# The Effect of Lexicon Composition in Pronunciation by Analogy

Tasanawan Soonklang[1], R.I. Damper[1], and Yannick Marchand[2]

[1] School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
ts03r,rid@ecs.soton.ac.uk
[2] Institute for Biodiagnostics (Atlantic)
1796 Summer Street, Suite 3900, Halifax
Nova Scotia, Canada B3H 3A7
Yannick.Marchand@nrc-cnrc.gc.ca

**Abstract.** Pronunciation by analogy (PbA) is a data-driven approach to phonetic transcription that generates pronunciations for unknown words by exploiting the phonological knowledge implicit in the dictionary that provides the primary source of pronunciations. Unknown words typically include low-frequency 'common' words, proper names or neologisms that have not yet been listed in the lexicon. It is received wisdom in the field that knowledge of the class of a word (common versus proper name) is necessary for correct transcription, but in a practical text-to-speech system, we do not know the class of the unknown word *a priori*. So if we have a dictionary of common words and another of proper names, we do not know which one to use for analogy unless we attempt to infer the class of unknown words. Such inference is likely to be error prone. Hence it is of interest to know the cost of such errors (if we are using separate dictionaries) and/or the cost of simply using a single, undivided dictionary, effectively ignoring the problem. Here, we investigate the effect of lexicon composition: common words only, proper names only or a mixture. Results suggest that high-transcription accuracy may be achievable without prior classification.

## 1 Introduction

Text-to-phoneme conversion is an integral part of several important speech technologies. The main strategy to determine pronunciation from spelling is to look up the word in a dictionary (or 'lexicon', or 'lexical database') to retrieve its pronunciation, since this is straightforward to implement and yields $\sim$100% accuracy. However, the set of all words of a language is unbounded, so is not possible to list them all. Missing words typically include low-frequency 'common' words, neologisms and proper names, i.e., of people, streets, companies, etc. Thus, there must be a backup strategy for pronouncing unknown words not in the dictionary.

One of the most successful backup strategies (vastly superior to expert rules [1]) is pronunciation by analogy (PbA), which exploits the phonological knowledge implicit in the dictionary of known words to generate a pronunciation for an unknown word. So

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 464–471, 2007.

far, many variants of PbA have been proposed and evaluated with different lexicons. With very few exceptions, previous works using PbA assumed that any missing words tend to be neologisms and so have used a lexicon of common words only. Yet there is a general consensus in the field that knowledge of word class (common word versus proper name) is essential to high-accuracy pronunciation.

In practice, when encountering an unknown word in the input to a text-to-speech (TTS) system, we would not know if it is a proper name or a common word. It should be possible to develop techniques for automatic classification, but these will never be entirely error-free. Therefore, one of several aspects to investigating the performance of PbA is whether or not it makes a difference when the system infers a pronunciation by analogy with a lexicon containing: (1) known common words only, (2) known proper names only, or (3) a mix of common words and proper names.

If high accuracy can be obtained in case (3), then automatic classification of unknown words (with attendant potential for errors) might be avoided. Since PbA infers pronunciations using lexical words most similar (in an analogical sense) to the unknown word, there is a reasonable chance of this. In the best case, the pronunciation of a proper name will be inferred predominantly by analogy with proper names in the dictionary, whereas the pronunciation of a common word will be inferred predominantly by analogy with common words in the dictionary, without having to separate the lexical entries into the two classes in advance. In this paper, we test this possibility, focusing on the effect that lexicon composition has on pronunciation accuracy for PbA.

## 2 Pronunciation by Analogy

An early, influential PbA system was PRONOUNCE, described by Dedina and Nusbaum [2]. Since then, there have been many variants, e.g., [3,4,5,6,7,8], more or less based on PRONOUNCE. The variant of PbA used in this work features several enhancements to PRONOUNCE as detailed in [8]. The pronunciation of an unknown word is assigned by comparing a substring of the input to a substring of words in the lexicon, gaining a phoneme set for each substring that matches, and then assembling the phoneme sets together to construct the pronunciation. As depicted in Figure 1, this process is comprised of four components briefly described as follows.

### 2.1 Aligned Lexical Database

PbA requires a dictionary in which the letters of each word's spelling are aligned in one-to-one fashion with the phonemes (possibly including "nulls") of the corresponding pronunciation. We use the algorithm of Damper et al. [9] for this.

