

Automatic conversion between pronunciations of different English accents[☆]

Linsen Loots, Thomas Niesler^{*}

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Received 11 January 2010; received in revised form 6 July 2010; accepted 25 July 2010

Abstract

We describe the application of decision trees to the automatic conversion of pronunciations between American, British and South African English accents. The resulting phoneme-to-phoneme (P2P) conversion technique derives the pronunciation of a word in a new target accent by taking advantage of its existing available pronunciation in a different source accent. We find that it is substantially more accurate to derive pronunciations in this way than directly from the orthography and available target accent pronunciations using more conventional grapheme-to-phoneme (G2P) conversion. Furthermore, by including both the graphemes and the phonemes of the source accent, grapheme-and-phoneme-to-phoneme (GP2P) conversion delivers additional increases in accuracy in relation to P2P. These findings are particularly important for less-resourced varieties of English, for which extensive manually-prepared pronunciation dictionaries are not available. By means of the P2P and GP2P approaches, the pronunciations of new words can be obtained with better accuracy than is possible using G2P methods.

© 2010 Elsevier B.V. All rights reserved.

Keywords: English accents; Pronunciation modelling; G2P; P2P; GP2P; Decision trees; South African English

1. Introduction

The pronunciation dictionary is a key component of an Automatic Speech Recognition (ASR) as well as a text to speech (TTS) system. However, the development of this dictionary involves a great deal of time and effort by linguistic experts, a process that may be prohibitively expensive for under-resourced languages and accents. Nevertheless phonetic transcriptions in the appropriate accent can improve both ASR accuracy (Humphries et al., 2001) and TTS quality (Damper et al., 1999).

We consider the scenario where the pronunciations of words not in the dictionary are sought, and where these pronunciations are already available in a different accent

of the language in question. It will be shown that, in this situation, data-driven techniques similar to those used in grapheme-to-phoneme (G2P) conversion can be used to obtain the desired pronunciation from the existing pronunciation, instead of from the graphemes. This amounts to a conversion between the two accents.

In our experimental evaluation, we focus on General American (GenAm), Received Pronunciation (RP) and Standard South African English (SSAE). The first two are commonly-used and widely-studied reference accents for American and British English, respectively (Wells, 1982). In South Africa, a number of different English accents are prevalent. The accent generally used by first-language speakers is commonly referred to as Standard South African English and has been described as fairly similar to RP (Bowerman, 2004; Bekker, 2009). An important distinguishing factor of SSAE is the characteristic split of Wells' *KIT* vowel. This vowel appears in words like *kit*, *thin* and *sit*, and in SSAE has two clear allophonic variations: one close and one more central (Wells, 1982). While these words

[☆] Extended version of a paper presented at Interspeech 2009, Brighton, UK (Loots and Niesler, 2009).

^{*} Corresponding author. Tel.: +27 21 808 4118.

E-mail addresses: linsen@sun.ac.za (L. Loots), trn@sun.ac.za (T. Niesler).

contain a single vowel sound in most English accents, they are pronounced using two different vowel sounds in SSAE.

From our perspective the objective of accent conversion is to determine whether British or American pronunciations can be used to complement our still fairly small SSAE pronunciation dictionary, and how this can be done best. However, we believe that the techniques we consider are generally applicable to any accent pair.

The following section will describe the data upon which our experimental evaluation is based. Sections 3–5 then describe three alternative techniques that can be used to obtain pronunciations of new words. Finally, Section 7 discusses the results and Section 8 concludes.

2. Dictionaries

Four dictionaries were used to represent the three accents which we consider, and are described shortly in the following.

CMUDICT and PRONLEX were used to represent GenAm pronunciations. CMUDICT was developed at Carnegie Mellon University (CMU, 2009). This dictionary contains words drawn from a variety of sources, including an initial subset of over 20,000 words developed by hand and extensively verified at CMU. Other sources of pronunciations for CMUDICT include the Shoup and the Dragon dictionaries. PRONLEX is the COMLEX English Pronouncing Lexicon, and was hand-transcribed at the Linguistic Data Consortium (LDC, 2009).

The BEEP dictionary was used to represent RP and is drawn primarily from the Medical Research Council (MRC) Psycholinguistic Database and the Computer Usable Oxford Advanced Learners Dictionary (BEEP, 2009). In addition, it contains pronunciations provided by sources at Durham and Oxford Universities.

SSAE pronunciations were obtained from SAEDICT, a pronunciation dictionary under development at Stellenbosch University. All pronunciations in SAEDICT were transcribed by the same linguistic specialist to ensure consistency. Transcriptions were chosen to reflect commonly accepted SSAE pronunciations.

Table 1 describes each dictionary in terms of the number of words it contains, the total number of pronunciations and the average number of pronunciations per word (P/W). The number of pronunciations and P/W is also given for a set of 23,031 words common to all four dictionaries, as will be described in Section 2.2.

2.1. Phoneset

ARPABET (Rabiner and Juang, 1993) was chosen as the common phoneset in which to analyse the four dictionaries. Both BEEP and CMUDICT already use ARPABET, while PRONLEX uses an ARPABET short-hand that can easily be converted to standard ARPABET. SAEDICT uses a phoneset, based on IPA, developed to describe the languages of Southern Africa (Niesler et al., 2005). This was converted to ARPABET by means of a mapping based on the closest IPA symbol.

A small number of phonemes was found to be present in only a single dictionary, and these were mapped to alternative phonemes as indicated in Table 2. This normalisation ensures that the phonesets used by the four dictionaries match exactly.

