INTERSPEECH 2009 BRIGHTON

Data-driven Phonetic Comparison and Conversion between South African, British and American English Pronunciations

Linsen Loots, Thomas Niesler

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

linsen@sun.ac.za, trn@sun.ac.za

Abstract

We analyse pronunciations in American, British and South African English pronunciation dictionaries. Three analyses are perfomed. First the accuracy is determined with which decision tree based grapheme-to-phoneme (G2P) conversion can be applied to each accent. It is found that there is little difference between the accents in this regard. Secondly, pronunciations are compared by performing pairwise alignments between the accents. Here we find that South African English pronunciation most closely matches British English. Finally, we apply decision trees to the conversion of pronunciations from one accent to another. We find that pronunciations of unknown words can be more accurately determined from a known pronunciation in a different accent than by means of G2P methods. This has important implications for the development of pronunciation dictionaries in less-resourced varieties of English, and hence also for the development of ASR systems.

Index Terms: English accents, pronunciation modelling, G2P, decision trees

10.21437/Interspeech.2009-74

1. Introduction

The pronunciation dictionary is a key component of any automatic speech recognition (ASR) system. Developing 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 such as South African English (SAE). Nevertheless a dictionary in the speaker's accent can greatly improve ASR accuracy. Grapheme-to-phoneme (G2P) conversion is often used to predict the pronunciations of words not yet in the dictionary, but has limited accuracy and needs a large training set.

This study compares the pronunciations of corresponding words in the General American (GenAm), Received Pronunciation (RP) and SAE accents. The first two are commonly-used and widely studied reference accents for American and British English respectively [1]. The objective is to determine whether British or American pronunciations can be used to complement an SAE pronunciation dictionary. Decision tree methods commonly applied to G2P will be used to analyse the individual dictionaries as well as to convert automatically between the different accents. Decision trees have been used in other studies to generate accent-specific pronunciation variants for ASR [2].

2. Dictionaries

Four dictionaries were used to represent the three accents. CMUDICT [3] and PRONLEX [4] were used for GenAm pronunciations and BEEP [5] for RP. SAE pronunciations were obtained from SAEDICT, a pronunciation dictionary under development at Stellenbosch University. All pronunciations in SAEDICT were transcribed by the same linguistic specialist, ensuring its consistency. Transcriptions were chosen to reflect commonly accepted SAE mother tongue pronunciations. SAEDICT is the smallest of the four dictionaries, currently containing 36 956 entries. The other dictionaries have between 90 000 and 250 000 entries.

2.1. Phoneset

ARPABET was chosen as the common phoneset in which to analyse the four dictionaries. Both BEEP and CMUDICT were already in ARPABET, and PRONLEX used an ARPABET short-hand that could easily be converted to standard ARPABET. SAEDICT was transcribed in a phoneset, based on IPA, developed to describe the languages of Southern Africa [6]. This was converted to ARPABET by means of a mapping based on the closest IPA symbol.

A small number of phonemes present only in a single dictionary were mapped to other phonemes, so as to match the different dictionaries' phonesets exactly. These were mostly diphthongs in BEEP, which were split into their constituent phonemes. PRONLEX contained a syllabic 'n', */en/*, and an aspirated 'w', */wh/*, which were replaced with */ax n/* and */w/* respectively.

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 converted to standard UK spelling. Separate tests were also carried out using US spelling, but the G2P accuracy differed by less than 0.1%. The final set of common words contained 23 034 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 average between 1.11 and 1.22 pronunciations per word. Of the 23 034 words, 9 368 have multiple pronunciations in at least one dictionary.

3. Grapheme-to-phoneme conversion

G2P conversion is the process of predicting a word's pronunciation given its spelling. There are several data-driven G2P methods, including decision trees, HMMs and pronunciation by analogy. Decision trees were chosen for this work, as they are easily implemented and analysed, and give competitive accuracy [7].

