Comparing direct G2P with G2P followed by accent conversion when determining pronunciations for South African English

Linsen Loots, Thomas Niesler
Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa.
linsen@sun.ac.za, trn@sun.ac.za

***Abstract*—It has been shown that techniques known as grapheme-and-phoneme-to-phoneme (GP2P) conversion can be used to derive pronunciations in a poorly-resourced accent, such as South African English, using available pronunciations in better-resourced accents of the same language, such as British and American English. However if the pronunciation is not available in either accent, it must be obtained using grapheme-to-phoneme (G2P) conversion in either the source or the target accent. The question therefore arises whether it is better to apply G2P in the source accent and then GP2P to obtain the desired pronunciation in the target accent, or to apply G2P directly to the target accent. This study finds that if the source dictionary used has a high G2P accuracy (due to the dictionary's size, regularity, or both), it is advantageous to generate a pronunciation in the source accent first using G2P, and subsequently convert this pronunciation to the target accent.**

Index Terms: English accents, pronunciation modelling, G2P, P2P, GP2P, decision trees, South African English

I. Introduction

Pronunciation dictionaries are key components of automatic speech recognition (ASR) as well as text to speech (TTS) systems. However, the development of such dictionaries involves a great deal of time and effort by linguistic experts, a process that may be prohibitively expensive for under-resourced languages and accents.

Where the pronunciations of words not in a dictionary are sought, and where these pronunciations are already available in a different accent of the language in question, phoneme-to-phoneme (P2P) conversion is an effective method of obtaining the desired pronunciations from the known pronunciations [1]. Grapheme-and-phoneme-to-phoneme (GP2P) conversion is an extension of P2P conversion, which includes graphemes as additional input. It has been shown in [1] that GP2P performs better than P2P, and as a result has been chosen as a focus for this study.

If the desired word is not present in the dictionary of either accent, however, GP2P conversion can not be applied. In this case, where the word is in neither the source nor the target dictionary, there are two possible approaches. The first is to use grapheme-to-phoneme (G2P) conversion in the target accent, and generate the pronunciation directly. The second approach is to generate the pronunciation in the source accent using G2P, and to then to convert this to the target accent using GP2P. The latter technique is potentially advantageous because G2P accuracy increases with training set size. If the source accent has a significantly larger dictionary, this approach may be able to leverage the increased training set size to improve pronunciation generation accuracy in the smaller target accent. This study investigates whether this is indeed the case. Figure 1 illustrates the relationship between the two approaches.

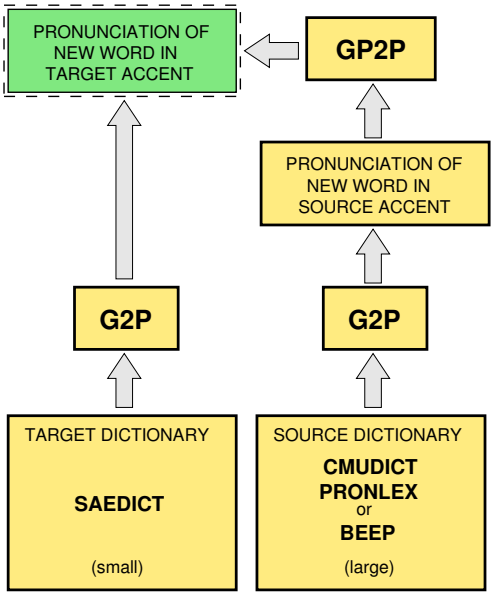

Fig. 1. An illustration of the two approaches to pronunciation generation compared in this paper.

For our investigation, 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 [2]. 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 [3], [4].

From our perspective the objective of this study is to determine whether existing and extensive British or American re-

sources can be taken advantage of when determining unknown pronunciations for our still fairly small SSAE pronunciation dictionary. However, we hope that the techniques we consider are generally applicable also to other accent pairs.