2.2. Wordlist

In order to compare pronunciations, the set of words common to all four dictionaries was determined. Before extracting this set, all words were automatically converted to standard UK spelling using tools developed at the University of Sheffield (Wan et al., 2008). The set of common words contained 23,031 entries.

There are a number of words in each dictionary with multiple pronunciations, either homographs (words with the same spelling but different pronunciations), or words with alternate pronunciations. The number of pronunciations vary, but all four dictionaries contain an average of between 1.11 and 1.22 pronunciations per word, as reflected in Table 1. Of the 23,031 words common to all four dictionaries, 9344 have multiple pronunciations in at least one dictionary.

Table 2
Phoneme mappings used to achieve a common phoneset.

Dictionary	Mapping
BEEP	/ohl/ ⇒ /la/
	/lial/ ⇒ /lih axl/
	/leal/ ⇒ /leh axl/
	/lual/ ⇒ /luh axl/
PRONLEX	/whl/ ⇒ /hw/
	/enl/ ⇒ /ax nl/

Table 1
Number of entries in the four pronunciation dictionaries.

Dictionary	Entire dictionary			Common set		
	Words	Prons	P/W	Words	Prons	P/W
SAEDICT	30,390	36,954	1.22	23,031	28,087	1.22
BEEP	237,592	256,998	1.08	23,031	25,576	1.11
CMUDICT	123,292	133,694	1.08	23,031	26,771	1.16
PRONLEX	90,919	99,978	1.11	23,031	25,471	1.11

2.3. Pronunciation alignments

The pronunciations of corresponding words in the three accents were aligned using a dynamic programming algorithm that will be described in Section 3.1. This was done so that accents could be compared directly, as will be described in the following section. However, the alignments will also be used to obtain the one-to-one correspondences between phonemes in different accents. These correspondences are needed to train the decision trees that will be responsible for inter-accent conversions (described in Sections 4 and 5).

As an example, the alignment between SSAE and RP pronunciations of the word “reactions” is shown in Table 3. Nulls have been introduced where a phoneme has no counterpart in the other string (i.e. insertions and deletions). In this case the /ax/ in the SSAE pronunciation is deleted in RP.

Words with multiple pronunciations pose a problem both when comparing two dictionaries, and when subsequently using the aligned pronunciations to train decision trees for accent conversion. Our solution has been to use only the single pair of pronunciations per word (one from each dictionary) that gives the best alignment. An alternative approach is to consider all the possible pairings of pronunciations between the two dictionaries. However, preliminary experiments indicated that this did not influence the relative accuracies obtained for the various accents. Hence, the simpler approach of using the best-aligned pair was chosen for all further experiments.

2.4. Direct pronunciation comparison

In order to obtain a baseline for inter-accent similarity against which accent conversion can later be compared, a direct phonetic comparison of the different accents was performed. A summary of the results obtained for the 23,031 common words described in Section 2.2 is given in Table 4. The phoneme accuracy was calculated according to Eq. (1), where N_c , N_i and N_t are the numbers of correct, inserted and total phonemes, respectively. This accuracy is given for all phonemes, and separately for vowels and for consonants. The last column indicates the percentage of words that have identical pronunciations.

$$Acc = \frac{N_c - N_i}{N_t}. \quad (1)$$

It is clear from Table 4 that SAEDICT and BEEP match most closely, while the poorest correspondence is found between CMUDICT and SAEDICT. On average, the American pronunciations are closer to RP than to SSAE.

Table 3
Two aligned pronunciations of the word “reactions”.

SAEDICT	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	<i>ax</i>	<i>n</i>	<i>s</i>
BEEP	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	–	<i>n</i>	<i>z</i>

Table 4

Accuracies (%) of direct pronunciation alignments between dictionaries.

Dictionary pair	Phoneme	Vowel	Cons	Word
BEEP $\leftrightarrow$ SAEDICT	92.1	84.5	96.9	59.6
CMUDICT $\leftrightarrow$ SAEDICT	81.6	59.8	95.2	29.4
PRONLEX $\leftrightarrow$ SAEDICT	87.3	72.8	96.2	43.0
CMUDICT $\leftrightarrow$ BEEP	84.7	67.7	94.6	37.8
PRONLEX $\leftrightarrow$ BEEP	89.4	79.0	95.6	50.6
PRONLEX $\leftrightarrow$ CMUDICT	89.0	73.2	98.6	50.7

Furthermore, vowels differ more strongly than consonants between accents. While at least 94.6% of consonants matched between any pair of dictionaries, this figure varied between 59.8% and 84.5% for vowels. A more detailed analysis of the phonetic differences between the three accents can be found in (Loots and Niesler, 2009).

3. Grapheme-to-phoneme conversion (G2P)

G2P conversion is the process of determining a word’s pronunciation from its spelling. Several data-driven G2P methods have been suggested in the literature, including decision trees (Black et al., 1998; Suontausta and Häkkinen, 2000; Kienappel and Kneser, 2001; Webster and Braunschweiler, 2008), HMMs (Taylor, 2005), pronunciation by analogy (Marchand and Damper, 2000) and memory-based learning (Daelemans et al., 1999). Decision trees have been shown to yield competitive accuracy in G2P conversion (Han and Chen, 2004) and to be a highly competitive approach to the generation of accent-specific pronunciation variants for ASR (Humphries et al., 2001). We have chosen decision trees as the method with which we will investigate the success of accent conversion.

