Extracting pronunciation rules for phonemic variants

Marelie Davel and Etienne Barnard

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

Abstract

Various automated techniques can be used to generalise from phonemic lexicons through the extraction of grapheme-to-phoneme rule sets. These techniques are particularly useful when developing pronunciation models for previously unmodelled languages: a frequent requirement when developing multilingual speech processing systems. However, many of the learning algorithms (such as Dynamically Expanding Context or Default&Refine) experience difficulty in accommodating alternate pronunciations that occur in the training lexicon.

In this paper we propose an approach for the incorporation of phonemic variants in a typical instance-based learning algorithm, Default&Refine. We investigate the use of a combined 'pseudo-phoneme' associated with a set of 'generation restriction rules' to model those phonemes that are consistently realised as two or more variants in the training lexicon.

We evaluate the effectiveness of this approach using the Oxford Advanced Learners Dictionary, a publicly available English pronunciation lexicon. We find that phonemic variation exhibits sufficient regularity to be modelled through extracted rules, and that acceptable variants may be underrepresented in the studied lexicon. The proposed method is applicable to many approaches besides the Default&Refine algorithm, and provides a simple but effective technique for including phonemic variants in grapheme-to-phoneme rule extraction frameworks.

1. Introduction

The growing trend towards multilingual speech systems implies an increasing requirement for linguistic resources in additional languages. A basic linguistic resource typically required when developing a speech processing system is the pronunciation model: a mechanism for providing a language-specific mapping of the orthography of a word to its phonemic realisation.

The process of creating a pronunciation model in a new language can be accelerated through bootstrapping [1, 2]. 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. The pronunciation model is further enhanced by extracting grapheme-to-phoneme rules from the final lexicon, in order to deal with out of vocabulary words.

A variety of techniques are available for the extraction of grapheme-to-phoneme prediction rules from pre-existing lexicons, including decision trees [3], pronunciation-by-analogy models [4] and instance-based learning algorithms [5, 6]. Unfortunately, some of these techniques, including Dynamically Expanding Context(DEC) [5] and Default&Refine [7], 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.

Pronunciation 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 for a previously unmodelled language, 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. It is these explicit alternate pronunciations that are currently not modelled effectively by many of the techniques used to generalise from an existing lexicon.

In this paper we investigate the incorporation of explicit phonemic variants in a typical instance-based learning algorithm, Default&Refine [7], by generating a combined 'pseudo-phoneme' and an associated set of 'generation restriction rules' to model alternate phonemic pronunciations. The remainder of this paper is structured as follows: in Section 2 we provide background on the Default&Refine rule extraction algorithm used, in Section 3 we describe our approach to the modelling of phonemic variation, and in Section 4 we evaluate the effectiveness of the method when applied to the Oxford Advanced

Learners Dictionary (OALD) [8], a publicly available English pronunciation lexicon that includes pronunciation variants. In the concluding section we discuss the implications of our results and future work.

2. Background: Default&Refine

The Default&Refine algorithm is a fairly straightforward 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 very 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 [7].

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

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

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 correct phoneme is extracted as a refined rule. Exceptions to this refined rule are similarly represented by further refinements, and so forth, leading to a rule set that describes the training set with complete accuracy. Further details can be found in [7].

3. 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.

3.1. Pseudo-phonemes

A pseudo-phoneme is used to model a phoneme that is consistently realised as two or more variants. In practise, we use the following process: we align the training lexicon (as discussed in Section 2), 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[9] symbol. The 'ϕ' symbol indicates phonemic nulls (inserted during alignment).

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	p1=ay‖ax	ae n ih m p1 t ϕ
delegate	d eh l ih g ay t ϕ d eh l ih g ax t ϕ	p1=ay‖ax	d eh l ih g p1 t ϕ
lens	l eh n z l eh n s	p2=s‖z	l eh n p2
close	k l ow z ϕ k l ow s ϕ	p2=s‖z	k l ow p2 ϕ

Once all pseudo-phonemes have been defined, the aligned training lexicon is regenerated in terms of the new phoneme set.

3.2. Generation restriction rules

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 p3 k p4 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.

