Developing Consistent Pronunciation Models for Phonemic Variants

Marelie Davel and Etienne Barnard

Human Language Technologies Research Group
CSIR Meraka Institute / University of Pretoria, Pretoria,0001
mdavel@csir.co.za, ebarnard@up.ac.za

Abstract

Pronunciation lexicons often contain pronunciation variants. This can create two problems: It can be difficult to define these variants in an internally consistent way and it can also be difficult to extract generalised grapheme-to-phoneme rule sets from a lexicon containing variants. In this paper we address both these issues by creating 'pseudo-phonemes' associated with sets of 'generation restriction rules' to model those pronunciations that are consistently realised as two or more variants. By pre-processing and post-processing the lexicon appropriately, grapheme-to-phoneme algorithms that were not able to deal with pronunciation variants previously can now be extended to incorporate variants easily, without requiring changes to the standard algorithms. We evaluate the effectiveness of this approach using the Default&Refine rule extraction algorithm, and apply the method to both the English Oxford Advanced Learners Dictionary (OALD) and the Flemish FONILEX pronunciation lexicon. We find that the approach generalises to different languages, is able to model phonemic variation accurately and is able to identify inconsistent variants in pre-existing lexicons.

Index Terms: pronunciation modelling, pronunciation variants, grapheme-to-phoneme rules, pseudo-phonemes.

1. Introduction

Pronunciation lexicons often contain pronunciation variants: words with the same orthography that are realised as different pronunciations in different contexts. These variants can occur in a continuum ranging from generally accepted alternate word pronunciations to pronunciation variants that only occur in limited circumstances: in effect ranging from true homonyms to dialect and accent variants, to phonological variants based on a variety of factors such as speaker and/or speaking style. It can be difficult to decide which of these variants to model, especially if different levels of variation are to be kept distinct. While phonological phenomena (such as /r/-deletion, schwa-deletion or schwa-insertion) can be modelled as predictive rewrite rules, phonemic variation is most often included in pronunciation lexicons as explicit alternate pronunciations. Including explicit alternate pronunciations in pronunciation lexicons introduces two challenges:

1. It is often difficult to include variants in a consistent way. When a lexicon grows through general usage (for example, the evolution of the CMU pronunciation dictionary[1]) it is easy to include one example of a variant as required for a specific application, without including the entire variant family. For example, if (using ARPABET) both the pronunciations */iy n k r iy s/* and */iy ng k r iy s/* are allowed for the verb 'increase', then similarly, both */iy n k r iy s t/* and */iy ng k r iy s t/* should be allowed for the word 'increased' in order for the dictionary to remain internally consistent.

2. A variety of techniques are available for the extraction of grapheme-to-phoneme prediction rules from pre-existing lexicons, including decision trees [2], pronunciation-by-analogy models [3] and instance-based learning algorithms [4, 5]. Unfortunately, many of these techniques, including Dynamically Expanding Context (DEC) [4] and Default&Refine [6], experience difficulty in accommodating alternate pronunciations during the machine learning of grapheme-to-phoneme prediction rules. For such techniques, the lexicon is typically pre-processed and pronunciation variants removed prior to rule extraction. The latter is a drawback when developing pronunciation models through bootstrapping, a useful technique for the accelerated development of pronunciation lexicons [7, 8]. Bootstrapping systems utilise automated techniques to extract grapheme-to-phoneme prediction rules from an existing lexicon and apply these rules to predict additional entries, typically in an iterative fashion. These systems can have difficulty incorporating learning from pronunciation variants.

In this paper we address both of the above issues. In prior work [9] we explored the incorporation of variants in a standard grapheme-to-phoneme rule extraction algorithm through the generation of a 'pseudo-phoneme' and an associated set of 'generation restriction rules' to model alternate phonemic pronunciations, and reported initial results obtained modelling the phonemic variation in the Oxford Advanced Learners Dictionary (OALD) [10]. In this paper we define the pseudo-phoneme model approach in further detail, verify the effectiveness of the method using both OALD and FONILEX [11], a large Flemish dictionary that includes significantly more phonemic variation than OALD, and investigate the applicability of the approach to identify inconsistent variants in a pre-existing lexicon. We focus on one typical instance-based learning algorithm, Default&Refine [6], but discuss how this approach generalises to other grapheme-to-phoneme frameworks.