The following section describes the data upon which our experimental evaluation is based, after which an overview of G2P and GP2P conversion is given. Section IV describes the manner in which experiments were conducted, as well as the experimental results. Finally, Sections V and VI discuss the results and give conclusions.

II. Dictionaries

Four dictionaries were used to represent the three accents which we consider: CMUDICT, PRONLEX, BEEP and SAEDICT.

CMUDICT and PRONLEX were used to represent GenAm pronunciations. CMUDICT was developed at Carnegie Mellon University [5]. 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. PRONLEX is the COMLEX English Pronouncing Lexicon, and was hand-transcribed at the Linguistic Data Consortium [6].

The BEEP dictionary was used to represent RP and is drawn primarily from the MRC Psycholinguistic Database and the Computer Usable Oxford Advanced Learners Dictionary [7]. 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.

Dictionary	Size (words)
SAEDICT	30 390
BEEP	237 184
CMUDICT	123 646
PRONLEX	90 632

TABLE I
NUMBER OF ENTRIES PER PRONUNCIATION DICTIONARY.

Table I describes each dictionary in terms of the number of words it contains (excluding abbreviations). It is evident that SAEDICT is the smallest dictionary by a considerable margin.

A. Phoneme set

ARPABET [8] was chosen as the common phoneme set 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 phoneme set, based on IPA, developed to describe the languages of Southern Africa [9]. This was converted to ARPABET by means of a mapping based on the closest IPA symbol.

B. 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 [10], and abbreviations were removed. The resulting set of common words contains 23 006 entries.

III. G2P and GP2P conversion

G2P conversion applies machine learning to convert the known graphemic representation of a word into its unknown pronunciation. Several data-driven G2P methods have been suggested in the literature, including decision trees [11], [12], [13], [14], HMMs [15], pronunciation by analogy [16], default&refine [17] and memory-based learning [18]. Decision trees have been shown to yield competitive accuracy in G2P conversion [19] and to be effective for GP2P conversion [1]. We have chosen decision trees as the method with which we will investigate the accuracy of G2P and GP2P conversion.

Our implementations employ deterministic binary decision trees, for which each node is associated with a true/false question regarding the input grapheme and/or phoneme, 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 [20]. 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 [12]. A more detailed discussion of decision trees and their use for P2P and GP2P can be found in [1].

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.

For GP2P conversion, the graphemes of the words are used as input in addition to the source pronunciation's phonemes. In order to do so, these graphemes must first be aligned with the source phonemes. This is achieved by means of dynamic programming.

IV. Experimental method and results

In order to allow direct comparison between results, most experiments were carried out using the 23 006 words common to all four dictionaries, as described in Section II-B. However, in order to investigate the effects of increasing the training set size for the source G2P converter, additional experiments

that include all the words available in the respective source dictionaries were also performed.

A. Decision tree training and testing

There are a number of parameters that need to be specified when training a decision tree, such as the context window, direction of conversion and amount of data reserved for pruning. To obtain values for these parameters, 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 23 006 words 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 divided 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 23 006-word data set. In general, different optimal parameter values were obtained for each partition. Finally, each of the 10 partitions was employed as a test set, while decision trees were trained using the remaining 9 partitions with the corresponding optimal parameters. All experiments carried out on the 23 006 set of common words utilised decision trees trained and tested with the same 10 disjoint partitions.

The same experiments were also carried out using the full BEEP, CMUDICT and PRONLEX dictionaries. In this case, the entire dictionary was subjected to the same 10-fold cross-validation process as for the 23 006-word set. This also means, however, that the test and training sets for the experiments using the 23 006 word set and the full dictionaries are necessarily different.

For words with multiple pronunciations, only a single pronunciation was used during training. For G2P, the earliest pronunciation in the dictionary was chosen, while for GP2P the pronunciation pair (one from each dictionary) that gave the best alignment was used. Multiple pronunciations were maintained for testing, however, so that any pronunciation present in the target dictionary was accepted as correct.