Although this paper does not focus on G2P systems, the methods used for accent conversion in Sections 4 and 5 are based on analogous techniques. Hence, G2P conversion by means of decision trees will be reviewed first, and used as a further baseline against which the later systems can be measured.

For G2P conversion, the graphemes and their context form the input of the decision tree classifier, and the phonemes the output class. Pronunciations are generated by sequentially passing graphemes and their context through the tree, assigning each to a phoneme class, and then concatenating these output phonemes. This process is illustrated in Fig. 1.

3.1. Aligning graphemes and phonemes

In order to sequentially generate an output phoneme for each grapheme, the decision tree requires a one-to-one correspondence between the graphemes and the phonemes. An example of such an alignment is shown in Table 5.

Discrepancies in length and alignment between the graphemes and phonemes are dealt with by inserting nulls into the phoneme string (for letters not corresponding to a phoneme), and by combining pairs of phonemes into

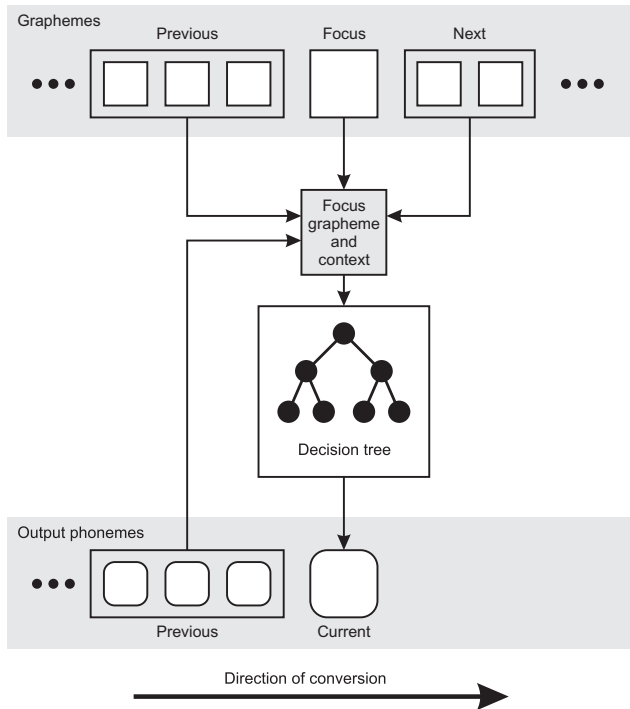


Fig. 1. Diagrammatic representation of grapheme-to-phoneme (G2P) conversion using decision trees.

double phonemes for the relatively few cases where one grapheme corresponds to two phonemes (Damper et al., 2004; Suontausta and Häkkinen, 2000).

The primary method of aligning graphemic and phonetic strings uses the Viterbi algorithm (Daelemans and van den Bosch, 1994; Andersen et al., 1996; Pagel et al., 1998). The process begins by estimating $P(G, P)$, the unigram probability that a given grapheme is associated with a given phoneme. This is estimated using those words that have graphemic and phonetic strings of equal length (Andersen et al., 1996). Optimal alignments based on these estimated probabilities are then generated by means of dynamic programming for all words in the corpus. The newly-aligned words are then used to recalculate the unigram probabilities, and the process is repeated until the distributions converge (Pagel et al., 1998; Webster and Braunschweiler, 2008).

Double phonemes can be determined automatically by iteratively repeating this process. First, suitable double phonemes are identified from unsuccessful alignments between grapheme and phoneme strings. Alignment is subsequently re-attempted, and the process repeated until all words align with their pronunciations.

Table 5
An example of grapheme-to-phoneme (G2P) alignment for the word “extreme”.

Grapheme	e	x	t	r	e	m	e
Phoneme	eh	k s	t	r	iy	m	–

3.2. Decision trees for G2P

We have employed deterministic binary decision trees for G2P conversion, for which each node is associated with a true/false question regarding the input grapheme and its context. The tree is traversed from the root by recursively using the answer of each node’s question to determine which child node to choose. Each leaf node is associated with an output phoneme, which constitutes the classification result (Breiman et al., 1984). This process produces a single pronunciation for the presented orthography.

Decision trees are grown recursively. For each new node the available training data are split according to all possible questions. The question which results in the greatest entropy gain is then chosen for that node (Suontausta and Häkkinen, 2000). For a node t with entropy $i(t)$, children t_L and t_R with respective entropies $i(t_L)$ and $i(t_R)$, and the proportion of node t ’s data associated with each child given by p_L and p_R such that both are non-negative and sum to unity, the entropy gain is (Breiman et al., 1984):

$$\Delta i = i(t) - p_L i(t_L) - p_R i(t_R). \quad (2)$$

When choosing the question with the largest Δi , $i(t)$ can be omitted from the calculation with no loss of generality as it remains constant for all questions at a given node. Furthermore p_k can be approximated by the relative frequency N_{t_k}/N_t for a child node t_k , while $1/N_t$ is constant for all questions.

$$\Delta i = \sum_{\forall \text{ children } k} \sum_{\forall \text{ phonemes } p} N_{t_k, p} \log \left(\frac{N_{t_k, p}}{N_{t_k}} \right). \quad (3)$$

It is therefore sufficient to maximise Eq. (3) in order to determine the optimal question for a new decision tree node (Kienappel and Kneser, 2001).