2. Background: Default&Refine

The Default&Refine algorithm is an instance-based learning algorithm that can be used to extract a set of grapheme-to-phoneme prediction rules from an existing pronunciation lexicon. It is highly competitive in terms of both learning efficiency (that is, the accuracy achieved with a limited number of training examples) and asymptotic accuracy when compared to alternative approaches [6].

The Default&Refine framework is similar to that of most multi-level rewrite rule sets. Each grapheme-to-phoneme rule con-

sists of a pattern:

$$(left\ context - g - right\ context) \rightarrow p \quad (1)$$

where g indicate the grapheme being predicted, $left\ context$ and $right\ context$ indicate the left and right graphemic context respectively, and p the phonemic realisation of g. Rules are ordered explicitly. The pronunciation for a word is generated one grapheme at a time: each grapheme and its left and right context as found in the target word are compared with each rule in the ordered rule set, and the first matching rule is applied.

During rule extraction, iterative Viterbi alignment is used to obtain grapheme-to-phoneme mappings, after which a hierarchy of rewrite rules is extracted per grapheme. The rule set is extracted in a straightforward fashion: for every letter (grapheme), a default phoneme is derived as the phoneme to which the letter is most likely to map. 'Exceptional' cases – words for which the expected phoneme is not correct – are handled as refinements. The smallest possible context of letters that can be associated with the next most frequently occurring phoneme is extracted as a refined rule. (The rule that describes the largest number of current exceptions accurately is selected next.) Exceptions to this refined rule are similarly represented by further refinements, and so forth, leading to a cascading rule set that describes the training set with complete accuracy. Further details can be found in [6].

3. Modelling phonemic variation

3.1. Approach

Our approach to the modelling of explicit pronunciation variants utilises two concepts that we refer to as *pseudo-phonemes* and *generation restriction rules*, respectively. These are discussed in the remainder of this section.

A pseudo-phoneme is used to model two or more phonemes that consistently occur as variant pronunciations of the same word. In practise, we use the following process: we align the training lexicon, extract all the words giving rise to pronunciation variants from the aligned lexicon, and analyse these words one grapheme at a time. Since the word-pronunciation pairs have already been aligned, there is a one-to-one mapping between each grapheme and its associated phoneme. For each word, we consider any grapheme that can be realised as two or more phonemes and map this set of phonemes to a new single pseudo-phoneme. If a set of phonemes has been seen before, the existing pseudo-phoneme – already associated with this set – is used. Table 1 lists examples of pseudo-phonemes generated from the *OALD* corpus. Phonemes are displayed simplified to the closest ARPABET[12] symbol. The 'ϕ' symbol indicates phonemic nulls (inserted during alignment). Once all pseudo-phonemes have been defined, the aligned training lexicon is regenerated in terms of the new phoneme set.

The generation restriction rules are used to restrict the number of possible variants generated when two or more pseudo-phonemes occur in a single word. For example the word 'second' can be realised as two variants */s eh k ih n d/* and */s ih k aa n d/*. According to the pseudo-phoneme generation process described above, these two variants will be combined as a single pronunciation: */s* p_3 *k* p_4 *n d/*. However, this new pronunciation implies four different variants, of which */s ih k ih n d/* and */s eh k aa n d/* are not included in the training lexicon. The generation restriction rules are used to identify and limit the expansion options for such cases, to ensure that the newly generated training lexicon encodes exactly the same information as the initial training lexicon.

Table 1: *Examples of pseudo-phonemes generated from the OALD corpus*

Word	Variants	Pseudo-phoneme	New pronunciation
animate	ae n ih m ay t ϕ ae n ih m ax t ϕ	p_1=ay||ax	ae n ih m p_1 t ϕ
delegate	d eh l ih g ay t ϕ d eh l ih g ax t ϕ	p_1=ay||ax	d eh l ih g p_1 t ϕ
lens	l eh n z l eh n s	p_2=s||z	l eh n p_2
close	k l ow z ϕ k l ow s ϕ	p_2=s||z	k l ow p_2 ϕ

