Category-based phoneme-to-grapheme transliteration

Willem D. Basson, Marelie H. Davel
Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
wlmbasson@gmail.com, marelie.davel@gmail.com

Abstract

Grapheme-based speech recognition systems are faster to develop but typically do not reach the same level of performance as phoneme-based systems. In this paper we introduce a technique for improving the performance of standard grapheme-based systems. We find that by handling a relatively small number of irregular words through a phoneme-to-grapheme (P2G) transliteration – transforming the original orthography of irregular words to an ‘idealised’ orthography – grapheme-based accuracy can be improved. An analysis of speech recognition accuracy based on word categories shows that P2G transliteration succeeds in improving certain word categories in which grapheme-based systems typically perform poorly, and that the problematic categories can be identified prior to system development. We evaluate when category-based P2G transliteration is beneficial and discuss how the technique can be implemented in practice.

Index Terms: speech recognition, P2G transliteration, phoneme-to-grapheme rules, grapheme-based ASR

1. Introduction

In an automatic speech recognition (ASR) system, words are traditionally represented as a sequence of acoustic sub-word units such as phonemes [1]. The mapping from these sub-word units to words is usually contained in some form of pronunciation dictionary. The overall performance of ASR systems is strongly dependent on the accuracy of the pronunciation dictionary and best results are usually obtained with hand-crafted dictionaries. Development of these dictionaries is a time-consuming, costly and labour-intensive process, often requiring expert knowledge. If expert knowledge is unobtainable, manually developed or statistical grapheme-to-phoneme (G2P) rules can be used to generalise from small data sets [1]; however, these methods are not available for many languages and typically produce less accurate results than manually developed phoneme-based (P-based) dictionaries.

Earlier work in grapheme-based (G-based) systems has shown that for regular languages – languages that exhibit a fairly close relationship between graphemes and phonemes – P-based dictionary development may be unnecessary, and that the letters of the word can be used directly as the acoustic sub-word units to model [1, 2, 3]. Using G-based sub-word units eliminates the need for expert knowledge and saves time and cost, results in a significantly simplified lexicon definition, and can result in relatively noise-free pronunciation models [4].

In [3] a G-based system was developed using polygraphs, that is, letter-based units constructed from the orthographic word form with arbitrary length left and right contexts. More recent work includes context-dependent G-based recognisers [1], as well as G-based model tying, using a decision tree based on graphemic acoustic sub-word units together with phonetic questions [2].

The regularity of a language can be measured based on G2P consistency: using the average accuracy that is obtained at a specific dictionary size when extracting G2P rules. According to this measure, languages vary considerably, from highly irregular languages such as English, to highly regular languages such as Spanish or Vietnamese, with Afrikaans – the language used as a case study for this paper – being of medium regularity [5].

In earlier work on Afrikaans ASR [6] it was observed that G-based ASR does not reach the same level of performance as that of a system developed using a manually verified dictionary, and that this discrepancy in performance is due to very specific words categories. Word categories such as spelled out words, proper names and foreign words all tend to have a highly irregular relationship between graphemes and phonemes, confusing both G2P- and G-based systems. In this paper we confirm some of the results of [6], using an improved data set, and then attempt to address the discrepancy in performance by ‘regularising’ the spelling of irregular words. Specifically, we experiment with the P2G transliteration of words from problematic categories: creating phone strings for a small subset of words, and then ‘re-spelling’ these with a P2G converter, in effect transforming the original orthography of problematic words to an ‘idealised’ orthography, more suited to G-based modelling.

2. P2G Transliteration Technique

When building a new G-based ASR system our first step is to distinguish between words with regular and irregular pronunciations using word orthography only. While this is not possible for all irregular words, there are some easily identifiable categories (such as spelled words or foreign words) that form the focus of this study. Our aim is to transform the original orthography of these words into an ‘idealised’ form that can easily be incorporated into a G-based system.

We use generic in-language words – that is, words from the ‘regular’ category – to generate crude (broadly applicable but not accurate in detail) P2G rules. In this study, we develop P2G rules using 2nd order joint sequence models (JSMs) [7]. These models are not refined on purpose, as capturing too much spelling detail will cause the P2G model to learn idiosyncratic spellings, resulting in irregular transliterations.