3.3. Decision tree training and testing

A G2P decision tree includes many parameters, the most important of which are the number of graphemes and previously generated phonemes included in the context, the direction of the conversion, the use of clusters, and the use of pruning (Webster and Braunschweiler, 2008).

To obtain values for these parameters in our G2P experiments, the phoneme accuracy of the decision tree was optimised on a held-out development set within a 10-fold cross-validation framework. This was achieved by dividing the 23,031 common words described in Section 2.2 into 10 non-overlapping and approximately equally-sized partitions. Reserving each partition in turn for later use as a test set, the remaining 9 partitions (corresponding to 90% of the data) were again subdivided into 10 equal partitions. The first of these sub-partitions served as a development set, and the remaining 9 as a training set for parameter optimisation. Decision trees were trained on this training set for a range of parameter values, and those that lead to the highest phoneme accuracy measured on the

associated development set were identified as optimal. This process was repeated for each of the 10 partitions of the 23,031-word data set. In general, different optimal parameter values were obtained in each case.

The data partitioning must be carefully considered, as training, development and testing datasets should be properly representative of the dictionary (and thus the language) as a whole. Furthermore, it is important that there is not too much overlap between the data partitions, as might occur between words with the same root but different affixes. This should be avoided because it can cause the test or development data to be unrealistically similar to the training data, leading to results that are not properly indicative of the system's performance on completely unknown words. Error rates have been reported to drop by a relative margin of up to 60% when training and test data contain words with overlapping roots (Demberg et al., 2007).

We have chosen to first sort the 23,031 words alphabetically, and then assign groups of 10 consecutive words at a time to the 10 cross-validation partitions. The same procedure was followed for the nested subdivision into development and training partitions for parameter optimisation. Manual inspection showed this strategy, described in (Kienappel and Kneser, 2001), to consign most related words to the same dataset. In comparison, the use of random selection gave more optimistic results, with an absolute increase in word accuracy of approximately 10% observed for G2P conversion.

The optimal number of graphemes and previously generated phonemes included in the G2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined by 10-fold cross-validation for each of the four dictionaries, leading to a total of 40 optimised parameter sets.

In agreement with the findings of others, the highest accuracy was obtained when including on average three graphemes to the left and four to the right of the focus grapheme, as well as the four phonemes most recently produced as part of the G2P process, while converting from right-to-left (Pagel et al., 1998; Torkkola, 1993). Note that, during training, the phonemes in question are taken from the dictionary pronunciation, while during G2P conversion the phonemes hypothesised by the decision tree are used. Since the G2P process is not perfect, this will lead to some mismatch between training and testing conditions.

The inclusion of questions relating to clusters of phonemes or graphemes, rather than merely individual phonemes or graphemes, was found to improve G2P performance in 27 of the 40 optimisations, measured in each case on the respective development set. The clusters were automatically determined using the greedy optimisation algorithm described in (Kienappel and Kneser, 2001). For each context position, this algorithm hierarchically groups graphemes or phonemes into clusters by means of repeated entropy-based binary splits. Optimal

performance was achieved when using on average 26 graphemic and 39 phonemic clusters, although the precise number depends on the context position as well the development set in question. Examples of frequent grapheme clusters include (c, k, q) , (a, o, u) and (s, z) , while $(laal, lael, lehl, laol, lyl)$ and $(lthl, ldhl)$ were common phoneme clusters.

Finally, decision tree pruning removes child nodes and their associated subtrees when these do not contribute at least a threshold improvement to the classification accuracy on a held-out set (Breiman et al., 1984). In 9 of the 40 optimisations, using 5% of the training data for pruning led to improved performance measured on the respective development set, but in most cases pruning was not used.

All experiments were performed using software developed at the University of Stellenbosch.

3.4. Experimental results for G2P

Using the optimal parameters determined in the previous section, G2P decision trees were trained for each dictionary individually. This allowed us to determine how accurately pronunciations could be produced based on training data drawn from the same accent. Together with the direct pronunciation comparison of Section 2.4, these results serve as a baseline for the accent conversion techniques that are presented in Sections 4 and 5.

The same 10 disjoint partitions of the 23,031-word data set described in Section 3.3 were used for training and testing the decision trees. Each of the 10 partitions was employed as a test set, while decision trees were trained using the remaining 9 partitions and the corresponding optimal parameters.

The generated pronunciations for each of the 10 test partitions were aligned with the pronunciations in the dictionary by dynamic programming. From these alignments the number of substitutions, insertions and deletions were determined. The phoneme accuracy was subsequently calculated using Eq. (1), and is presented in Table 6 for each of the four dictionaries. Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical. Where words have multiple pronunciations in the target accent, any of these are considered as correct. In each case the reported percentages are averages for the 10 cross-validation splits, while 95% confidence intervals were calculated using the bootstrap method described in (Bisani and Ney, 2004).

Table 6

Grapheme-to-phoneme (G2P) accuracies for the four dictionaries, with 95% confidence intervals and decision tree sizes.

Dictionary	Phoneme (%)	Word (%)	Size (nodes)
SAEDICT	90.67 ± 0.54	59.64	18,064
BEEP	91.76 ± 0.53	65.24	14,164
CMUDICT	91.26 ± 0.54	63.29	15,328
PRONLEX	92.22 ± 0.51	65.82	14,044

The results show little variation among the three accents. It is clear, however, that SAEDICT has the lowest G2P accuracy, while PRONLEX has the highest. This suggests that SAEDICT has the least regular relationship between its graphemes and phonemes. This is also reflected by the substantially larger size of SAEDICT decision trees. The results are comparable to those found by other researchers, where accuracies of approximately 91% are common (Black et al., 1998; Suontausta and Häkkinen, 2000).