Results are given in terms of phoneme accuracy and word accuracy. The generated pronunciations for each of the 10 test partitions were aligned with the dictionary pronunciation by dynamic programming. From these alignments the number of substitutions, insertions and deletions were determined. The phoneme accuracy was subsequently calculated using Equation 1, where N_c, N_i and N_t are the numbers of correct, inserted and total phonemes respectively.

$$Acc = \frac{N_c - N_i}{N_t} \tag{1}$$

Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical. In each case the reported percentages are averages for the 10 cross-validation splits, with 95% confidence intervals calculated using the bootstrap method described in [21].

B. G2P conversion

A first set of experiments applied G2P rules to the individual dictionaries. In the case of SAEDICT, this provides a baseline performance against which later experiments can be compared. Results were obtained for all four dictionaries using the common 23 006-word set described in Section II-B, as well as for the full BEEP, CMUDICT and PRONLEX dictionaries.

Dictionary	Size (words)	Phoneme acc.	Word acc.
SAEDICT	23 006	90.63 $\pm$ 0.53%	59.35%
BEEP	23 006	91.66 $\pm$ 0.54%	65.04%
CMUDICT	23 006	91.24 $\pm$ 0.54%	62.74%
PRONLEX	23 006	92.08 $\pm$ 0.51%	64.93%
BEEP	237 184	93.72 $\pm$ 0.15%	69.97%
CMUDICT	123 646	90.00 $\pm$ 0.26%	59.97%
PRONLEX	90 632	92.14 $\pm$ 0.26%	65.30%

TABLE II
G2P CONVERSION ACCURACIES FOR VARIOUS DICTIONARIES AND TRAINING SET SIZES.

Table II presents the results for these G2P experiments. The first column of the table indicates the dictionary used for training and testing, while the second indicates the total number of words in the training and test data, and hence whether the entire dictionary or the set of 23 006 common word were used. The last two columns give the phoneme and word accuracies respectively, as described in Section IV-A. It is clear from the results in Table II that SAEDICT has the least regular relationship between its graphemes and phonemes.

The parameters determined when optimising against the respective development sets matched those found in earlier studies [1]. For G2P conversion, this entailed a window of three graphemes to the left and four to the right. The size of the decision trees was on average 15532 nodes when using the common set of words. When using the full dictionaries, the trees were considerable larger, varying between 58 982 nodes for PRONLEX and 120 575 for BEEP.

C. GP2P conversion

A second set of experiments applied GP2P conversion to the various source dictionaries, with SSAE as target accent in each case. This was done using the common set of 23 006 words, and using the dictionary pronunciations as input to the converter. While the results of these experiments are not directly applicable to the question being considered in this study, it is useful to know the accuracy of GP2P conversion in isolation, before combining it with G2P conversion. Training and testing was accomplished using the same 10-fold cross-validation approach described in Section IV-A and

Source dictionary	Phoneme acc.	Word acc.
BEEP	96.58 ± 0.31%	81.74%
CMUDICT	95.59 ± 0.34%	76.58%
PRONLEX	96.04 ± 0.32%	78.90%

TABLE III
GP2P CONVERSION ACCURACIES WITH SSAE AS TARGET ACCENT FOR THE 23 006 COMMON WORDS.

Source	Size (words)	Phoneme acc.	Word acc.
BEEP	23 006	88.08 ± 0.63%	54.51%
CMUDICT	23 006	87.29 ± 0.64%	51.91%
PRONLEX	23 006	88.63 ± 0.58%	54.53%
BEEP	237 184	92.21 ± 0.53%	68.98%
CMUDICT	123 646	89.53 ± 0.60%	59.87%
PRONLEX	90 632	91.10 ± 0.54%	62.98%

TABLE IV
COMBINED G2P AND GP2P CONVERSION ACCURACIES WITH SSAE AS TARGET ACCENT.

the same partitions used in the G2P experiments described in Section IV-B.