In practice, all words that contain two or more pseudo-phonemes are extracted from the training lexicon and the pseudo-phoneme combinations analysed. If a pseudo-phoneme combination (such as p_3-p_4 above) is realised as one or more specific phoneme combinations (*/eh-ih/* or */ih-aa/*) for all words in the training lexicon, the p_3-p_4 combination will always be expanded as these two phoneme combinations, and these only. If a specific phoneme combination exists for some words in the training lexicon and not for others, more complex generation restriction rules are required. Fortunately the Default&Refine algorithm is well suited to extracting such rules from the pseudo-phoneme combination information. The smallest possible rule is extracted to indicate the context in which a pseudo-phoneme combination is realised as one phoneme combination or another. For example, the extracted rule '$-p_3-, -p_4- : eh\_ih, ih\_aa'$ specifies that whenever the two pseudo-phonemes p_3 and p_4 occur together in a word, in any graphemic context, only two variants are allowed, namely expanding the pseudo-phonemes to */eh-ih/* and */ih-aa/*, and these combinations only. A more complicated rule, specifying that this should only occur if p_4 is followed by an 'n' would be written as '$-p_3-, -p_4 - n : eh\_ih, ih\_aa'$. Luckily the default rule is typically sufficient; more complex rules are seldom required.

The new rule extraction process consists of the following steps: We align the original training lexicon, generate a set of pseudo-phonemes and rewrite the aligned lexicon in terms of the new pseudo-phonemes. Next, we extract Default&Refine rules for the rewritten lexicon, and extract generation restriction rules based on the original lexicon (in comparison with the rewritten lexicon). We then use these two rule sets to predict the pronunciation of the test word lists: standard Default&Refine prediction is used to generate a test lexicon specified in terms of pseudo-phonemes, and the pseudo-phonemes are expanded to regular phonemes according to the generation restriction rules, resulting in the final test lexicon.

3.2. Evaluation and Results

In order to evaluate whether the proposed approach is practical and generalises to different languages, we model the pronunciation variants occurring in two different lexicons:

1. The Oxford Advanced Learners Dictionary (OALD) [10] is a publicly available English pronunciation lexicon that includes pronunciation variants. We use the exact 60,399 word version of the lexicon as used by Black *et al* [2]. For this set of experiments we do not utilise the part-of-speech tags and predict pronunciations without stress assignment.
2. FONILEX, a publicly available pronunciation dictionary of

Dutch words as spoken in the Flemish part of Belgium[11]. We use the exact 173,873-word pre-aligned version of the dictionary as used by Hoste *et al* [13].

Statistics with regard to the phonemic variation occurring in these two lexicons are provided in Table 2.

Table 2: *Phonemic variation in OALD and FONILEX*

	OALD	FONILEX
Number of pronunciations	60,399	173,873
Number of unique words	59,696	166,786
Remaining words if variants removed	59,001	160,284
Number of words with variants	695	6,502
Average pronunciations per variant	2.01	2.09
Variant words as % of unique words	1.16%	3.90%

In all experiments we perform 10-fold cross-validation, based on a 90% training and 10% test set. We report on phoneme correctness (the number of phonemes identified correctly), phoneme accuracy (number of correct phonemes minus number of insertions, divided by the total number of phonemes in the correct pronunciation) and word accuracy (number of words completely correct). We also report on the standard deviation of the mean of each of these measurements, indicated by σ_{10}. (If the mean of a random variable is estimated from n independent measurements, and the standard deviation of those measurements is σ, the standard deviation of the mean is $\sigma_n = \frac{\sigma}{\sqrt{n}}$.)

3.2.1. Benchmark systems

In previous experiments in which Default&Refine was applied to the OALD corpus [14], the first version of each pronunciation variant was kept and other variants deleted prior to rule extraction. Results for this approach are listed in Table 3 as '*1 var*'. Before applying the new approach, we evaluate the effect on predictive accuracy if all variants are simply removed from the training lexicon (as this is what in effect happens when variants are modelled separately using the pseudo-phoneme approach), and list the results in Table 3 as '*no var*'. As can be seen, results are comparable, with the variant-containing scores consistently somewhat lower because of the extra complexity introduced by variants. Comparable results are listed for the FONILEX corpus, retaining one variant during training. During testing, results are slightly different if, for test words that have more than one variant, the first variant is consistently used (*1 var first*), or any variant is selected at random (*1 var random*). These systems are used as benchmarks to evaluate the effect of the new approach to variant modelling on the accuracy with which both variants and non-variants can be predicted. The accuracy of the Default&Refine benchmark systems are high, as can be seen by comparing with other results obtained in literature, specifically using decision trees (*dtrees*) [2] and IB1-IG (*IB1-IG*) [13].

