

Improving Letter-to-Pronunciation Accuracy with Automatic Morphologically-Based Stress Prediction

Gabriel Webster

Toshiba Research Europe Ltd
Cambridge Research Laboratory
gabriel.webster@crl.toshiba.co.uk

Abstract

Robust text-to-speech (TTS) systems require a letter-to-pronunciation module for generating the pronunciations of words missing from the system lexicon. These pronunciations must specify not only the phone sequence that corresponds to an input orthography, but also the location of lexical stress. However, letter-to-pronunciation modules that make use of a window of context letters around a target letter normally cannot “see” larger-context morphological information that is highly correlated with stress location. The present work demonstrates that by adding a new component that uses morphological information to predict which letter of a word might receive primary stress, and then using the resulting “stressed letters” as input to a decision tree stressed-letter-to-pronunciation component, improvements to both stress accuracy and phone accuracy are obtained in American English, British English, and German. Furthermore, using stressed letters as the input to the decision tree also improves phone accuracy when stress is not required in the output pronunciation, as is conventionally the case for automatic speech recognition (ASR).

1. Introduction

Robust text-to-speech (TTS) systems must be able to generate a pronunciation for any input word. A pronunciation dictionary is inadequate for this task, since the system must be able to generate pronunciations for an indefinitely large set of possible input words. TTS systems therefore require a module for generating pronunciations for words missing from the system pronunciation dictionary. This letter-to-pronunciation module must generate a sequence of phones for an input sequence of letters (often referred to as *grapheme-to-phoneme* or *letter-to-sound* conversion), and for languages with lexical stress must also generate the stress information needed to pronounce the word correctly. A letter-to-pronunciation module is also useful for automatic speech recognition (ASR) systems, because it allows word-level recognition grammars to be input to the system, where they are internally converted to phone-level grammars. In ASR, stress information is typically not necessary.

Letter-to-pronunciation mappings are often learnt automatically via one of several machine learning (ML) algorithms [1], [2], [3], [4], [5]. The present work is concerned with letter-to-pronunciation mapping using decision trees, an ML technique commonly used for this task [1], [2], [3].

Decision trees map from a set of input features to an output class [6]. Training is an iterative process which finds the input feature that partitions the set of training instances

into subsets with the best gain in entropy relative to the unpartitioned set. Then, each subset is taken as the set to be partitioned, and a new best input feature is found. When the entropy gain of the best partition falls below a threshold, no partition is made, but rather a best output class is chosen for that set. The result is a tree that is traversed from root to leaf to map a set of input features to an output class. At every branching node, the input value for the feature specified by that node determines which daughter to choose; when a leaf node is reached, the class specified by that node is output by the tree.

In the case of letter-to-pronunciation conversion, each letter of a target orthography is mapped to a (possibly null) sequence of phones via a separate call to the decision tree. Phones may be “stressed phones” in which there are separate output classes for each relevant combination of phone identity and stress value. The set of input features are the target letter, a context window of the letters directly preceding and following, and possibly the phone sequences output by earlier calls to the tree made for preceding or following letters [3]. Creating the training data thus first requires aligning each letter of each training orthography to a portion of the corresponding pronunciation (see [3] for discussion).

A drawback of generating stress information in this way is that lexical stress correlates with morphology in ways that are difficult to capture using a context window of letters. For example, in English words ending in *-ation* are extremely likely to be stressed on the penultimate syllable, while for words in German that begin with *ver-*, stress is likely to fall on the second syllable (see [7] for a review of stress in English). However, a typical context window length of seven letters ([1], [2], [3]) is not large enough to “see” all of the relevant morphological information. In a word such as *conversation*, the seven-letter context window for the letter *a*, [e,r,s,a,t,i,o], is not large enough to capture the fact that in this case *a* is part of the morphological suffix *-ation*, and the decision as to whether *a* is stressed is therefore made without information that we know to be relevant. Using a larger context window might help solve the problem, but since the cost would be a larger decision tree with more data sparsity, it is worthwhile looking for a more clever solution.

Another possible solution might be to develop a wholly separate model for predicting stress that is sensitive to wider-context morphological information. However, [2] and [3] have both found that using a separate model for stress prediction results in lower accuracy than using a single “stressed phone” model. This decrease in accuracy may be related to the fact that lexical stress interacts with vowel identity. The classic example is vowel reduction in English, where vowels in unstressed syllables are often reduced to [ə], but vowels in stressed syllables are never reduced.

This state of affairs delineates the problem dealt with by the present work. If the letter-to-pronunciation module were to have access to the morphological information lost by the decision tree context window, while still capturing the interaction of stress with vowel identity, improvements might be expected not only in the accuracy of stress prediction, but also in the accuracy of phone sequence prediction.