Decision trees are a classification method that assigns an input pattern to a class. For G2P, the graphemes and their context form the input and the phonemes the output. Decision trees

thus require a one-to-one correspondence between graphemes (including context) and phonemes. Pronunciations are generated by sequentially passing graphemes (with context) through the tree, and then concatenating the output phonemes. 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 pseudophonemes for the relatively few cases where one grapheme corresponds to two phonemes [8].

The decision trees we used are binary trees with each node containing a true/false question regarding the input. 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 tree's output [9].

Decision trees are grown recursively. For each new node the available training data is split according to all possible questions. The question which results in the greatest information gain is then chosen. Information gain is the difference between a node's information entropy and the weighted entropy of its children [10], where information entropy is given by:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

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 proportions p_L and p_R of t's data associated with each child, the entropy gain is given by [9]:

$$\Delta i = i(t) - p_L i(t_L) - p_R i(t_R) \tag{2}$$

When choosing the question with the largest Δi, $i(t)$ can be omitted with no loss of generality as it remains constant for all questions at a given node. Furthermore p_k can be approximated by N_{t_k}/N_t for a child node t_k and $1/N_t$ is constant for all questions. Finding the maximum of Equation 3 therefore allows the optimal question to be found [11].

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

Decision trees were used with a context of 2 graphemes to the left, 3 to the right, and the 3 most recently generated phonemes. These parameters had been found to give optimal performance in preliminary experiments. G2P conversion took place from right to left. Clusters, used to find questions relating to groups of phonemes or graphemes, were automatically determined using the algorithm described in [11]. Trees were grown to their maximum size, and then pruned using a held-out dataset to improve generalisation [9].

4. G2P-analysis

As a first experiment G2P was applied to each dictionary individually, to gauge how accurately pronunciations could be produced based on training data drawn from the same accent. Of the available data, 80% were used for training, 10% for pruning and 10% for testing. 10-Fold cross-validation was then applied to minimise the effect of the training/test data split.

$$Acc = \frac{N_c - N_i}{N_t} \tag{4}$$

The correct and generated pronunciations were aligned using dynamic programming (DP), allowing the number of substitutions, insertions and deletions to be determined. The phoneme accuracy was then calculated using Equation 4, where N_c, N_i and N_t are the numbers of correct, inserted and total phonemes respectively. Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical.

Dictionary	Phoneme	Word
SAEDICT	88.97%	53.16%
BEEP	89.81%	57.82%
CMUDICT	89.27%	55.68%
PRONLEX	90.35%	57.83%

Table 1: *G2P accuracy for the different dictionaries.*

The results in Table 1 show little variation between the different accents. It is however clear that SAEDICT has the lowest G2P accuracy, while PRONLEX has the best. This suggests that SAEDICT has the least regular relationship between graphemes and phonemes. All the results are comparable to those found in the literature, where accuracies of approximately 91% are common [12, 10].

The most frequently confused phonemes in all accents were found to be */uw/* and */uh/*; */ay/* and */ih/*; */th/* and */dh/*; and the schwa */ax/* with a variety of other vowels. The voiced affricate */jh/* was also often predicted as */g/*. All these correspond to phonemes that are commonly associated with the same grapheme. Additionally, */iy/* is often incorrectly given as */ih/*. Both SAEDICT and BEEP confuse */ey/* and */ae/*.

5. Pronunciation alignments

A second experiment involved a direct phonetic comparison of the different accents, in an attempt to find systematic correspondences. In order to do this, the pronunciations of corresponding words in different accents were aligned by DP.

SAEDICT	r	ih	ae	k	sh	ax	n	s
BEEP	r	ih	ae	k	sh	-	n	z

Table 2: Two aligned pronunciations of *reactions*.

An example of such an alignment is shown in Table 2. Nulls have been introduced where a phoneme has no match in the other string (i.e. insertions and deletions).