Table III presents the results of the isolated GP2P experiments. The first column indicates the source dictionary used, and the last two columns give the phoneme and word accuracies respectively, as described in Section IV-A.

The optimal parameters found for GP2P again matched those of earlier studies [1], with a context window of two graphemes to the left and two to the right. The size of the decision trees was on average 2556 nodes, a considerable reduction when compared to G2P.

D. Combined G2P and GP2P conversion

A third and final set of experiments first applied G2P rules to a source dictionary (BEEP, CMUDICT or PRONLEX), and then used GP2P to convert the generated pronunciation to SSAE, as illustrated in Figure 1. Both the G2P and the GP2P conversion was carried out using the 23 006 set of words common to all four dictionaries, with 10-fold cross-validation as described in Section IV-A. The same data partitions used for G2P in Section IV-B were used again to train and test the combined G2P and GP2P approach. In particular, this ensures that the SAEDICT words held out for testing were not present in the training data of either G2P or the GP2P converters.

In order to determine the effect of a larger G2P training set size on the combined use of G2P and GP2P, experiments were also carried out in which the full source dictionary was used to train the G2P converter. This increases the training set available to the source G2P converter to between 90 632 and 237 184 words, as indicated in Table I. In this case, for each of the 10 partitions of the 23 006 common words, the corresponding test set entries were removed from the source dictionary in question. Of the remaining words, 10% were reserved as a development set with which to obtain decision tree parameters, and the remaining 90% were used as training data for the G2P converter. The GP2P converter was however still trained using the set of 23 006 common words, so that the experimental results reflect only the effect of an increase in the size of the G2P training set.

Table IV presents the results of the combined G2P and GP2P experiments. The first column indicates the source dictionary used, while the second indicates the number of words in this dictionary. The last two columns give the phoneme and word accuracies respectively, as described in Section IV-A.

V. DISCUSSION

In this paper, two methods of determining the pronunciation of an unknown word were considered: G2P within the target accent, or G2P within the source accent followed by GP2P to convert the generated pronunciation to the target accent. The latter method was evaluated with G2P rules trained using both a subset of the source dictionary corresponding to words that are common to all four dictionaries, as well as using the full source dictionary.

Before considering the success of these two methods, it is worth examining the results in Table II. From these results it can be seen that for BEEP and PRONLEX, a decision tree trained on the full dictionary is more accurate than one trained only on the 23 006-word common set. This improvement is statistically significant for BEEP. In the case of CMUDICT, however, the G2P performance deteriorates when using the full dictionary. This indicates that, for CMUDICT, there are irregular words present in the full dictionary that are not in the common set.

GP2P conversion on its own gives high conversion accuracies, as reflected in Table III. This would suggest that the accuracies of combining G2P and GP2P conversion should be fairly close to the accuracies of G2P conversion on its own.

When considering Tables II and IV, it is clear that if only the common subset of 23 006 words is used, combining G2P and GP2P leads to a statistically significant deterioration in the accuracy obtained when applying G2P directly to the target dictionary. Applying G2P to SAEDICT gives a phoneme accuracy of 90.63%, while applying G2P and GP2P in succession with PRONLEX as a source dictionary lead to a phoneme accuracy of 88.63% (with lower accuracies with the other two source dictionaries). This occurs because this process is more indirect, involving two consecutive prediction steps, each of which introduces errors, while G2P alone involves only a single step. As the G2P conversion in the source dictionary (when using the 23 006 subset) is only slightly more accurate than that in the target dictionary, the benefit is not enough to compensate for the accuracy lost during GP2P conversion.

Using the full source dictionary to train the G2P decision trees, however, leads to considerably better result. For all three dictionaries, improved word accuracies are obtained relative to the direct application of G2P to SAEDICT. When using either BEEP or PRONLEX as source dictionary, also the phoneme

accuracy is higher, and in the case of BEEP this improvement is statistically significant. For this scenario, the full BEEP dictionary, comprising 237 184 words, is used to train the G2P converter, and gives a phoneme accuracy of 92.21% for SSAE pronunciations.

