IMPROVED PREDICTION OF STRESS IN OUT-OF-VOCABULARY WORDS

John Coleman

1. Introduction

Thanks to the availability of inexpensive, large hard disks and affordable semiconductor memory, most currently-available speech synthesis systems employ large pronouncing dictionaries to determine the pronunciation of written words. This method reserves spelling-to-sound rule systems for use as a back-up measure, to be used only for determining the pronunciation of words that are not in the dictionary. (In the remainder of this paper, I shall call such items OOV — that is, "out-of-vocabulary" — words. Though other methods of computing, rather than storing, the spelling-to-sound mapping, such as the use of "neural nets", or pronunciation-by-analogy, are applicable in the same circumstances, I shall use the term "rule-based" to refer to all such methods, as well as to my particular approach.)

The utility of rule-based speech synthesis mechanisms is most evident in applications where the vocabulary is open-ended, which can cause problems even for systems with very large dictionaries. In some of the more forward-looking applications of speech synthesis, such as systems that read email messages or web pages aloud, whether in mobile internet, disabled access tools or other "eyes free" circumstances, OOV words occur extremely frequently. To estimate *how* frequently, I sampled a broad spectrum of web pages as catalogued by the Alta Vista (UK) search engine. At the time of writing, 15 categories (entertainment, business and finance, computers …), each divided into 3 or 4 sub-categories (movies, TV, music …) are listed on Alta Vista's "front page": I examined the top-ranked web site in each sub-category (a fascinating sociological exercise), and determined the proportion of words that were listed in neither the 70,123-word Mitton (1992) dictionary, nor in Weide (1993), which contains a further 102,721 words. The results were: up to 14.5% of OOV words per web page, mean 3.2%, median 2.2%. By any of these measures, OOV words occur often enough that even with a large dictionary, a typical current speech synthesis system would need to rely on its back-up pronunciation rules in order to read almost any web page. I estimate the frequency of OOV words in email messages to be similarly troublesome.

In a closer examination of the kinds of OOV words that are encountered in web pages, the following categories were identified:

a) proper nouns, including foreign names and commercial names (i.e. names of products and businesses, such as *nintendo, Bibendum, Kathmandu, Houllier*)
b) internet neologisms (e.g. *web site, e-mail*)
c) a novel kind of compound word, shown by a specific pattern of capitalisation, often used for commercial names (e.g. *AccountingWeb, CornerStone, myTalk*)
d) conventional compounds (e.g. *centre-forward*); these were counted as OOV because despite their relatively innocuous appearance the stress of compounds is notoriously difficult to determine by rule
e) abbreviations (e.g. *BeOS, Mgt, Int'l*)
f) misspelled and mis-punctuated words (e.g. *lattitude, I?ll, Government''s*)

To synthesize such words, therefore, some kind of computation of their pronunciation will be necessary.

Author's address: University of Oxford Phonetics Laboratory, 41 Wellington Square, Oxford OX1 2JF.
Email john.coleman@phonetics.oxford.ac.uk, Web pages http://www.phon.ox.ac.uk/~jcoleman

Many grapheme-to-phoneme conversion methods have been developed and described previously, and their level of performance evaluated. For example, Damper et al. (1999) and Bagshaw (1998) describe "pronunciation-by-analogy" systems, in which the pronunciation of OOV words is estimated by dividing OOV words into substrings, obtaining the pronunciation of those substrings in words that *are* listed in the dictionary, and putting the substring pronunciations together to obtain a best estimate of the likely pronunciation of the OOV words. Naturally, the pronunciations obtained in this way are less accurate than the 100% pronunciation accuracy of *known* words, because the approach is imperfect, but an approximately correct pronunciation is better than no pronunciation at all, and may even be good enough for the purpose of the system to be fulfilled.

Bagshaw's pronunciation-by-analogy system is noteworthy in that as well as addressing the problem of grapheme-to-phoneme translation, it does not neglect the problem of stress assignment. In contrast, most other studies of deriving pronunciations from text set prosody generation aside as a problem for another day. Bagshaw's system used a set of stress assignment rules which he kindly attributes to an earlier paper of mine (Coleman 1993), which essentially just codified the results of a large body of prior research on the principles of English stress placement. In this paper, I will report on an improved method of stress assignment, based on that earlier work: the improvements arise from the addition of probabilities to the rules, enabling the most probable pattern of stress assignment to be estimated in a principled manner. Rule probabilities are estimated by using the dictionary as a training corpus. As with rule-based systems in general, although the method is far from perfect, the improvements it brings are particularly useful when it comes to the pronunciation of OOV words.

2. Background: linguistic principles of English stress assignment

Before discussing the use of a probabilistic grammar for stress assignment, in this section I shall present a brief summary of what is currently known about English stress placement.