2. Automatic Morphological Stress Prediction

The present work provides the letter-to-pronunciation module with the morphological information necessary to predict stress by preceding the decision tree step with a new algorithm called Automatic Morphological Stress Prediction (AMSP).¹ AMSP uses the presence of automatically determined pseudo-morphemes to predict which letter of a target orthography has primary stress. The resulting “stressed orthography” (similar to words in which stress is explicitly marked, as in the Italian *città*) is used as the input to an otherwise standard decision tree, and phone sequences with stress information are output. Figure 1 shows the overall module architecture.

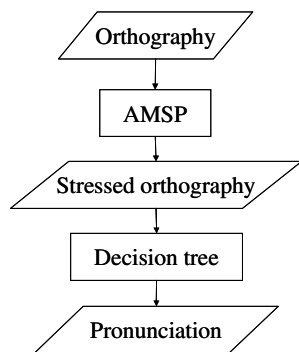


Figure 1: Architecture of the letter-to-pronunciation module with AMSP

Because AMSP predicts which letter should receive primary stress, the first training step is creating training data that consists of stressed orthographies, rather than stressed phone sequences. This is done with the help of the same letter/phone alignments that are used to create the decision tree training data. Letters that are aligned with stressed phones are assumed to be stressed, with the restriction that only vowel letters (possibly accented *a,e,i,o,u*, and *y* when followed by a consonant) may be stressed. In cases where a non-vowel letter is aligned with stress, a preceding vowel letter is stressed instead.

The training algorithm for AMSP then finds correlations between orthographic affixes and stress locations in its training set. A greedy algorithm is used to iteratively find the prefix or suffix that correlates most strongly with stress in a particular location. A minimum frequency threshold is imposed to prevent, for example, affixes that only appear once in the training data, since these always correlate perfectly with the stress of the word they appear in. For prefixes, the location of stress of a training word is counted as the number of vowel letters from the beginning of the word to the stressed

vowel; for suffixes, the location of stress is calculated relative to the end of the word. For example, in English *-ation* correlates highly with stress 3, since the often-stressed *a* in the suffix *-ation* is three vowel letters from the end of words ending in the suffix. In German, *ver-* correlates with stress 2, since words beginning in *ver-* are very often stressed on the second vowel letter (whatever that letter happens to be). Training ends after no prefix or suffix that correlates significantly with stress location is found.

To predict the location of stress for an unknown word, the set of affixes found during training is matched against the word, with multiple matches settled in favor of the affix that had the highest correlation with stress during training. Words that match no affixes are assigned a default stress value, which is the overall most common stress location for the training corpus (1 in English and German).

After AMSP has predicted the location of primary stress, the target orthography and the predicted stress location are then used as input to the decision tree. The location of stress is actually encoded as a separate feature that represents the distance, counted in vowel letters, from the stressed vowel to the target vowel. (Target letters that are consonants receive no value for this feature.) This “distance to stressed vowel” feature makes metrical information available to the decision tree. For example, English meter is based on trochaic (strong-weak) feet, and as a consequence vowels directly adjacent to stressed vowels are much more likely to be reduced than vowels that are two vowels away.

The decision tree is trained on the stresses as assigned by AMSP, rather than on the actual stresses as determined by the letter/phone alignment. In this way, the decision tree training process is presented with incorrect as well as correct stressed letter input, as well as the correct stressed phone output. Since the decision tree is free to output whatever stress location it sees fit, the decision tree can and does try to “correct” errors made by AMSP. Thus AMSP can be seen as generating a preliminary location of stress, which the decision tree accepts or changes as necessary to produce a final prediction of stress location.

It is this twofold prediction of stress that solves the problem described in section 1. To overcome the lack of morphological information in the decision tree’s context window, a different kind of model is used to predict a preliminary location of stress. Then, to capture the interaction between stress and phone identity, the final stress values and phone sequences are predicted in a single model. The notion of “stressed orthography” makes this doubled stress prediction possible.

3. Method

To compare the performance of systems incorporating AMSP to existing systems, the accuracies of four different letter-to-pronunciation systems were compared. The first two systems test whether the addition of AMSP leads to higher stressed phone accuracies. The first system is a standalone letter-to-stressed-phone decision tree (DT) in which a single decision tree predicts both phone sequences and stress, and no AMSP is used. This system is the benchmark system, as it is the type of system that was the most accurate in [2] and [3]. The second system is the system described in section 2 (AMSP-DT), where AMSP is used to predict stressed orthographies,

¹ Patents pending.

and then a stressed-letter-to-stressed-phone decision tree predicts final stress and phone sequences.