Previous studies investigating the nature of the errors when converting pronunciations between accents did not reveal any clear, systematic differences [22]. The nature of the prediction errors were not investigated further during this study.

VI. Conclusion

The results indicate that under certain circumstances it is better to generate pronunciations in a well-resourced source accent, and convert those pronunciations to the target accent, rather than using G2P rules to generate them in the source accent. This is especially true for source accents with large dictionaries that have a very regular relationship between graphemes and phonemes, i.e. dictionaries that yield more accurate G2P converters. Furthermore, it can be seen from Table III that GP2P conversion has a relatively high level of accuracy. As a result, increased G2P accuracy when generating pronunciations in the source accent translates into an overall improvement when generating pronunciations in the target accent.

Future research will consider more fully the impact that the size of a dictionary has when training decision trees for G2P, P2P and GP2P conversion.

VII. 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

[1] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different English accents," *Speech Communication*, p. DOI: http://dx.doi.org/10.1016/j.specom.2010.07.006, 2010.

[2] J. Wells, *Accents of English*. Cambridge: Cambridge University Press, 1982.

[3] S. Bowerman, "White South African English: phonology," in *A Handbook of Varieties of English*, E. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds. Mouton de Gruyter, Berlin, 2004, vol. 1, pp. 931 – 942.

[4] I. Bekker, *The Vowels of South African English*. Potchefstroom, South Africa: PhD Thesis, North-West University, 2009.

[5] CMU, "Carnegie Mellon University Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict," 2009, (accessed March 2009).

[6] LDC, "COMLEX English pronouncing lexicon, http://www.ldc.upenn.edu," 2009, (accessed March 2009).

[7] BEEP, "The British Example Pronunciation (BEEP) Dictionary, http://svr-www.eng.cam.ac.uk/comp.speech/section1/lexical/beep.html," 2009, (accessed March 2009).

[8] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. New Jersey, USA: Prentice Hall, 1993.

[9] T. Niesler, P. Louw, and J. Roux, "Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases," *Southern African Linguistics and Applied Language Studies*, vol. 23, no. 4, pp. 459–474, 2005.

[10] V. Wan, J. Dines, A. El Hannani, and T. Hain, "Bob: A lexicon and pronunciation dictionary generator," in *Proc. Spoken Language Technology Workshop*, Goa, India, 2008.

[11] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. ESCA Speech Synthesis Workshop*, Jenolan Caves, Australia, 1998.

[12] J. Suontausta and J. Häkkinen, "Decision tree based text-to-phoneme mapping for speech recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[13] A. K. Kienappel and R. Kneser, "Designing very compact decision trees for grapheme-to-phoneme transcription," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[14] G. Webster and N. Braunschweiler, "An evaluation of non-standard features for grapheme-to-phoneme conversion," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[15] P. Taylor, "Hidden markov models for grapheme to phoneme conversion," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[16] Y. Marchand and R. Damper, "A multistrategy approach to improving pronunciation by analogy," *Computational Linguistics*, vol. 26, no. 2, pp. 195–219, 2000.

[17] M. Davel and E. Barnard, "Pronunciation prediction with default&refine," *Computer Speech and Language*, vol. 22, no. 2, pp. 374–393, 2008.

[18] W. Daelemans, A. van den Bosch, and J. Zavrel, "Forgetting exceptions is harmful in language learning," *Machine Learning*, vol. 34, pp. 11–41, 1999.

[19] K.-S. Han and G.-L. Chen, "Letter-to-sound for small-footprint multilingual TTS engine," in *Proc. ICSLP*, Jeju, Korea, 2004.

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Pacific Grove: Wadsworth & Brooks, 1984.

[21] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, Montreal, Canada, 2004.

[22] L. Loots and T. Niesler, "Data-driven phonetic comparison and conversion between South African, British and American English pronunciations," in *Proc. Interspeech*, Brighton, UK, 2009.