3.2.2. Prediction of non-variants

First, we consider whether the additional modelling of the variants may have a detrimental effect on the prediction of non-variants. Using both the generated lexicon and the reference lexicon, we generate a list of all variants in the test set. We remove these words from the test word list, and compare the accuracy of the best baseline systems (*OALD no var*, *FONILEX 1 var*) with that of the pseudo-phoneme systems (*pseudo*), when measured using the reduced test set. Results are listed in Table 4. We see that the pseudo-phoneme modelling approach does not negatively influence the accuracy with which non-variants are predicted.

Table 3: *Predictive accuracy of different systems.*

Approach	Word accuracy	σ_{10}	Phoneme accuracy	σ_{10}	Phoneme correct	σ_{10}
OALD						
dtrees [2]	76.92	-	-	-	96.36	-
1 var	86.46	0.15	97.41	0.03	97.67	0.03
no var	86.87	0.16	97.49	0.03	97.74	0.03
FONILEX						
IB1-IG [13]	86.37	-	-	-	98.18	-
1 var random	92.03	0.06	98.78	0.04	98.87	0.01
1 var first	95.64	0.05	98.36	0.01	99.43	0.01

Table 4: *The pseudo-phoneme approach does not have a detrimental effect on the accuracy with which non-variants are predicted. (Tested on test set without variants.)*

Approach	Word accuracy	σ_{10}	Phoneme accuracy	σ_{10}	Phoneme correct	σ_{10}
OALD						
no var	86.93	0.16	97.50	0.03	97.75	0.03
pseudo	86.92	0.15	97.50	0.03	97.76	0.03
FONILEX						
1 var	95.54	0.06	99.35	0.04	99.42	0.01
pseudo	95.54	0.08	99.33	0.03	99.41	0.01

3.2.3. Prediction of variants

Given the modelling process, it is clear that the original training lexicon and the training lexicon rewritten using pseudo-phonemes are equivalent. (This can be verified by expanding the rewritten training lexicon with the same process used to expand the test lexicon, and comparing the expanded lexicon with the original version.) The pseudo-phoneme approach therefore provides a technique to encode pronunciation variants within the Default&Refine framework without requiring any changes to the standard algorithm. While this in itself is a useful capability, we are more interested in the effectiveness with which the approach is able to generalise from variants in the training data. In order to evaluate the above, we count the number of variants occurring in the reference lexicon and the generated test lexicon according to the number of variants *correctly* identified in the test lexicon, the number of variants *missing* from the test lexicon, and the number of *extra* variants occurring in the test lexicon, but not in the reference lexicon.

On average we find that, for OALD, 58% of expected variants are correctly generated and that 67% of generated variants are correct. For FONILEX, 41% of expected variants are correctly generated and 83% of generated variants are correct. In Table 5 we list the detailed results for three example cross-validation sets per lexicon. These results indicate that the pseudo-phoneme approach indeed generalises from the training data and can generate a significant percentage of the variants occurring in the reference lexicon.

Table 5: *Correct, missing and extra variants generated during cross-validation. The percentage of expected variants that were correctly generated, and percentage of generated variants that were correct are also displayed.*

Correct	Missing	Extra	% correct of expected	% correct of generated
OALD				
58	43	23	57.43	71.60
56	40	20	58.33	73.68
53	34	28	60.92	65.43
FONILEX				
1214	1639	277	42.55	81.42
1117	1674	240	40.02	82.31
1145	1609	219	41.58	83.94

4. Verifying the consistency of variants

When the variants classified as *'extra'* in the above experiment are analysed, it soon becomes clear that some of the generated variants may be legitimate variants that have simply not been included in the original lexicon. For example, *OALD* contains the two pronunciations */r iy p ae t r ia t s /* and */r iy p ae t r iy ey t s/* as variants of the word 'repatriates', but allows only the single pronunciation */r iy p ae t r ia t/* as a pronunciation of the word 'repatriate'. When the prediction system generates the alternative pronunciation */r iy p ae t r iy ey t /*, it is flagged as erroneous. These two pronunciations are close to each other, and will not necessarily affect the quality of a speech recognition or text-to-speech system developed using these pronunciations. However, inconsistencies in the pronunciation lexicon lead to unnecessarily complex pronunciation models, and consequently, suboptimal generalisation.