Words with multiple pronunciations pose a problem both when aligning the pronunciations of two dictionaries for analysis, and when subsequently using the aligned pronunciations to train decision trees for accent conversion.

The first approach taken was to use only the single pair of pronunciations per word (one from each dictionary) that gave the best alignment. Variations on this approach, whereby the remaining pronunciations were paired according to the quality of their alignments, were also attempted, but these led to deteriorated results.

Two alternate approaches were also considered. The first was simply to discard all words having multiple pronunciations in at least one of the dictionaries. The second was to use a heuristic to attempt to determine the "standard" pronunciation. It is known that for SAEDICT the most common pronunciation is given first, and the same principle is thought to hold for the other dictionaries. While both these alternate approaches affected the results, they had little or no effect on the relative accuracies for the different accents. Removing ambiguous words entirely increased accent-to-accent prediction accuracies

(see Section 6 below) by approximately 0.3%. Using only the first pronunciation reduced the proportion of identically aligned phonemes by approximately 1.2%, and reduced the accent-to-accent prediction accuracy for all accents. The first approach discussed, choosing the best-aligned pair of words, was used for the following analysis.

5.1. Broad comparison

A summary of the results obtained when aligning the various dictionaries is given in Table 3. The phoneme accuracy, calculated according to Equation 4, is given for all phonemes, for just vowels and for just consonants. The last column gives the percentage of words with identical pronunciations.

	Phoneme	Vowel	Cons	Word
SAE - BEEP	92.2	82.7	98.2	59.6
BEEP - SAE	92.1	86.2	95.6	59.6
SAE - CMU	81.6	59.4	95.5	29.4
CMU - SAE	81.6	60.2	94.8	29.4
SAE - PRON	87.2	72.3	96.5	43.0
PRON - SAE	87.3	73.3	95.8	43.0
BEEP - CMU	84.5	69.2	93.7	37.8
CMU - BEEP	84.8	67.1	95.5	37.8
BEEP - PRON	89.3	80.2	94.7	50.6
PRON - BEEP	89.5	77.8	96.6	50.6
CMU - PRON	89.0	73.2	98.6	50.7
PRON - CMU	89.0	73.2	98.6	50.7

Table 3: Accuracies of alignments between dictionaries.

It is clear from Table 3 that the closest matching dictionaries are SAEDICT and BEEP, while the American pronunciations differ to a greater extent from SAEDICT, with CMUDICT differing more strongly than PRONLEX. It was suspected that this was due to CMUDICT's use of */ah/* in many places where other dictionaries (including the American PRONLEX) use a schwa, */ax/*. Recomputing the results while ignoring */ax/* and */ah/* made no notable difference to the results however.

The phonemes of the accents differ most strongly in the vowels. This was clearly visible when comparing the alignments of the vowels and consonants separately. While at least 93.7% of consonants matched between any pair of dictionaries, this figure varied between 59.4% and 86.2% for vowels.

5.2. Phoneme shifts

RP, GenAm and to a lesser extent SAE are well-documented in phonetic and linguistic literature. Wells [1] especially gives a detailed analysis of the different accents. This was compared to the phoneme confusion statistics obtained from the alignments.

5.2.1. Consonants

There is essentially no difference in the consonant systems of the three accents [13]. This is confirmed by Table 3. There are however a few differences in the frequency and environments in which certain consonants are used.