Finally we obtain pronunciations for the words considered irregular and transliterate these pronunciations using the P2G rules. Though pronunciations are still needed for these irregular words, typically their frequency of occurrence is lower than regular words, and a much smaller investment in manual lexicon development is hoped to achieve the same level of accuracy as a fully manually developed lexicon.
3. Experimental Design

We develop comparable G- and P-based ASR systems for different training data sizes ranging from 5 to 40 hours, and use independent test sets and 4-fold cross validation in order to establish baseline performance. (G2P-based ASR is included as a reference.) We then determine whether it is possible to improve the performance of the G-based system using P2G transliteration of specific word categories. We first evaluate the effectiveness of our P2G transliteration technique itself, and then apply it to all words one category at a time. ASR results are obtained per category, before the different transliterations are combined into a single system and performance evaluated.

All test sets are initially decoded using the same flat language model containing all the words in the entire data set. While better recognition accuracy can be obtained using a statistical language model, we specifically want to evaluate the effect of the acoustic models without recognition being guided by a female language model. This means that the systems are evaluated and compared in terms of WER with the only difference between systems being their pronunciation dictionaries. The impact of a language model is revisited in Section 4.4.

3.1. Data preparation and selection

3.1.1. Afrikaans - NCHLT corpus

Afrikaans was selected as the experimental language due to its G2P regularity (fairly regular without being highly regular). A 64.5 hour subset of the NCHLT corpus [8] was selected, consisting of 75 150 utterances from 167 speakers, with a male to female ratio of 48.5/51.5. Every utterance in this dataset passed basic quality control checks; namely, clipping detection, volume detection and speech cutting detection [9]. To ensure a well-balanced dataset, every speaker contributed exactly 450 utterances. A development set of approx. 2.75 hours was reserved for ASR system parameter tuning. The remaining utterances were split into 4 folds with 4 mutually exclusive test sets. The training set of each fold is roughly 46 hours long and contains 54 000 utterances from 120 different gender-balanced speakers. All 4 the training sets are individually partitioned into random segments of approx. 5, 10, 20 and 40 hours.

3.1.2. Pronunciation dictionaries

To ensure a fair category-based comparison between systems, it is important to ensure that the pronunciation dictionary is not only as accurate as possible but also that it is correctly categorised. The same data as used in [6] (based on recrLapd [10]) was used in this study, but all pronunciations and category assignments were manually reviewed by an external phonetician, and corrected as necessary. A small improvement (less than 1% absolute) was obtained when comparing the newly verified pronunciation dictionary with the previous gold standard dictionary. The dictionary contains a total of 9 375 unique words.

Every word is assigned to a category. Categories include (1) generic Afrikaans words, (2) proper names, (3) spelled out words, (4) foreign words, (5) unknown words and (6) a combined category for erroneous words. The erroneous words category contains (a) partial words, (b) concatenations and (c) spelling errors. For this study, words that belong to more than one category are classified as (7) multiple category words and any remaining words are classified as (8) other.

3.2. Technical ASR system description

To evaluate the effect of category-based P2G transliteration, we develop each ASR system using the same relatively standard approach. We use the hidden Markov model toolkit (HTK) [11] and develop context-dependent tied-state acoustic models. 13 Mel Frequency Cepstral Coefficients (MFCCs) with their first and second order derivatives are combined to produce 39-dimensional feature vectors, with cepstral mean normalisation applied at speaker level. With regard to modelling structure, each triphone or trigraph has three emitting states with eight Gaussian mixtures per state and a diagonal covariance matrix.

3.3. Evaluation metrics

Two evaluation metrics, Word Error Rate (WER) and Matching Error Rate (MER) are used, according to their standard definitions [12]:

\[
WER = \frac{S + D + I}{H + S + D} \quad (1)
\]

and

\[
MER = \frac{S + D + I}{H + S + D + I} \quad (2)
\]

where S, D, I and H denote substitutions, deletions, insertions and word hits (number of correct words). As WER is a more standard measure in ASR literature, we use this to report on all results except those in category-based analysis. Since WER does not have an upper bound of 1, we find MER to be a more intuitive metric when performing category-based analysis, as many of the irregular categories produce more insertions than hits.

4. Experiments and Results

A baseline for each ASR system is established in Section 4.1. The P2G transliteration technique is analysed in Section 4.2 and the results when incorporating the transliterated pronunciations are analysed in Section 4.3. Per category results are presented in Section 4.3.1 before the best-performing categories are combined into a single system. A comparative category-based analysis of this system is presented in Section 4.3.2 and the total gain in performance is presented in Section 4.3.3. Finally, the effect of language modeling is explored in Section 4.4.