In order to evaluate the consistency of the OALD lexicon, we create a list of all variants flagged as *extra* during the 10 cross-validations, and have this list evaluated by a linguist. We find that 249 words generate *extra* variants (498 additional variants were generated in total). Of the 498 pronunciations, 251 were valid pronunciations according to the OALD lexicon. However, of the remaining 247 pronunciations, 84 were identified as valid by the linguist, that is 34% of the variants classified as *extra* may indeed be valid pronunciations, simply not included in the lexicon. The variants generated by the pseudo-phoneme approach therefore provides a good candidate list when verifying the consistency of an existing lexicon. This process can be repeated a number of times (each time including the new variants in the training set) to identify additional variants that may be valid.

5. Conclusions

In this paper we described a process that allows for the incorporation of explicit phonemic variants in the Default&Refine algorithm. This is done in a way that requires no adjustments to the standard algorithm, but rather utilises pre- and post-processing of the training data and testing data. As the data is re-configured to a format expected by the standard algorithm, the same approach can be used for other grapheme-to-phoneme learning algorithms such as Dynamically Expanding Context (DEC).

Evaluated on both the *OALD* and the *FONILEX* corpus, we find that the incorporation of variants does not have a detrimental effect on the accuracy with which non-variants can be predicted. In addition, the proposed approach is able to describe all variants occurring in the training set and identify a significant percentage of variants occurring in the test set (58% in the case of OALD, 41% in the case of FONILEX). Of the variants generated 67% were correct in the case of OALD, and 83% correct in the case of FONILEX. These results do not take into account that some of the variants identified as incorrect may be legal variants not included in the version of the lexicons used here.

Utilising the list of *'extra'* variants as a candidate list for potential missing variants, the dictionary can be evaluated by a linguist to determine consistency. In the case of OALD, 34% of variants on the candidate list were deemed legal by a linguist, but missed by the lexicon. This therefore provides a useful tool for the verification of the consistency of phonemic variation in existing lexicons.

6. References

[1] "The CMU pronunciation dictionary," 1998, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[2] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, November 1998, pp. 77–80.

[3] F. Yvon, "Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks," in *Proceedings of NeMLaP*, Ankara, Turkey, 1996, pp. 218–228.

[4] K. Torkkola, "An efficient way to learn English grapheme-to-phoneme rules automatically," in *Proceedings of ICASSP*, Minneapolis, USA, April 1993, vol. 2, pp. 199–202.

[5] W. Daelemans, A. van den Bosch, and J. Zavrel, "Forgetting exceptions is harmful in language learning," *Machine Learning*, vol. 34, no. 1-3, pp. 11–41, 1999.

[6] M. Davel and E.Barnard, "A default-and-refinement approach to pronunciation prediction," in *Proceedings of PRASA*, South Africa, November 2004, pp. 119–123.

[7] M. Davel and E. Barnard, "Bootstrapping for language resource generation," in *Proceedings of PRASA*, South Africa, November 2003, pp. 97–100.

[8] S. Maskey, L. Tomokiyo, and A.Black, "Bootstrapping phonetic lexicons for new languages," in *Proceedings of Interspeech*, Jeju, Korea, October 2004, pp. 69–72.

[9] M. Davel and E.Barnard, "Extracting pronunciation rules for phonemic variants," in *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing*, Stellenbosch, South Africa, April 2006.

[10] R. Mitten, "Computer-usable version of Oxford Advanced Learner's Dictionary of Current English," Tech. Rep., Oxford Text Archive, 1992.

[11] P. Mertens and F. Vercammen, "Fonilex manual," Tech. Rep., K.U.Leuven CCL, 1998.

[12] J. S. Garofolo, Lori F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus, NIST order number PB91-100354," February 1993.

[13] V. Hoste, W. Daelemans, E.T.K. Sang, and S. Gillis, "Meta-learning for phonemic annotation of corpora," in *Proceedings of ICML-2000*, Stanford University, USA, 2000.

[14] M. Davel, *Pronunciation modelling and bootstrapping*, Ph.D. thesis, University of Pretoria, 2005.