- GenAm is a rhotic accent, whereas RP and SAE are non-rhotic [1]. Rhotic accents articulate the *'r'* in non-prevocalic contexts such as *farm* and *far*, which is silent in non-rhotic accents. In GenAm these words are pronounced */f aa r m/* and */f aa r/*, while in SAE and RP they are */f aa m/* and */f aa/*. In the alignments, some 17% of */r/* phonemes in PRONLEX and CMUDICT are deleted in BEEP and SAEDICT.
- The use of the semi-vowel */y/* after a consonant varies, for example in *tune*, *duke*, *new* and *temperature*. SAE uses this phoneme in almost all cases. RP uses */y/* in most cases, except after */t/* where it sometimes merges to form a */ch/*. GenAm drops the */y/* in all contexts except after labials and velars, such as *cute* and *beauty*. This phenomenon, described by Wells [1] as Later Yod Dropping, manifests itself in the confusion statistics: approximately 25% of */y/*'s in SAEDICT and BEEP are deleted in CMUDICT and PRONLEX.
- The sibilants */s/* and */z/* have a slightly different distribution in SAE when compared with the other two accents: words like *holds* and *levels* contain */s/* rather than */z/*. This is especially prevalent in word-final positions. This change, not mentioned by Wells, occurs for approximately 16% of */z/* sounds in RP and GenAm. A related shift occurs for */dh/*, where some 15% are substituted with */th/* in SAE. Examples include *earthenware*, *wreaths* and *baths*.
- RP uses a syllabic consonant for */l/* and */n/* in words like *bubble* or *sudden*, while GenAm and SAE use a schwa and a consonant, that is */ax l/* and */ax n/* [1]. Of the schwas in SAEDICT, 15% are deleted in BEEP (with comparable results between PRONLEX and BEEP). Of these deletions, almost all occur before */n/* or */l/* (51% and 45% respectively in SAEDICT).

5.2.2. Vowels

Vowels exhibit a much larger and less systematic pronunciation variation than consonants. There are however certain well-known differences.

- The words *cot* and *caught* are pronounced in the same way in GenAm, but not in SAE and RP [1, 13]. This reflects what Wells terms the THOUGHT/LOT merge, where the vowels present in these words have merged in American speech. Our analysis finds that about 12% of */aa/* phonemes in SAEDICT and BEEP become */ao/* in CMUDICT and PRONLEX. A related shift is the approximately 11% of */aa/*'s in BEEP and SAEDICT that become */ae/* in GenAm, such as the vowel in *can't*.
- SAE uses */ih/*, primarily at the end of words, for example *happy* and *cheeky*, but also in words like *barrier*, while other accents use the higher vowel */iy/* [1]. More than 55% of the */iy/* phonemes in the other accents map to */ih/* in SAEDICT.
- Other variations within vowel correspondences are mostly related to */ax/*. These shifts show no obvious structure, but seem to reflect the accents' different stress patterns, as unstressed vowels are frequently weakened to */ax/*.

6. Automatic conversion between accents

As a final experiment we considered the prediction of the pronunciation in one accent using the pronunciation in another. The same decision tree based algorithm used for G2P conversion was used, with the phonemes of the source accent taken as the "graphemes" and those of the target accent as the "phonemes". The only notable difference observed in the functioning of the system was a higher incidence of insertions when aligning two pronunciations. While there are relatively few cases in G2P of a

single grapheme corresponding to multiple phonemes, there are considerably more cases where a single source phoneme corresponds to multiple target phonemes. This was dealt with by allowing the decision tree to use a larger number of pseudo-phonemes (double phonemes treated as a single symbol).

	Phoneme	Word
SAE to BEEP	95.3	76.2
BEEP to SAE	95.1	74.5
SAE to CMU	93.3	66.7
CMU to SAE	93.6	67.7
SAE to PRON	94.2	70.1
PRON to SAE	94.4	71.1
BEEP to CMU	94.8	73.4
CMU to BEEP	95.2	75.4
BEEP to PRON	95.4	75.8
PRON to BEEP	96.1	79.5
CMU to PRON	96.6	80.5
PRON to CMU	97.0	83.8

Table 4: Accuracies of conversion between accent pairs.

The accent conversion accuracies are given in Table 4. The phoneme accuracy (calculated using Equation 4) and word accuracy are both given as percentages.

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 the pronunciations (Table 3) are at least in part due to systematic differences in the transcription conventions followed, rather than fundamental differences in the pronunciations.