4. Phoneme-to-phoneme conversion (P2P)

Grapheme-to-phoneme conversion can be used to estimate the pronunciations of new words, given a set of training pronunciations in the same accent or language. It is sometimes the case, however, that pronunciations for the desired words are already available, but in a different accent of the same language. For example, the pronunciations of many words currently not in SAEDICT are available in CMUDICT, PRONLEX or BEEP. We now address the question of whether such existing pronunciations in a different (source) accent can be used to estimate the pronunciation in the desired (target) accent. Since this process involves the conversion of one phoneme string (the source pronunciation) to another (the target pronunciation), it will be referred to as phoneme-to-phoneme (P2P) conversion.

P2P conversion has been used successfully to correct the output of a G2P system when applied to the derivation of pronunciations for proper names (van den Heuvel et al., 2007). We will not perform an explicit G2P conversion, but employ the same decision tree based algorithm used by this approach. In the case of P2P conversion, the phonemes of the source accent take the role of the “graphemes” and those of the target accent of the “phonemes”. This process is shown diagrammatically in Fig. 2.

Parameter optimisation for the P2P decision trees was carried out using a held-out development set within a 10-fold cross-validation framework, in an analogous fashion to the process described in Section 3.3. The optimal number of phonemes included in the P2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined for each of the 12 possible dictionary pairs, leading to a total of 120 optimised parameter sets.

It was found that the inclusion of on average two phonemes to the left and two to the right of the focus phoneme was optimal. Including previously generated output phonemes was beneficial in 65 of the 120 optimisations, while right-to-left conversion was best in 63 cases. The use of phoneme clusters improved performance in all 120 cases. In contrast to the G2P findings, decision tree pruning also led to improved performance, with the best results obtained when between 40% and 50% of the dictionary was used for training and the remainder for pruning. A

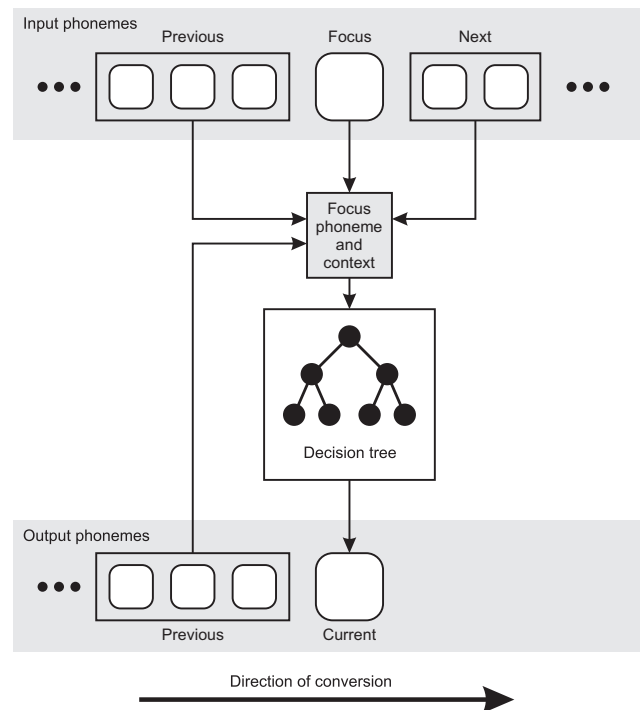


Fig. 2. Diagrammatic representation of phoneme-to-phoneme (P2P) conversion using decision trees.

notable difference between the G2P and P2P approaches is a higher incidence of insertions when aligning two pronunciations. While there are relatively few cases of a single grapheme corresponding to multiple phonemes, there are many more cases of a single source phoneme corresponding to multiple target phonemes. This led to the use by the decision tree of a larger number of double phonemes (treated as a single symbol). A total of 151 double phones was needed across all dictionary pairs, compared with just 63 for G2P. Double phonemes were determined automatically by iteratively aligning the training pronunciations of all words in a dictionary pair, finding pronunciations that could not be aligned, and adding the double phonemes required for the alignment to succeed. This process was also employed for G2P conversion, as set out in Section 3.1.

The P2P accent conversion accuracies are given in Table 7. The phoneme and word accuracies are again calculated as described in Section 3.4. The same 10-fold split into training and testing data was used in order to make Tables 6 and 7 directly comparable.

It is clear that the pronunciations that can most accurately be derived from each other are CMUDICT and PRONLEX. This suggests that the apparent differences when directly comparing their pronunciations (Table 4) are at least in part due to systematic differences in the transcription conventions followed, rather than due to fundamental differences in the pronunciations.