4.1. Comparing G-based, P-based and G2P-based system performance

To evaluate the effect of category-based P2G transliteration on G-based ASR and to determine how much performance gain is possible relative to P-based ASR, we first establish a baseline for each system. Figure 1 shows the baseline results for G-, P- and G2P-based ASR at four different training sizes of 5, 10, 20 and 40 hours. Comparative to observations in [6], the more training data is available, the less the performance degradation that is observed with G-based ASR becomes. The P-based system outperforms both other systems at all training set sizes. Initially performing second best, G2P-based ASR is overtaken by G-based ASR just after the 20 hour mark. Using a paired $t$-test, nearly all performance differences are statistically significant at the $p=0.01$ level, except between G- and G2P-based ASR at 10 and 20 hours.

4.2. Verifying the transliteration technique

P2G rules are trained using all available generic Afrikaans words. (Comparable rules can be obtained with smaller sub-
Figure 1: Average WER of baseline G-based, G2P-based and P-based ASR systems as training data is increased (prior to transliteration.)

sets as the low order of the JSM models implies that they approach asymptotic performance fairly quickly – this will be investigated further in future work.) In order to evaluate the appropriateness of this P2G approach to the task at hand we first transliterate the entire phonetic pronunciation dictionary and measure the similarity between the transliterated and original spelling. (Note that we are only analyzing the effectiveness of the transliteration approach here – these specific transliterations are not utilised in further ASR system development.) We calculate the difference using PDP scoring[13]: a phone-based similarity measure that produces a value between -1 and 1: 1 indicates that there is no difference between the transliterated and original orthographical form; -1 indicates that every single letter is different. An ideal transliteration mechanism will produce a large number of ‘1’ values for words with a regular orthography, and a small proportion of lower values for more irregular spellings.

Figure 2 depicts the P2G similarity score for each word category: for each similarity score listed on the x-axis, it displays the percentage of words of each category that achieves more than that score on the y-axis. According to this measure, generic Afrikaans words can be regarded as the most regular category with approx. 95% of words having a similarity score of greater than 0.5 and almost 50% of words with a similarity score of 1. In contrast, the spelled word category is the most irregular, with only 10% of words achieving a score of greater than 0.

4.3. Incorporating transliterated forms in the graphemic system

Experimenting with the development set indicated that transliterating specific categories in isolation resulted in a reduction in WER. These categories included: proper names (1), spelled out words (2) and foreign words (3). Proper names and spelled out words tend to have fairly long phone strings, but foreign words may be quite short. As additional pronunciation variants for already easily confusable short words are expected to be detrimental, we only transliterate foreign words with more than 4 characters. (Additional experiments on the development set indicated transliteration of shorter words to be detrimental.)

When transliterating a specific category, two different approaches can be used: the transliterations can either replace the original spelling or be added in addition to the original spelling as a ‘spelling variant’. As observed in Figure 2, a large percentage of the proper names category consists of fairly regular Afrikaans proper names, and therefore the system benefits by keeping their default grapheme strings as variants. It was found that for spelled out words a greater performance gain is achievable when the original grapheme strings are removed and only the transliterations are made available. Note that all initial experiments were performed on the development set, and a transliteration strategy selected prior to obtaining the first results on the evaluation sets. Only a single strategy was implemented once cross-validation started, as described in the following sections.

4.3.1. Improvement per word category

Figure 3 shows the reduction in WER when transliterating categories in isolation. As a reference, the WERs of the default G- and P-based systems are included. The total size of the test set is 207 174 tokens. Transliterating foreign words resulted in the smallest reduction in WER of 0.11% absolute. This is understandable due to the fact that foreign words only make up 0.82% of the test set. Transliterating proper names, that comprise 2.72% percent of the test set resulted in a reduction in WER of 0.34% absolute. The largest reduction in WER of 0.94% absolute is achieved when transliterating spelled out words which comprises of 1.99% of the total test set.
4.3.2. Analysing the effect of word categories

Table 1 provides a detailed view of our findings at 40 hours of training data for G- and P-based ASR as well as a transliterated G-based system consisting of a combination of transliterated proper names, spelled out words and foreign words. Scores are given as a percentage of how many times words from a specific category are mis-recognised out of the total number of words in that category, in other words, category-specific MER. MER percentages are colour coded, with the worst performing system in red, 2nd best in orange and best in green. When compared to the P-based system, G-based ASR performs worse in 4 categories; namely, proper names (1), multi-category words (2), spelled out words (3), and foreign words (4). Combining transliterated categories significantly lowers the MER for each of the individual categories while only slightly increasing the MER of other categories with the worst case being 5.3% absolute for multi category words which is also the worst performing category for transliterated G-based ASR.