Phonologists distinguish between only three levels of stress within single words. Although many finer variations in loudness can be distinguished, in terms of the linguistic organisation of stress there is no evidence that more than three stress levels are linguistically, systematically distinguished. First, in words of two or more syllables, one syllable is more prominent than all the others: this is called *primary* stress, often marked with an acute accent, e.g. *ínner*, *inért*, *inértial*. In these examples, the syllables which do not have primary stress are all unstressed. Words of only a single syllable may be stressed or unstressed according to their position in the sentence and the speaker's intentions: we shall not deal with either of those issues in this paper. Many polysyllabic words have several stressed syllables, of which one is more prominent than the others. In such cases, a distinction between primary and secondary stress (marked with a grave accent) may be made. For example, *àntidìsestàblishmentárianism* has three secondary stresses, a primary stress, and eight unstressed syllables. Secondary stresses are also found in some two-syllable words, e.g. *pòntóon*: compare the prominence of the first syllable with that of *platóon*, which has an unstressed, short, central vowel. As a result, the two words have quite different rhythms: *pòntóon* is "dum-DUM", whereas *platóon* is "da-DUM". In short, words are made of syllables, and stresses in English are associated with syllables (even though, as a notational convenience, we place stress marks on vowel letters).

In calculating the location of stress, not every part of a syllable is relevant. Each syllable has an onset (the initial consonants), and a rime (the vowel and final consonants). In many languages, including English, stress placement depends in part on the complexity of the rime. Heavy rimes (those with a long vowel, diphthong or final consonant(s)) are preferred for stress assignment, whereas onsets, however big, do not affect stress placement. The facts about the three levels of stress mentioned above are explained by the theory of metrical phonology (e.g. Goldsmith 1990) by reference to the following linguistic principles.

i) Syllables are classified as strong or weak.
ii) A strong syllable and any following weak syllables forms a metrical foot.
iii) Feet are classified as strong or weak.
iv) A strong syllable in a strong foot has primary stress.
v) A strong syllable in a weak foot has secondary stress.
vi) Weak syllables are unstressed.

Thus, the *àntidìsestàblishmentárianism* consists of four metrical feet, each of which begins with a strong syllable: *ànti, dìse, stàblishmen,* and *tárianism*. Of these four feet, the last is more prominent than the first three, so that the strong syllable of the last foot has primary stress, whereas the strong syllable of the initial feet have secondary stress.

The situation is somewhat complicated by the existence of compound words. In this case, each simple word in the compound has a primary stress (e.g. *réader*, *méter*, *fórward*, *céntre*), but with two simple words put together into a compound word, one of the primary stresses is subordinated to the other, and is often transcribed as a secondary stress, as if it has been "demoted". We do not yet understand why in some cases the first word of a compound is the more prominent one (e.g. *méter-rèader*), whereas in other instances the second word is more prominent (e.g. *cèntre-fórward*). A consequence of this uncertainty is that if we encounter an out-of-vocabulary *compound* word in some document, it is very difficult to tell which member of the compound has primary stress and which has secondary stress. Even if both simple words are familiar and listed in the dictionary, we will still have a problem with their combination.

An extensive study of noun–noun compounds by Sproat (1994) detailed a number of tendencies that obtain in the placement of stress in compounds, but unfortunately for text-to-speech applications some of these make reference to properties of the text that are extremely difficult to determine computationally. (Recall that compound words were identified in the introduction as a rich source of OOV words in web pages.) This illustrates a more general point, which is that beside the phonological, structural aspects of stress placement listed above, English stress depends also in part on a variety of other linguistic properties, including i) whether the word is derived according to Latinate or Germanic word-formation principles, or a combination of the two; ii) its part of speech; and iii) specific, idiosyncratic contributions to stress placement made by a small number of morphemes (such as the fact that *–ic* requires the preceding syllable to be stressed, as in *eléctric*, unless it is itself stressed, as in *electrícity*). For the majority of words these factors may be determined, or at least guessed at, if a morphological and syntactic analysis of the input is also conducted. But many OOV words, especially proper nouns, cannot be decomposed into prefixes, suffixes etc., so that the phonemic transcription provided by grapheme-to-phoneme translation may be the only information available to the stress assignment component. In the evaluation of the system described below, therefore, I shall test the accuracy of stress assignment from phonemic transcriptions alone, in the absence of any other kind of linguistic information. Naturally, it is hoped that future research may improve the system's performance by also taking such additional information into account.

3. Formalisation of stress placement principles

The principles of stress distribution set out in the preceding section can be expressed formally as a context-free phrase structure grammar instantiating the following rules.

First, a word of text may be simple, consisting of one "prosodic word", or compound, with several constituent prosodic words. Some compounds are spelled with hyphens, though unfortunately not always. Compare, for example, the spellings "e-mail" and "email".