The third and fourth systems test the hypothesis that because stress and vowel identity interact, AMSP can also be useful where stress information is not needed in the final output, such as in a typical ASR system. The third system is identical to the DT system except that it outputs phone sequences without any stress values rather than stressed phone sequences (DT-NS). Finally, the fourth system is similar to the AMSP-DT system, but the decision tree maps stressed letters to phones without any stress values (AMSP-DT-NS).

The decision trees were all built using C4.5 [6]. Feature subsetting was turned on, and all other parameters were kept at their default values. The input features to all decision trees consisted of the seven features of a seven-letter context window centered around the target letter, and the phone sequences that were output for the two letters to the left of the target letter (letters were sent to the decision tree in a left-to-right order). In addition, in the AMSP-DT and AMSP-DT-NS systems an extra feature representing the distance to the stressed vowel was added.

Results were calculated for American (U.S.) English, British (U.K.) English, and German. The U.S. and U.K. English dictionaries were corrected versions of the 90,000 word Cambridge English Pronunciation Dictionary [8], and the German dictionary was a corrected version of the 140,000 word Bonn Machine Readable Pronunciation Dictionary [9]. Each dictionary was randomly divided into 80% training data and 10% testing data (the remaining 10% was not used).

Phone accuracy was measured at the word level and at the letter level. At the letter level, a letter was judged correct if it was assigned the same phone sequence that was aligned with the letter in the letter/phone alignment (the same alignments were used for all systems). This measurement of accuracy was used to be able to compare the accuracy of vowel letters and consonant letters, since improvements to stress prediction might be expected to increase vowel letter accuracy more than consonant letter accuracy.

Prediction of secondary stress was not attempted for these results; no secondary stress information was present in the training or testing data. Since each word has exactly one primary stress, stress prediction accuracy was measured at the word level.

In addition to accuracy, the size of each system is also included in the results, since for many applications, particularly embedded systems, the amount of memory required is an important consideration. For the systems without AMSP, the sizes given are simply those of the C4.5 pruned decision trees in a moderately compressed format. For the AMSP systems, the sizes given are those of the decision trees plus the sizes of the AMSP lists of affixes and their corresponding stress values, again in a moderately compressed format. Since data structures can almost always be further compressed or decompressed, in these results the *relative* sizes are of primary interest.

4. Results

The overall word-level accuracies of all four systems are given in Table 1. These accuracies represent the percentage of words for which the entire phone sequence, as well as stress

for the systems outputting stressed phones, were correctly predicted.

	DT	AMSP-DT	DT-NS	AMSP-DT-NS
U.S. English	64.1%	68.2%	70.9%	72.5%
U.K. English	62.8%	67.9%	70.9%	72.2%
German	78.5%	82.9%	84.1%	85.6%

Table 1: Overall word level accuracy

Overall, the systems with AMSP are always more accurate. In the systems outputting stress, in each language AMSP-DT is more accurate than DT by 4%-5% absolute. For the systems not outputting stress, AMSP leads to improvements of 1%-1.5% absolute. These improvements suggest that the addition of AMSP to the letter-to-pronunciation module succeeds in making more useful information available to the decision tree.

More detailed accuracies for these four systems are reported by language in Tables 2-4. These tables give separate accuracies for phone sequence prediction and stress prediction (when relevant) at the word level, the letter-level accuracies of the phone sequences predicted for consonant letters and vowel letters, and the sizes of the systems. For the systems using AMSP, the accuracies of the raw stress predictions made by AMSP are also given.

	DT	AMSP-DT	DT-NS	AMSP-DT-NS
size	580K	634K	449K	492K
raw AMSP stress	-	80.3%	-	80.3%
word level phone sequence	70.1%	71.6%	70.9%	72.5%
letter level: consonants	97.1%	97.2%	97.3%	97.3%
letter level: vowels	88.6%	89.4%	89.1%	89.5%
final DT stress	81.7%	87.1%	-	-
phone sequence/stress	64.1%	68.2%	-	-

Table 2: Detailed results for U.S. English

Several parts of these detailed results help shed light on why the AMSP-DT system is more accurate than the DT system. Firstly, the gains in combined phone sequence/stress accuracy of AMSP-DT are due more to an increase in stress prediction accuracy than to an increase in phone sequence accuracy. In U.S. English (Table 2), for example, word level phone sequence accuracy increases from 70.1% in the DT system to 71.6%, while final stress prediction accuracy increases from 81.7% to 87.1%. The improvement in phone sequence accuracy is about the same as that of the AMSP-DT-NS system over the DT-NS system (1.5% versus 1.6% absolute). Furthermore, in both types of AMSP systems, improvements in phone sequence accuracy are due almost entirely to improvements in the mapping of vowel letters to their phone sequences; the accuracy of consonant letter phone sequences does not appreciably increase when AMSP is added. And finally, in the AMSP-DT systems, large increases

in stress accuracy are seen between the raw AMSP stress prediction and the final DT stress predictions.