SAEDICT and BEEP also have a high conversion accuracy. What is particularly interesting when referring back to Table 1, is that it is substantially more accurate to obtain an unknown pronunciation from an existing pronunciation in a different accent than by G2P methods from the same dictionary. For SAE specifically, BEEP should be used as a first recourse when determining the pronunciation of missing words, since SAEDICT and BEEP deliver high accuracies in Table 4.

Analysis of the confusion statistics produced by the accent conversion process revealed the same patterns described in Section 5.2. The most commonly confused phoneme is the schwa, which is confused with most other vowels, but particularly */ih/*. Other common confusions are between */uh/* and */uw/*, between */z/* and */s/* and between */zh/* and */sh/*. When converting to SAE pronunciations, many */iy/* phonemes are incorrectly predicted as */ih/* and */z/* as */s/*. During conversion to or from American pronunciations, */aa/* is often confused with */ao/* and */ae/*.

7. Summary and Conclusions

This study has investigated properties of and correspondences between three different English accents (RP, GenAm and SAE), as captured by four pronunciation dictionaries (BEEP, CMUDICT, PRONLEX and SAEDICT). The internal consistency of the dictionaries was studied by means of G2P algorithms, and found to be comparable in all four cases.

Common phoneme substitutions, insertions and deletions were determined by pairwise alignment of the dictionaries' pronunciations. This analysis confirmed many of the differences between the accents that have been desribed in the phonetic literature.

Lastly, the G2P algorithm was modified to allow conversion of pronunciations from one accent to another. The results showed that this process is much more accurate than the generation of pronunciations within the same accent using G2P. Hence when the pronunciation of a word is not available, as is fairly common in non-major accents, it is better to first search for it in a different pronunciation dictionary, and if it is found, to convert it to the target accent. G2P methods should be reserved as a last resort.

For the specific case of SAE, our results clearly show that RP pronunciations (such as those in the BEEP dictionary) are most similar. Hence BEEP would provide a better source of unknown pronunciations than an American dictionary. Our results indicate that by using this approach, the pronunciations of almost 80% of words should be correctly predicted, while for G2P this figure drops to below 60%.

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.

8. Acknowledgements

This material is based on work supported financially by the South African National Research Foundation (NRF) under Grants FA2007022300015 and TTK2007041000010.

9. References

[1] Wells, J.C., *Accents of English*, Cambridge University Press, 1982.

[2] Humphries, J.J., Woodland, P.C. and Pearce, D., "Using accent-specific pronunciation modelling for robust speech recogntion", Proc. ICSLP 1996.

[3] CMU. Carnegie Mellon Pronouncing Dictionary. Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed Mar 2009.

[4] PRONLEX, COMLEX English pronouncing lexicon from the LDC. Online: http://www.ldc.upenn.edu/, accessed Mar 2009.

[5] BEEP, "The British Example Pronunciation (BEEP) dictionary". Online: http://svr-www.eng.cam.ac.uk/comp.speech/ Section1/Lexical/beep.html, accessed Mar 2009.

[6] Niesler, T.R., Louw, P. and Roux, J., "Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases", Southern African Linguistics and Applied Language Studies, 23(4): 459-474, 2005.

[7] Ke-Song Han and Gui-Lin Chen, "Letter-to-Sound for Small-footprint Multilingual TTS Engine", Proc. ICSLP 2004.

[8] Damper, R.I., Marchand, Y., Marsters, J.D. and Bazin, A., "Aligning letters and phonemes for speech synthesis", Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA, 2004.

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

[10] Suontausta, J., and Häkkinen, J., "Decision tree based text-to-phoneme mapping for speech recognition", Proc. ICSLP 2000.

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

[12] Black, A. W., Lenzo, K. and Pagel, V., "Issues in building general letter to sound rules", Proc. ESCA Synthesis Workshop, Australia, 1998.

[13] Ladefoged, P., *Vowels and Consonants, Second Edition*, Blackwell Publishing, MA., 2005.