Overall, the accuracy of the P2P process is considerably higher than that of G2P conversion. It is also evident that

Table 7

Phoneme-to-phoneme (P2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP $\Rightarrow$ SAEDICT	96.19 $\pm$ 0.32	79.53	2708
CMUDICT $\Rightarrow$ SAEDICT	94.76 $\pm$ 0.36	72.79	3471
PRONLEX $\Rightarrow$ SAEDICT	95.62 $\pm$ 0.34	76.69	2951
SAEDICT $\Rightarrow$ BEEP	96.30 $\pm$ 0.32	80.77	2794
CMUDICT $\Rightarrow$ BEEP	95.43 $\pm$ 0.36	76.89	3143
PRONLEX $\Rightarrow$ BEEP	96.25 $\pm$ 0.33	80.82	2776
SAEDICT $\Rightarrow$ CMUDICT	94.54 $\pm$ 0.39	72.46	3511
BEEP $\Rightarrow$ CMUDICT	95.28 $\pm$ 0.37	75.91	3308
PRONLEX $\Rightarrow$ CMUDICT	97.64 $\pm$ 0.26	87.20	1995
SAEDICT $\Rightarrow$ PRONLEX	95.16 $\pm$ 0.36	74.44	3221
BEEP $\Rightarrow$ PRONLEX	95.66 $\pm$ 0.35	77.31	3005
CMUDICT $\Rightarrow$ PRONLEX	96.81 $\pm$ 0.29	82.37	2219

the decision trees used for P2P conversion are substantially smaller than those used for G2P conversion. A detailed discussion of these results is deferred to Section 7.

5. Grapheme-and-phoneme-to-phoneme conversion (GP2P)

For some words, there may be information present in the orthography that is not represented phonetically. For example, words with different spellings can have identical pronunciations in one accent but not in another. In our case this occurs for “spokesmen” and “spokesman”, which are both pronounced */s p o w k s m a x n/* in BEEP, while “spokesmen” is pronounced */s p o w k s m e h n/* in SAEDICT.

The hypothesis that the orthography can provide information not already captured by the pronunciation is tested in the following by using both the graphemes and the phonemes in the source accent to derive the pronunciation in the target accent. This modification will be referred to as grapheme-and-phoneme-to-phoneme conversion (GP2P).

In order to employ the source graphemes as additional context for the decision tree, they are first aligned with the source phonemes, as described in Section 3.1. In this case the alignment was used to associate a sequence of zero or more graphemes with every source phoneme (rather than associating zero or more phonemes with each grapheme).

The decision tree question set was extended to include questions of the type “Does the grapheme sequence aligned with the phoneme contain the letter X?”. These questions could be asked of the focus phoneme or of any phoneme in the input context of the source accent. Fig. 3 illustrates the GP2P conversion process.

Optimal decision tree parameters were found using the same process described in Section 4. It was found that the parameters that produced the best performance for P2P systems did so also for GP2P systems. The only departure from this finding was that between 55% and 65% of the

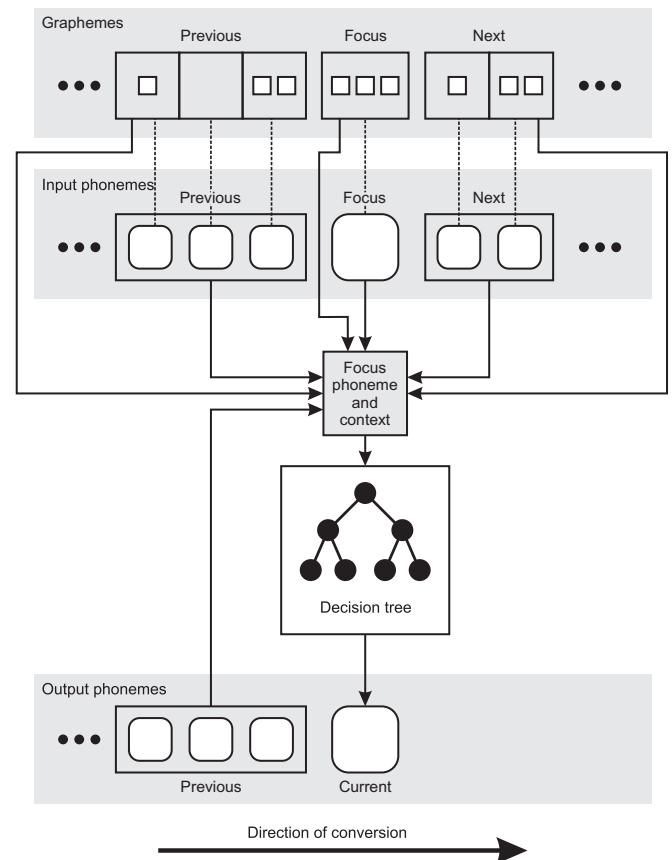


Fig. 3. Diagrammatic representation of grapheme-and-phoneme-to-phoneme (GP2P) conversion using decision trees.

training data is best used for pruning, an absolute increase of approximately 5% relative to the P2P systems. Further refinements that were considered were allowing questions to be based on grapheme clusters obtained in an analogous fashion to the phoneme clusters, including the length of the grapheme sequence aligned with a source phoneme, and allowing questions to include fixed grapheme sequences (n -tuples). However, none of these were found to yield any further performance improvements during cross-validation. The results of the GP2P experiments are presented in Table 8, and will be discussed in relation to the performance of G2P and P2P conversion in Section 7.