	DT	AMSP -DT	DT- NS	AMSP -DT- NS
size	571K	591K	422K	476K
raw AMSP stress	-	80.0%	-	80.0%
word level phone sequence	69.8%	71.9%	70.9%	72.2%
letter level: consonants	97.4%	97.5%	97.5%	97.5%
letter level: vowels	88.8%	89.8%	89.2%	89.9%
final DT stress	80.8%	86.6%	-	-
phone sequence/stress	62.8%	67.9%	-	-

Table 3: Detailed results for U.K. English

	DT	AMSP -DT	DT- NS	AMSP -DT- NS
size	599K	613K	418K	452K
raw AMSP stress	-	81.2%	-	81.2%
word level phone sequence	83.7%	85.4%	84.1%	85.6%
letter level: consonants	98.7%	98.7%	98.7%	98.7%
letter level: vowels	95.0%	95.6%	95.3%	95.8%
final DT stress	88.3%	92.2%	-	-
phone sequence/stress	78.5%	82.9%	-	-

Table 4: Detailed results for German

This pattern of improvement, seen for all languages tested here, supports the idea that the addition of AMSP to the letter-to-pronunciation module supplements the information available to the decision tree with complementary, wider context morphological information that the decision tree alone is not sensitive to. With this extra information the decision tree is then able to “correct” the raw AMSP predictions in many cases, with the net result being higher stress prediction accuracy. This increase in turn allows better prediction of vowel identity due to the interaction between stress and vowel identity.

This interaction between stress and vowel identity is the reason why an improvement in phone sequence accuracy is seen even in the AMSP-DT-NS systems, which do not output stress information. In AMSP-DT-NS, the stress information output from AMSP is used by the decision tree to better predict vowel identity alone, rather than to improve prediction of both vowel identity and stress value.

The addition of AMSP results in small to moderate increases in system size. Whether the improvement in accuracy justifies the increase in size is of course a matter of opinion. However, in U.K. English and German, an improvement of about 5% absolute stress prediction accuracy

at a size increase of just about 3% must be considered a strong candidate for efficient use of extra memory.

5. Conclusion

The results presented in this paper demonstrate that the addition of AMSP to a letter-to-pronunciation decision tree considerably improves the accuracy of primary stress prediction as well as phone sequence prediction. The improvements in phone sequence accuracy hold even for systems that do not output any stress information.

At a higher level, this work demonstrates the effectiveness of using two iterations of prediction to combine two types of models that are sensitive to different, complementary kinds of information. The success of this method suggests that it may be used independently of the specific models and task used for this research. For example, syllabification information might similarly be predicted twice, once by a standalone syllabification model, and then again by a decision tree or other data-driven model that uses syllabified orthographies as input. Thus, this work may be an example of a general technique for many types of task in addition to being a method for improving letter-to-pronunciation conversion.

References

- [1] O. Anderson, R. Kuhn, A. Lazaridès, P. Dalsgaard, J. Haas, and E. Nöth, “Comparison of two tree-structured approaches for grapheme-to-phoneme conversion,” in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1700-1703.
- [2] A. van den Bosch, “Learning to pronounce written words: A study in inductive language learning,” Ph.D. dissertation, Universiteit Maastricht, The Netherlands, 1997.
- [3] V. Pagel, K. Lenzo, and A.W. Black, “Letter to sound rules for accented lexicon compression,” in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 2015-2018.
- [4] L. Galescu and J.F. Allen, “Bi-directional conversion between graphemes and phonemes using a joint N-gram model,” in *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.
- [5] M. Bisani and H. Ney, “Investigations on joint-multigram models for grapheme-to-phoneme conversion,” in *Proceedings of the International Conference on Spoken Language Processing*, Denver, USA, 2002, pp. 105-108.
- [6] J.R. Quinlan, *C4.5: Programming for Machine Learning*. San Mateo, CA: Morgan Kaufman, 1993.
- [7] J. Coleman, “English word-stress in unification-based grammar,” in *Edinburgh Working Papers in Cognitive Science*, No. 8, *Computational Phonology*, T.M. Ellison and J.M. Scobbie, Eds., 1993, pp. 97-106.
- [8] D. Jones, *Cambridge English Pronunciation Dictionary, 15th edition*. Cambridge: Cambridge University Press, 1996.
- [9] T. Portele, J. Krämer, and D. Stock, “Symbolverarbeitung im Sprachsynthesystem Hadifix,” in *Proc. 6. Konferenz Elektronische Sprachsignalverarbeitung*, Wolfenbüttel, 1995, pp. 97-104.