### 2.2 Substring Matching

Substring matching is performed between the input letter string and dictionary entries, starting with the initial letter of the input string aligned with the end letter of the dictionary entry. If common letters in matching positions in the two strings are found,

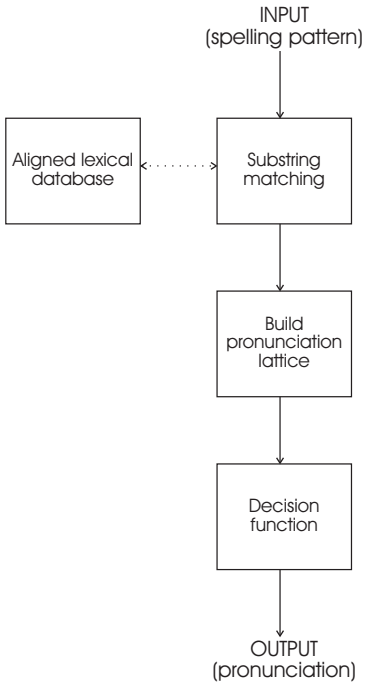


**Fig. 1.** Dedina and Nusbaum's PRONOUNCE

their corresponding phonemes (according to the prior alignment) and information about their positions in the input string are used to build a pronunciation lattice, as detailed next. One of the two strings is then shifted relative to the other by one letter and the matching process continued, until the end letter of the input string aligns with the initial letter of the dictionary entry. This process is repeated for all entries in the dictionary.

### 2.3 Building the Pronunciation Lattice

The pronunciation lattice is a directed graph in which information on matching substrings is used to construct nodes and arcs in the lattice for the particular input string. A lattice node represents a matched letter, $L_i$, at some position, $i$, in the input. The node is labelled with its position $i$ in the string and the phoneme corresponding to $L_i$ in the matched substring, $P_{im}$ say, for the $m$th matched substring. An arc is labelled with the phonemes intermediate between $P_{im}$ and $P_{jm}$ in the phoneme part of the matched substring and the frequency count, increasing by one each time the substring with these phonemes is matched during the search through the lexicon. If the arcs correspond to bigrams, the arcs are labelled only with the frequency. The phonemes of the bigram label the nodes at each end. Additionally, there is a *Start* node at position 0, and an *End* node at position equal to the length of the input string plus one.

### 2.4 Decision Function

Finally, the decision function finds the complete shortest path(s) through the lattice from *Start* to *End*. The possible pronunciations for the input correspond to the output strings assembled by concatenating the phoneme labels on the nodes/arcs in the order that they are traversed. In the case of only one candidate pronunciation corresponding to a unique shortest path, this is selected as the output. If there are tied shortest paths, then the five strategies of heuristic scoring of candidate pronunciations proposed in [8] and [10] are used, and combined by rank fusion to give a final result.

## 3 Lexical Databases

Two publicly-available dictionaries have been used in this work: the British English Example Pronunciation (BEEP) of common words and the Carnegie-Mellon University Dictionary (CMU) of common words and proper names. The former is intended to document British English pronunciations, whereas the latter contains American English pronunciations. We have also studied proper-name and common-word subsets of CMU and mixtures of BEEP and CMU.

### 3.1 BEEP

BEEP is available as file `beep.tar.gz` from `ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/`. It contains approximately 250,000 word spellings and their transcriptions. After removing some words that contain non-letter symbols and/or words with multiple pronunciations, the number of words used in this work is 198,632. The phoneme set for BEEP consists of 44 symbols.

### 3.2 CMU Dictionary

CMU contains both common words and proper names, and their phonemic transcriptions. The phoneme set for CMU contains 39 symbols. The latest version (CMU version 0.6) can be downloaded from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. There are some duplicate words, some containing non-letter symbols and some where the pronunciations obviously does not match the spelling. These were removed to leave 112,102 words. CMU can be partitioned into two subsets as follows.

**Proper Name Subset.** There is no single, easily-available list of proper names and their pronunciations. However, a proper-name dictionary can be developed by using a list of proper names (without pronunciations) together with the standard CMU version 0.6. The list of names can be downloaded as file `cmunames.lex.gz` from `http://www.festvox.org`. It includes the most frequent names and surnames in the USA and their pronunciations [11], from a wide variety of origins. The procedure was simply to extract from CMU pronunciations for the names on the first list. (Note, however, that some names on this list were not found in CMU.) We refer to this extracted subset as Names. The number of proper names in Names is 52,911.

**Common Word Subset.** After extracting the proper names from CMU as above, the remaining words form the common word subset of 59,191 words. We call this dictionary Com.