6. Multiple source dictionaries

A final question addressed by our work is whether the use of more than one source dictionary could further improve the performance of the GP2P approaches described in the previous section. As a specific example, we considered the two GP2P systems most accurately able to derive SSAE pronunciations. From Table 8, these are the BEEP to SAEDICT and PRONLEX to SAEDICT configurations. The GP2P algorithm was extended by allowing the decision trees to query the graphemes and phonemes in both BEEP and PRONLEX. To achieve this, BEEP was considered the primary and PRONLEX the

Table 8

Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP $\Rightarrow$ SAEDICT	96.55 $\pm$ 0.31	81.58	2362
CMUDICT $\Rightarrow$ SAEDICT	95.60 $\pm$ 0.34	76.79	2831
PRONLEX $\Rightarrow$ SAEDICT	96.05 $\pm$ 0.33	78.99	2591
SAEDICT $\Rightarrow$ BEEP	97.46 $\pm$ 0.27	86.78	1880
CMUDICT $\Rightarrow$ BEEP	96.58 $\pm$ 0.32	82.28	2596
PRONLEX $\Rightarrow$ BEEP	96.97 $\pm$ 0.30	84.41	2078
SAEDICT $\Rightarrow$ CMUDICT	96.61 $\pm$ 0.32	82.38	2313
BEEP $\Rightarrow$ CMUDICT	96.85 $\pm$ 0.31	83.96	2362
PRONLEX $\Rightarrow$ CMUDICT	97.87 $\pm$ 0.25	88.44	1613
SAEDICT $\Rightarrow$ PRONLEX	97.26 $\pm$ 0.27	84.60	2075
BEEP $\Rightarrow$ PRONLEX	97.30 $\pm$ 0.28	85.44	1943
CMUDICT $\Rightarrow$ PRONLEX	97.75 $\pm$ 0.25	87.22	1551

additional source, and PRONLEX pronunciations were aligned with BEEP pronunciations as described in Section 4.

The decision tree question set was extended to include dictionary-specific questions, such as “Is the preceding phoneme in PRONLEX a plosive?”. This set of questions is a superset of those used for GP2P conversion with a single source dictionary, as was described in Section 5. Optimal decision tree parameters, again obtained by 10-fold cross-validation, were found to coincide with those that were optimal for the single source dictionary GP2P systems. The results, presented in Table 9, indicate that a small improvement in performance over the baseline BEEP to SAEDICT GP2P system was achieved, but that this improvement is not statistically significant.

It therefore appears that further significant gains using multiple source dictionaries are not achievable for SSAE. However, it is possible that the result would be different for other English accents, or for other languages.

Table 9

Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies for SSAE when using one (BEEP) and two (BEEP and PRONLEX) source dictionaries, with 95% confidence intervals and decision tree sizes.

GP2P system	Phoneme (%)	Word (%)	Size (nodes)
BEEP $\Rightarrow$ SAEDICT	96.55 $\pm$ 0.31	81.58	2362
BEEP + PRONLEX $\Rightarrow$ SAEDICT	96.62 $\pm$ 0.31	81.84	2272

Table 10

Average accuracies achieved by the G2P, P2P and GP2P conversion approaches.

Dictionary	Phoneme (%)			Word (%)		
	G2P	P2P	GP2P	G2P	P2P	GP2P
$\Rightarrow$ SAEDICT	90.67	95.52	96.07	59.64	76.34	79.12
$\Rightarrow$ BEEP	91.76	96.00	97.00	65.24	79.50	84.49
$\Rightarrow$ CMUDICT	91.26	95.82	97.11	63.29	78.52	84.93
$\Rightarrow$ PRONLEX	92.22	95.88	97.44	65.82	78.04	85.75
Average	91.48	95.80	96.90	63.50	78.10	83.57

7. Discussion

When comparing Tables 6 and 7, it is clear that the pronunciation of a word in a new accent is in all cases substantially more accurately determined from its pronunciation in another accent by means of P2P conversion, than by means of G2P conversion within the same accent. Furthermore, Table 8 shows that the use of source graphemes in addition to source phonemes in GP2P conversion is rewarded with a further improvement in accuracy. This additional improvement is beyond the 95% confidence interval for all dictionary pairs except CMUDICT to PRONLEX, which is also the pair for which the highest P2P conversion accuracy was obtained.

Converting between different accents also gives considerably more accurate results than merely using another accent's dictionary, as can be seen from Table 4. Substituting one English dictionary for another with no modifications gives at best a 60% and at worst a 30% word accuracy. For GP2P, however, even the poorest matching pair of dictionaries achieves a word accuracy above 76%.

Table 10 summarises the results presented in Tables 6–8, and indicates the average performance across all four dictionaries. The results in the table show that, on average, P2P conversion leads to a 4.3% absolute improvement in phoneme accuracy over G2P conversion, which corresponds to a 51% relative decrease in error rate. Furthermore, the percentage of words with pronunciation errors drops by 14.6%, which is a 40% relative reduction in error. The move from P2P to GP2P conversion results in an additional average absolute phoneme accuracy improvement of 1.1%, which corresponds to a 26% relative reduction in the error rate. The corresponding percentage of incorrectly pronounced words drops by 5.5%, which is a relative improvement of 25%. Overall, GP2P conversion achieves relative reductions of 64% and 55% in the phoneme and word error rates, respectively, when compared with G2P conversion.

The introduction of source graphemes in GP2P conversion also led to a considerable reduction in the size of the decision trees. While P2P conversion employed 2925 nodes on average, the corresponding figure for GP2P falls to 2183. However, both P2P and GP2P trees are much smaller than G2P trees which contained 15,400 nodes on average.

Hence, when a word's pronunciation is not available, as is fairly common in less prominent accents, it is better to first search for it in a different pronunciation dictionary. If it is found, it should be converted to the target accent by means of GP2P or P2P techniques. G2P methods should be reserved as a last resort.

8. Summary and conclusions

This study has investigated methods that can be used to derive the pronunciation of a word in a new accent from its known pronunciation in another accent of the same language.

The performance of these techniques has been evaluated by applying them to three accents of English: RP, GenAm and South African English. In order to achieve this, four pronunciation dictionaries were employed: BEEP for RP, CMUDICT and PRONLEX for GenAm, and SAEDICT for South African English.