1) Word → PrWd

2) Word → PrWd PrWd e.g. water-meter

3) Word → PrWd PrWd PrWd e.g. figure-of-eight

Second, a prosodic word consists of one or more feet. In words consisting of more than one foot, one of the feet is more prominent than the others. In this grammar, more prominent feet are indicated by the subscript "s", and less prominent feet by the subscript "w":

4) PrWd → Foot

5) PrWd → Foot_s Foot_w e.g. sécre tàry

6) PrWd → $Foot_w$ $Foot_s$ e.g. sècre tárial

7) PrWd → $Foot_w$ $Foot_w$ $Foot_s$

8) PrWd → $Syllable_w$ $Foot_s$ e.g. ap póintment

9) PrWd → $Syllable_w$ $Syllable_w$ $Foot_s$ e.g. ap poin tée

Third, a foot consists of a stressed ("s") syllable and zero or more unstressed ("w") syllables:

10) Foot → $Syllable_s$

11) Foot → $Syllable_s$ $Syllable_w$ Trochees

12) Foot → $Syllable_s$ $Syllable_w$ $Syllable_w$ Dactyls

13) Foot → $Syllable_s$ $Syllable_w$ $Syllable_w$ $Syllable_w$

Fourth, each syllable consists of an onset and a rime:

14) Syllable → Onset Rime

In addition to these structural rules, there are some others for rarer patterns: a complete listing is given in Coleman (in press). As well as subscripting prosodic categories according to prominence (s/w), rules were also annotated for the position of syllables in the word (initial/medial/final position) and, in stressed syllables, the primary/secondary distinction. Thus, eight subcategories of syllable are in fact used in the complete grammar, such as $Syllable_{s2m}$, secondary stressed syllables that are word-medial, and $Syllable_{s1if}$, primary stressed syllables that are both initial and final, which are monosyllabic words. A long list of onsets and rimes is also provided. The correlation between syllable weight (i.e. rime complexity) and stress is taken care of by allowing the list of stressed rimes (i.e. rimes in stressed syllables) to differ from the list of unstressed rimes. For example, in Mitton (1992), the heavy rime /ɑɪd/ occurs as a rime in the initial syllable of some words, but only if it is stressed: it is listed in my phonological grammar, therefore, as a possible $Rime_{s1i}$ but not as a $Rime_{wi}$.

Using such a context-free grammar, a variety of well-known parsing algorithms can be employed to analyse strings of phonemes. Parse trees obtained in this way show the grouping of phonemes into onsets, rimes, syllables, feet, and prosodic words, and the relative prominence of syllables, that is, stress. If the grammar is correct, the stress assigned to each syllable (even in an OOV word) will be correct. If the stress assigned to some syllable(s) in a phonemic input is incorrect, the structure must be wrong in some way, and hence the grammar must be incorrect. To debug the grammar, therefore, I first took a machine-readable dictionary containing stress marks and used the stress marks to force the correct analysis for each word in the dictionary. In this way, it was ensured that the grammar was complete, though it does not guarantee correctness. In the search for correctness, I then employed the following development cycle:

1) From the machine-readable dictionary containing stress marks (Mitton 1992), a copy was made with the stress marks removed.
2) Every un-stress-marked word was parsed, giving a predicted stress pattern for each word.
3) The predicted stress was automatically checked against the stresses in the dictionary.
4) The errors were manually inspected and the grammar amended in order to remove them.

The cycle was repeated a few times until the parses were all consistent with the standard theory of stress placement, summarised above. Though erroneous stress assignments continued to be generated for some words, those stress assignments were consistent with the theory. Thus, the causes of error lay in the theory, not the specific implementation of the theory. Errors were expected because the grammar is unable to take account of morphology, syntax, and semantics, and in the case of compound words we are in a state of ignorance, as I have already mentioned.

Without stress marks in the input, a parser employing this grammar performs quite poorly: for the dictionary as a whole, it assigns stress correctly in only 45.2% of words, though for a subset of 3,943 capitalized words (e.g. proper nouns and acronyms), it performs rather better, and correctly assigns stress to 67% of that subset. A third test used a set of new words that had not been employed in the debugging cycle, representing the challenge of parsing out-of-vocabulary items. By comparing the Mitton (1992) dictionary to a much larger lexicon (a version of Weide 1993 with transcriptions converted to RP British English), 10,267 "unseen words" were obtained. This set consists mostly of proper nouns not contained in Mitton (1992), variant pronunciations of words in Mitton (1992), as well as some additional words, including new words. The size of this test set is a little over 10% of the size of the dictionary used in the debugging cycle (the "training dictionary", we could say). In this test, stress was correctly assigned to 57.2% of the test words. The results of these three tests give us an appraisal of the general approach to stress assignment: *could do better*.