Finally, we have used a 'Mixture' dictionary; a combination of the BEEP and Names dictionaries. Because different phoneme sets are used by the two dictionaries, we need to collapse the larger of the two (BEEP) onto the smaller (CMU), so that there is a uniform inventory of phonemes. This is the process of harmonisation [1]. Precise details of the harmonisation scheme are omitted for the sake of space.

## 4 Results

Performance was evaluated using a leave-one-out strategy. That is, each word was removed in turn from the dictionary and a pronunciation derived by analogy with the remaining words. Results are reported in terms of words correct, i.e., the number of words for which all phonemes of the transcription exactly match all the phonemes of the corresponding word in the lexicon. Stress assignment has been ignored for simplicity.

Table 1 shows the results of PbA with BEEP, Names and the Mixture dictionary for all combinations of the three dictionaries as test set and lexical database. It should be noted that all entries are significantly different from one another (binomial tests, one-tailed, $p \sim 0$). As can be seen, best results for a given test-set dictionary are achieved when the same dictionary is used as the lexical database. Much higher accuracy is achieved when BEEP is used as the test set and lexical database (87.50% words correct) than when Names is used as the test set and lexical database (68.35% words correct). This is to be expected in view of the diversity of origin of the proper names and different degrees of asssimilation into English [12,13], making their pronunciations harder to infer. Cross-lexicon testing leads to a very large deterioration in performance. Although it is tempting to think that this indicates that proper name transcription is a harder problem than common name transcription, the difference could be due primarily or solely to the different sizes of lexicon, since PbA transcription accuracy is a strong function of dictionary size, increasing as the size of dictionary increases [14].

Using the Mixtures dictionary as both test set and lexical database reflects the practical situation in which no attempt is made to classify the word class, merely treating all words as from the same class. Here the relevant result is 78.08% words correct, a long way below the performance when words from BEEP are pronounced by analogy with the entire BEEP dictionary. Note that a simple weighted linear sum of the BEEP/BEEP and Names/Names results (where the weights are the proportions of the two classes of word) would predict a result of 83.5% words correct, some way above the 78.08% result actually obtained. In effect, this weighted linear sum forms an upper bound on the performance that could be obtained if we had a perfect means of identifying the class of any input word.

In the results of the previous paragraph, the Mixture dictionary is of course heterogeneous, consisting of a British English lexicon of common words (whose phoneme set has had to be harmonised to CMU) and an American English dictionary of proper names. This was done to have the largest possible dictionaries. We have also studied

**Table 1.** Percentage words correctly transcribed by PbA for BEEP, Names and Mixture dictionaries

| | Lexicon | | |
|---|---|---|---|
| Test set | BEEP | Names | Mixture |
| BEEP | 87.50 | 15.93 | 83.62 |
| Names | 23.57 | 68.35 | 55.08 |
| Mixture | 73.34 | 26.62 | 78.08 |

**Table 2.** Percentage words correctly transcribed by PbA with Com, Names and CMU dictionaries

| | Lexicon | | |
|---|---|---|---|
| Test set | Com | Names | CMU |
| Com | 75.67 | 28.20 | 75.94 |
| Names | 38.63 | 68.35 | 51.10 |
| CMU | 64.36 | 39.18 | 72.13 |

the performance of PbA when the three dictionaries (common words, proper names and mixture) are homogeneous, all being derived from CMU. That is, we have used Com, Names and CMU as the three dictionaries. In this case, CMU acts as the dictionary of 'mixtures' (containing both common words and proper names.) Table 2 shows the corresponding results.

Here, the pattern of results is quite similar except for the case of common words tested against the full CMU dictionary. The Com vs. Com result of 75.67% words correct is not significantly different from the Com vs. CMU result of 75.94% words correct (binomial test, two-tailed, $p = 0.876$). That is, extending the lexical database from Com to CMU when testing Com did *not* lead to any deterioration in performance, unlike the corresponding BEEP/Mixture case where there was a large deterioration. We are inclined to believe that the difference is due to the inhomogeneity of the latter (Mixture) dictionary, and the avoidance of harmonisation for Com/CMU. Thus, we give more credence to the results of Table 2 than to those of Table 1.

The very positive Com/CMU result is intriguing. Why does Com vs. CMU, where there is a partial mismatch of the test set and lexicon, perform as well as Com vs. Com, where there is not? It cannot be because proper names are similar in some way to common words with respect to pronunciation by analogy, since this interpretation is denied by all the other results. For instance, there is a huge drop in performance (binomial test, one-tailed, $p \sim 0$) when testing Names against the full CMU dictionary, indicating that proper names have some special characteristics different from common words, as expected from their diversity. The most likely explanation is that PbA is somehow successful in forming strong analogies between common test words and common words in the CMU lexicon, while analogies between these words and the proper names in the lexicon (i.e., the 'wrong' class) are much weaker. This interpretation is currently under investigation.

