placement is controlled by the most likely distribution.

[4]http://www.ims.uni-stuttgart.de/phonetik/g2p/

g2p system	3-dim baseline	3-dim classes	5-dim baseline	5-dim classes
word accuracy	66.8 %	67.4 %	72.5 %	75.3 %

Figure 7: Evaluation of g2p systems using probabilistic syllable models

accuracy of two g2p systems using 3- and 5-dimensional syllable models, respectively.

The g2p system using 5-dimensional syllable models achieved the highest performance (75.3%), which is a gain of 3% over the performance of the 5-dimensional baseline system and a gain of 8% over the performance of the 3-dimensional models[5].

5 Discussion

We have presented an approach to unsupervised learning and automatic detection of syllable structure, using EM-based multivariate clustering. The method yields phonologically meaningful syllable classes. These classes are shown to represent valuable input information in a g2p conversion task.

In contrast to the application of two-dimensional EM-based clustering to syntax (Rooth et al., 1999), where semantic relations were revealed between verbs and objects, the syllable models cannot *a priori* be expected to yield similarly meaningful properties. This is because the syllable constituents (or phones) represent an inventory with a small number of units which can be combined to form meaningful larger units, viz. morphemes and words, but which do not themselves carry meaning. Thus, there is no reason why certain syllable types should occur significantly more often than others, except for the fact that certain morphemes and words have a higher frequency count than others in a given text corpus. As discussed in Section 4.2, however, we do find some interesting properties of syllable classes, some of which apparently represent high-frequency function words and productive affixes, while others are typically found in lexical content words. Subjected to a pseudo-disambiguation task (Section 4.1), the 3-dimensional models confirm the intuition that the onset is the most variable part of the syllable.

In a feasibility study we applied the 5-dimensional syllable model obtained for German to a g2p conversion task. Automatic conversion of a string of characters, i.e. a word, into a string of phonemes, i.e. its pronunciation, is essential for applications such as speech synthesis from unrestricted text input, which can be expected to contain words that are not in the system's pronunciation dictionary or otherwise unknown to the system. The main purpose of the feasibility study was to demonstrate the relevance of the phonological information on syllable structure for g2p conversion. Therefore, information and probabilities derived from an alignment of grapheme and phoneme strings, i.e. the lowest two levels in the trees displayed in Figure 6, was deliberately ignored. Data-driven pronunciation systems usually rely on training data that include an alignment of graphemes and phonemes. Damper et al. (1999) have shown that the use of unaligned training data significantly reduces the performance of g2p systems. In our experiment, with training on unannotated text corpora and without an alignment of graphemes and phonemes, we obtained a word accuracy rate of 75.3% for the 5-dimensional German syllable model.

Comparison of this performance with other systems is difficult: (i) hardly any quantitative g2p performance data are available for German; (ii) comparisons across languages are hard to interpret; (iii) comparisons across different approaches require cautious interpretations. The most direct point of comparison is the method presented by Müller (2000). In one of her experiments, the standard probability model was applied to the hand-crafted cfg presented in this paper, yielding 42% word

[5] 45 resp. 95 words could not be disambiguated by the 3- resp. 5-dimensional empirical distributions. The reported relatively small gains can be explained by the fact that our syllable models were applied only to this small number of ambiguous words.

accuracy as evaluated on our test set. Running the test set through the pronunciation rule system of the IMS German Festival TTS system (Möhler, 1999) resulted in 55% word accuracy. The Bell Labs German TTS system (Möbius, 1999) performed at better than 94% word accuracy on our test set. This TTS system relies on an annotation of morphological structure for the words in its lexicon and it performs a morphological analysis of unknown words (Möbius, 1998); the pronunciation rules draw on this structural information. These comparative results emphasize the value of phonotactic knowledge and information on syllable structure and morphological structure for g2p conversion.

In a comparison across languages, a word accuracy rate of 75.3% for our 5-dimensional German syllable model is slightly higher than the best data-driven method for English with 72% (Damper et al., 1999). Recently, Bouma (2000) has reported a word accuracy of 92.6% for Dutch, using a 'lazy' training strategy on data aligned with the correct phoneme string, and a hand-crafted system that relied on a large set of rule templates and a many-to-one mapping of characters to graphemes preceding the actual g2p conversion.

We are confident that a judicious combination of phonological information of the type employed in our feasibility study with standard techniques such as g2p alignment of training data will produce a pronunciation system with a word accuracy that matches the one reported by Bouma (2000). We believe, however, that for an optimally performing system as is desired for TTS, an even more complex design will have to be adopted. In many languages, including English, German and Dutch, access to morphological and phonological information is required to reliably predict the pronunciation of words; this view is further evidenced by the performance of the Bell Labs system, which relies on precisely this type of information. We agree with Sproat (1998, p. 77) that it is unrealistic to expect optimal results from a system that has no access to this type of information or is trained on data that are insufficient for the task.

References

Harald R. Baayen, Richard Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database—Dutch, English, German. (Release 1)[CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, Univ. Pennsylvania.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Math. Statistics*, 41(1):164–171.

Gosse Bouma. 2000. A finite state and data-oriented method for grapheme to phoneme conversion. In *Proc. 1st Conf. North American Chapter of the ACL (NAACL)*, Seattle, WA.

Robert I. Damper, Y. Marchand, M. J. Adamson, and Kjell Gustafson. 1999. Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech and Language*, 13:155–176.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statistical Soc.*, 39(B):1–38.

Bernd Möbius. 1998. Word and syllable models for German text-to-speech synthesis. In *Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves)*, pages 59–64.

Bernd Möbius. 1999. The Bell Labs German text-to-speech system. *Computer Speech and Language*, 13:319–358.

Gregor Möhler. 1999. IMS Festival. [http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html].

Karin Müller, Bernd Möbius, and Detlef Prescher. 2000. Inducing probabilistic syllable classes using multivariate clustering - GOLD. In *AIMS Report 6(2)*, IMS, Univ. Stuttgart.

Karin Müller. 2000. PCFGs for syllabification and g2p conversion. In *AIMS Report 6(2)*, IMS, Univ. Stuttgart.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proc. 37th Ann. Meeting of the ACL*, College Park, MD.

Richard Sproat, editor. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic, Dordrecht.