4. Improvements arising from the use of rule probabilities

A major source of error in the grammar listed above arises from structural ambiguity i.e. the fact that there may be multiple possible parses of a string, some of which give the wrong stress. The strategy adopted in most theoretical studies of English stress is to seek for heuristics in order to limit the number of analyses of each string to exactly 1, that is, to make the theory deterministic. Probably the most important heuristic used in theoretical phonology is the "maximal onset principle", which prefers intervocalic consonants to be parsed in the onset of the second syllable rather than in the rime of the first. For example, /ˈæspɹɪn/ (*aspirin*) could be syllabified as /ˈæ.spɹɪn/, /ˈæs.pɹɪn/, or /ˈæsp.ɹɪn/: the first of these is selected by the maximal onset principle, as it places /spɹ/ in the onset of the second syllable rather than dividing it between the two syllables. In short, if any consonants could either be in the coda or the onset, put them in the onset. The principle was employed in the implementation of the approach described above, in order to reflect and implement the textbook theoretical account of stress assignment.

For this study, however, I examined an alternative solution to the selection of a single analysis when more than one is possible: determine the *most likely parse* on the basis of an empirical estimate of the probability of each analysis. In particular, since rule frequencies are available as a spin-off of parsing the whole dictionary with stress-marks retained, it is easy to associate each rule with a probability estimate (Suppes 1972), reflecting the actual incidence of each rule in a parsed corpus. Although this is not the only approach to probabilistic grammar, it is the simplest probabilistic extension of a relatively straightforward encoding of metrical phonology using a context-free grammar. Other methods currently being investigated in the hope of even better results employ a stochastic "strip grammar", as in Coleman and Pierrehumbert (1997), or "data-oriented parsing" (Bod 1998).

With the addition of rule probability estimates, all possible parses of an input word may be found and ranked by overall probability, so that the most probable analysis can be determined in each case. (The maximal onset heuristic is dispensed with, of course.) In a re-run of the evaluations described above, the probabilistic grammar fares rather better in assigning stress. On test 2, using 3,943 capitalised words, the probabilistic grammar assigned stress correctly to 75% of test words (vs. 67% for the deterministic grammar). On the test of 10,267 OOV words, the probabilistic grammar assigned stress correctly to 69.8% of the test words (vs. 57.2%). For comparison, note that one of the best-performing grapheme-to-phoneme conversion methods (Bagshaw 1998) obtains 65%–83.3% correct assignment of stress, calculated per *syllable*, rather than per word, as here. (Note that scoring per word is less forgiving of errors, as *every* syllable of a word is required to be correctly stressed.) Test 1, evaluation of stress assignment for every word in the dictionary, has not yet been completed. With a probabilistic grammar, it is a much bigger test to run, as all possible parses of every word are computed, not just the single parse of each word that the deterministic grammar provides. Perhaps for commercially viable applications of this approach it will be necessary to employ a sub-optimal parse selection method, such as Viterbi search, but for academic research purposes I am quite happy to wait while the system cranks out all the possible parses of each word! Although it is clear from the test results that we still do not have a 100% correct theory of stress assignment, I have shown that there an improvement over the standard deterministic theory of stress assignment can be demonstrated by obtaining and exploiting rule probabilities.

References

Bagshaw, P. C. 1998. Phonemic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression. *Computer Speech and Language* **12**, 119–142.

Bod, R. 1998. *Beyond Grammar: an experience-based theory of language*. Stanford: CSLI Publications.

Coleman, J. (1993) English word-stress in unification-based grammar. In T. M. Ellison and J. M. Scobbie, eds. *Computational Phonology.* Edinburgh working papers in cognitive science, no. 8. University of Edinburgh. 97–106.

Coleman, J. (in press) Candidate selection. To appear in *The Linguistic Review.*

Damper, R. I., Y. Marchand, M. J. Adamson, and K. Gustafson (1999) Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech and Language* **13**, 155–176.

Goldsmith, J. A. (1990) *Autosegmental and metrical phonology.* Oxford: Blackwell.

Mitton, R. (1992) A computer-readable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English. ftp://ota.ox.ac.uk/pub/ota/public/dicts/710/text710.dat.

Sproat, R. (1994) English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language* **8**, 79–94.

Suppes, P. (1972) Probabilistic grammars for natural languages. In Davidson, D. and Harman, G. (eds.) *Semantics of Natural Language*. Dordrecht: D. Reidel. 741–762.

Weide, R. (1993) The Carnegie Mellon Pronouncing Dictionary [cmudict.0.3]. Pittsburg, PA: Department of Computer Science, Carnegie Mellon University.