Let us turn finally to the result of most practical interest; that is, the comparison of Com vs. Com with CMU vs. CMU. This reflects the situation where we have a

single, undivided lexicon in the TTS system. Here, the relevant figures are 75.67% and 72.13% words correct, respectively. This latter figure is almost exactly what we would predict from a weighted linear sum of the Com vs. Com and Name vs. Names results. This is an important finding, since it constitutes compelling evidence for independent errors for the two different classes of word. It supports the working hypothesis of strong analogies between test words of a particular class and lexical entries of the same class and weak analogies between test words of a particular class and lexical entries of the other class. If correct, this means there would be no advantage to attempting automatic inference of input-word class, since the analogy process itself.

## 5 Conclusions

Pronunciation by analogy has been described and tested with different lexicon compositions: common words only, proper names only, and a mixture of the two. Although we attempted to exploit the existence of the large BEEP dictionary, the attempt was complicated by the absence of a list of proper names and their pronunciations for British English. Thus, we believe that our most credible results are those for American English using the CMU dictionary, and common-word and proper-name subsets thereof. In this case, excellent performance has been obtained on the mixture, comparable to that on common names alone. This intriguing result suggests that there may be no need for automatic word class categorisation (common word versus proper name) to be attempted, with its attendant dangers of mis-classification. This interpretation is greatly strengthened by the observation that the result when testing all available words is almost exactly that predicted by a linear sum of the individual word accuracies, weighted by the relative proportions of common words and proper names, respectively, in CMU. As this prediction is based on assuming independence of errors for the two classes of word, it can be viewed as an upper bound on performance for a mixed lexicon.

## References

1. Damper, R.I., Marchand, Y., Adamson, M.J., Gustafson, K.: Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches. Computer Speech and Language 13(2), 155–176 (1999)
2. Dedina, M.J., Nusbaum, H.C.: Pronounce: A program for pronunciation by analogy. Computer Speech and Language 5(1), 55–64 (1991)
3. Sullivan, K.P.H., Damper, R.I.: Novel-word pronunciation: A cross-language study. Speech Communication 13(3-4), 441–452 (1993)
4. Federici, S., Pirrelli, V., Yvon, F.: Advances in analogy-based learning: False friends and exceptional items in pronunciation by paradigm-driven analogy. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'95) Workshop on New Approaches to Learning for Natural Language Processing, Montreal, Canada, pp. 158–163 (1995)
5. Yvon, F.: Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks. In: Proceedings of Conference on New Methods in Natural Language Processing (NeMLaP-2'96), Ankara, Turkey, pp. 218–228 (1996)

6. Damper, R.I., Eastmond, J.F.G.: Pronunciation by analogy: Impact of implementational choices on performance. Language and Speech 40(1), 1–23 (1997)
7. Pirrelli, V., Yvon, F.: The hidden dimension: A paradigmatic view of data-driven NLP. Journal of Experimental and Theoretical Artificial Intelligence 11(3), 391–408 (1999)
8. Marchand, Y., Damper, R.I.: A multistrategy approach to improving pronunciation by analogy. Computational Linguistics 26(2), 195–219 (2000)
9. Damper, R.I., Marchand, Y., Marsters, J.-D.S., Bazin, A.I.: Aligning text and phonemes for speech technology applications using an EM-like algorithm. International Journal of Speech Technology 8(2), 149–162 (2005)
10. Damper, R.I., Marchand, Y.: Information fusion approaches to the automatic pronunciation of print by analogy. Information Fusion 71(2), 207–220 (2006)
11. Font Llitjós, A.: Improving pronunciation accuracy of proper names with language origin classes, Master's thesis, Carnegie Mellon University, Pittsburgh, PA (2001)
12. Vitale, T.: An algorithm for high accuracy name pronunciation by parametric speech synthesizer. Computational Linguistics 17(3), 257–276 (1991)
13. Spiegel, M.F.: Proper name pronunciations for speech technology applications. International Journal of Speech Technology 6(4), 419–427 (2003)
14. Soonklang, T., Damper, R.I., Marchand, Y.: Effect of lexicon size on pronunciation by analogy of English, submitted to Interspeech 2007, Antwerp, Belgium (August 2007)