As a baseline for comparison, G2P techniques were first applied to the dictionaries individually, with comparable results in all four cases. The G2P algorithm was then modified to allow conversion of pronunciations from one accent to another, first using only the phonemes of the source accent (P2P) and then using both the phonemes and the graphemes (GP2P).

If we consider the specific case of South African English, for which pronunciation resources are limited, our experiments show that the pronunciations of 82% of new words obtained using the developed GP2P conversion technique are correct. In comparison, this figure is 60% when using conventional G2P conversion.

Our research can be extended by performing listening tests, as well as by performing Automatic Speech Recognition, to determine on a perceptual and an acoustic level how well the generated pronunciations match the desired accent. This would provide a further indication of the validity of our findings. Furthermore, it is possible that the techniques which we have presented would bear fruit if applied to other accents of the same language, or even to related languages, such as Dutch and Afrikaans.

Acknowledgements

This material is based on work supported financially by the South African National Research Foundation (NRF) under Grants FA2007022300015 and TTK2007041000010, and was executed using the High-Performance Computing (HPC) facility at Stellenbosch University.

References

- Andersen, O., Kuhn, R., Lazarides, A., Dalsgaard, P., Haas, J., Nöth, E., 1996. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proc. ICSLP, Philadelphia, USA.
- BEEP, 2009. The British Example Pronunciation (BEEP) Dictionary. <<http://svr-www.eng.cam.ac.uk/comp.speech/section1/lexical/bEEP.html>> (accessed March 2009).
- Bekker, I., 2009. The vowels of South African English. Ph.D. Thesis, North-West University, Potchefstroom, South Africa.
- Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. ICASSP, Montreal, Canada.
- Black, A., Lenzo, K., Pagel, V., 1998. Issues in building general letter to sound rules. In: Proc. ESCA Speech Synthesis Workshop, Jenolan Caves, Australia.
- Bowerman, S., 2004. White South African English: phonology. In: Schneider, E., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), . In: A Handbook of Varieties of English, Vol. 1. Mouton de Gruyter, Berlin, pp. 931–942.
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove.
- CMU, 2009. Carnegie Mellon University Pronouncing Dictionary. <<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>> (accessed March 2009).
- Daelemans, W., van den Bosch, A., 1994. Language-independent data-oriented grapheme-to-phoneme conversion. In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer, New York, pp. 77–89.
- Daelemans, W., van den Bosch, A., Zavrel, J., 1999. Forgetting exceptions is harmful in language learning. Machine Learn. 34, 11–41.
- Damper, R., Marchand, Y., Adamson, M., Gustafson, K., 1999. Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. Comput. Speech Lang. 13 (2), 155–176.
- Damper, R., Marchand, Y., Marsters, J., Bazin, A., 2004. Aligning letters and phonemes for speech synthesis. In: Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA.
- Demberg, V., Schmid, H., Möhler, G., 2007. Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. In: Proc. ACL, Prague, Czech Republic.
- Han, K.-S., Chen, G.-L., 2004. Letter-to-sound for small-footprint multilingual TTS engine. In: Proc. ICSLP, Jeju, Korea.
- Humphries, J., Woodland, P., Pearce, D., 2001. Using accent-specific pronunciation modelling for robust speech recognition. In: Proc. ICASSP, Salt Lake City.
- Kienappel, A.K., Kneser, R., 2001. Designing very compact decision trees for grapheme-to-phoneme transcription. In: Proc. Eurospeech, Aalborg, Denmark.
- LDC, 2009. COMLEX English pronouncing lexicon. <<http://www ldc.u-penn.edu>> (accessed March 2009).
- Loots, L., Niesler, T., 2009. Data-driven phonetic comparison and conversion between South African, British and American English pronunciations. In: Proc. Interspeech, Brighton, UK.
- Marchand, Y., Damper, R., 2000. A multistrategy approach to improving pronunciation by analogy. Comput. Linguist. 26 (2), 195–219.
- Niesler, T., Louw, P., Roux, J., 2005. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. Southern African Linguist. Appl. Lang. Studies 23 (4), 459–474.
- Pagel, V., Lenzo, K., Black, A., 1998. Letter to sound rules for accented lexicon compression. In: Proc. ICSLP, Sydney, Australia.
- Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, USA.
- Suontausta, J., Häkkinen, J., 2000. Decision tree based text-to-phoneme mapping for speech recognition. In: Proc. ICSLP, Beijing, China.
- Taylor, P., 2005. Hidden Markov models for grapheme to phoneme conversion. In: Proc. Interspeech, Lisbon, Portugal.

- Torkkola, K., 1993. An efficient way to learn grapheme-to-phoneme rules automatically. In: Proc. ICASSP, Minneapolis, USA.
- van den Heuvel, H., Martens, J.-P., Konings, N., 2007. G2P conversion of names. what can we do (better)? In: Proc. Interspeech, Antwerp, Belgium.
- Wan, V., Dines, J., El Hannani, A., Hain, T., 2008. Bob: a lexicon and pronunciation dictionary generator. In: Proc. Spoken Language Technology Workshop, Goa, India.
- Webster, G., Braunschweiler, N., 2008. An evaluation of non-standard features for grapheme-to-phoneme conversion. In: Proc. Interspeech, Brisbane, Australia.
- Wells, J., 1982. Accents of English. Cambridge University Press, Cambridge.