It is interesting to note that G-based ASR performs the best of all systems in the generic Afrikaans words category. Though this merits further investigation it is believed this indicates that G-based ASR succeeds in modeling pronunciation variation at the acoustic level. The higher P-based MER might also be caused by unnecessary confusion introduced by pronunciation variants.

Table 1: Size and MER per category for G-based, transliterated G-based and P-based ASR with 40 hours of training data.

<table>
<thead>
<tr>
<th>Category</th>
<th>% Total</th>
<th>graph</th>
<th>g-trans</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic Afr words</td>
<td>89.47</td>
<td>26.1</td>
<td>27.1</td>
<td>28.0</td>
</tr>
<tr>
<td>Proper names</td>
<td>2.72</td>
<td>52.6</td>
<td>47.2</td>
<td>44.7</td>
</tr>
<tr>
<td>Multi-category words</td>
<td>2.16</td>
<td>77.2</td>
<td>82.5</td>
<td>74.2</td>
</tr>
<tr>
<td>Erroneous</td>
<td>2.13</td>
<td>57.7</td>
<td>59.8</td>
<td>64.4</td>
</tr>
<tr>
<td>Spelled out word</td>
<td>1.99</td>
<td>93.7</td>
<td>79.3</td>
<td>77.5</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.82</td>
<td>77.0</td>
<td>65.8</td>
<td>62.4</td>
</tr>
<tr>
<td>Other</td>
<td>0.70</td>
<td>70.9</td>
<td>73.7</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Figure 4: WER for G2P-based, G-based, transliterated G-based, and P-based ASR at 40 hours of training data for baseline systems (utilising a flat language model) and the same systems utilising a basic statistical language model.

4.3.3. Total gain

Combining all the systems that outperformed the baseline G-based system; namely, proper names, spelled out words and foreign words caused a 1.54% absolute drop in total WER, exceeding the sum of its parts (1.39%). This results in G-based ASR performance comparable to that of P-based ASR with a difference of 0.12% absolute between systems. Figure 4 (left-hand-side) shows the total reduction in WER achieved with category-based P2G transliteration. Baseline G-, P- and G2P-based ASR results are included as reference.

4.4. Effect of language modelling

Based on the results in the prior sections, we were interested in determining whether gains from transliteration disappear when a more realistic language model is used. We therefore compare WER when using the flat language model (used in all prior sections) and a basic statistical language model during decoding. A bigram language model with modified Kneser-Ney discounting [14] was developed using the training data of the 40-hour corpus and SRILM [15]. Initial results are presented in Figure 4 (right-hand-side). When using a language model, G2P-based ASR now outperforms the baseline G-based system at 40 hours of training data, with phone-based ASR still performing best. It is encouraging to note that the gains obtained from the transliterated system are retained.

5. Conclusion

In this paper, we proposed a technique for improving G-based ASR by transliterating words from irregular word categories. Our focus is on word categories, as this makes it possible to identify a large percentage of problematic words prior to system development. Identifying irregular pronunciations from orthography alone is not necessarily possible, but foreign words can typically be identified using known word lists in various languages and spelled words typically have a known structure. Similarly, related work from text-based named entity recognition might be useful in identifying possible proper names in transcriptions.

We compared the performance of G-, P-, G2P- and transliterated G-based ASR systems for Afrikaans. The initial baseline experiments showed that as more training data becomes available, at a context-level of three (using triphones or trigrams), minimal effort G-based ASR approaches the performance of P-based ASR. As the remaining discrepancy in WER of G-based ASR is primarily caused by very specific word categories, we demonstrate how these irregular words can be transliterated and incorporated in a G-based ASR system, with much less effort than is required to develop a fully P-based system.

In future work, we aim to investigate the extent to which these results are transferable to different languages. We would also like to determine objective measures for when transliterations should be added as variants, and when the transliterations should replace the original graphemic form of a word. Finally, we are interested in the sensitivity of the proposed technique to different P2G algorithms.

6. Acknowledgements

This work is based upon research supported by the National Research Foundation. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF do not accept any liability in regard thereto. The authors would also like to thank the Center for High Performance Computing (www.chpc.ac.za) for providing sufficient processing resources in support of this work.
7. References