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 p3-p4 above) is realised as one or more specific phoneme combinations (eh-ih or ih-aa) for all words in the training lexicon, the p3-p4 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 com-

plex 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.

4. Evaluation and Results

In order to evaluate the practicality of the proposed approach, we model the pronunciation variants occurring in the Oxford Advanced Learners Dictionary (OALD) [8]. We use the exact 60,399 word version of the lexicon as used by Black [3], and do not utilise stress assignment.

In all experiments we perform 10-fold cross-validation, based on a 90% training and 10% test set. The exact training and test word lists are used as reported on in [7]. 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}[1].

4.1. Benchmark systems

In previous experiments in which Default&Refine was applied to the OALD corpus [7], 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 2 as '*one variant*'. Before applying the new approach, we evaluate the effect on predictive accuracy if all variants are simply removed from the training lexicon, and list the results in Table 2 as '*no variants*'. As can be seen, results are comparable, with the variant-containing scores consistently somewhat lower because of the extra complexity introduced by variants. These two systems are used as benchmarks to evaluate the effect that the new approach to variant modelling has on predictive accuracy.

Table 2: *Predictive accuracy is comparable whether one variant is retained or all variants removed during training. (Tested on full test set.)*

Approach	Word accuracy		Phoneme accuracy		Phoneme correct	
		σ_{10}		σ_{10}		σ_{10}
one variant	86.46	0.15	97.41	0.03	97.67	0.03
no variants	86.87	0.16	97.49	0.03	97.74	0.03

[1]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 $\sigma_n = \frac{\sigma}{\sqrt{n}}$.

4.2. 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. We align the original training lexicon, generate a set of pseudo-phonemes, rewrite the aligned lexicon in terms of the new pseudo-phonemes, extract Default&Refine rules for the rewritten lexicon, and extract generation restriction rules based on the original 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. 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 baseline system *('no variants')* and the pseudo-phoneme approach *('pseudo-phone')*. Results are listed in Table 3. We see that the accuracy with which non-variants are predicted is not negatively influenced by the pseudo-phoneme modelling approach.

Table 3: *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		Phoneme accuracy		Phoneme correct	
		σ_{10}		σ_{10}		σ_{10}
no variants	86.93	0.16	97.50	0.03	97.75	0.03
pseudo-phone	86.92	0.15	97.50	0.03	97.76	0.03

4.3. 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 both 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 58% of expected variants are correctly generated and that 67% of generated variants are correct. In Table 4 we list the detailed results for four of the cross-validation sets.

Table 4: *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
58	43	23	57.43	71.60
56	40	20	58.33	73.68
64	45	32	58.72	66.67
53	34	28	60.92	65.43

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. When the variants classified as *'extra'* 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 'iy n k r iy s' and 'iy ng k r iy s' as variants of the word 'increase', but allows only the single pronunciation 'iy n k r iy s t' as a pronunciation of 'increased'. When the prediction system generates the alternative pronunciation 'iy ng k r iy s 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.

The above discussion suggests an interesting approach for the validation of variants: all variants that are generated using the pseudo-phoneme approach and are marked as erroneous can be verified manually. Correct variants can be added to the training set and the process repeated until all generated variants have been verified, resulting in a more consistent lexicon.

5. Conclusions

In this paper we have 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 the *OALD* corpus, we find that the incorporation of variants does not have a detrimental effect on the accuracy with which non-variants can be predicted. Additionally, the proposed approach was able to identify 58% of expected variants, and of the variants generated 67% were correct. These results do not take into account that some of the 33% variants identified as incorrect may be legal variants not included in the version of *OALD* used here.

Initial results indicate that the approach is similarly applicable to other lexicons studied (the Flemish FONILEX lexicon [10] and the Carnegie Mellon Pronunciation dictionary [11]). In future work we would like to verify this more rigorously, and also obtain a more quantitative indication of the amount of possibly inconsistent variants occurring in these dictionaries.

6. References

[1] M. Davel and E. Barnard, "Bootstrapping for language resource generation," in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, South Africa, 2003, pp. 97–100.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] P. Mertens and F. Vercammen, “Fonilex manual,” Tech. Rep., K.U.Leuven CCL, 1998.

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