Subjective Evaluation of Techniques for Proper Name Pronunciation

R. I. Damper *Senior Member, IEEE* and T. Soonklang

***Abstract*— Automatic pronunciation of unknown words of English is a hard problem of great importance in speech technology. Proper names constitute an especially difficult class of words to pronounce because of their variable origin and uncertain degree of assimilation of foreign names to the conventions of the local speech community. In this paper, we compare four different methods of proper name pronunciation for English text-to-speech (TTS) synthesis. The first (intended to be used as the primary strategy in a practical TTS system) uses a set of manually-supplied pronunciations, referred to as the 'dictionary' pronunciations. The remainder are pronunciations obtained from three different data-driven approaches (intended as candidates for the back-up strategy in a real system) which use the dictionary of 'known' proper names to infer pronunciations for unknown names. These are: pronunciation by analogy (PbA), a decision tree method (CART), and a table look-up method (TLU). To assess the acceptability of the pronunciations to potential users of a TTS system, subjective evaluation was carried out, in which 24 listeners rated 1200 synthesized pronunciations of 600 names by the four methods using a 5-point (opinion score) scale. From over 50,000 proper names and their pronunciations, 150 so-called *one-of-a-kind* pronunciations were selected for each of the four methods (600 in total). A *one-of-a-kind* pronunciation is one for which 1 of the 4 methods disagrees with the other 3 methods, which agree among themselves. Listener opinions on *one-of-a-kind* pronunciations are argued to be a good measure of the overall quality of a particular method. For each *one-of-a-kind* pronunciation, there is a corresponding so-called *rest* pronunciation (another 600 in total), on which the remaining 3 competitor methods agree, for which listener opinions are taken to be indicative of the general quality of the competition. Non-parametric tests of significance of mean opinion scores show that the dictionary pronunciations are rated superior to the automatically-inferred pronunciations with little difference between the data-driven methods for the *one-of-a-kind* pronunciations, but for the *rest* pronunciations there is suggestive evidence that PbA is superior to both CART and TLU, which perform at approximately the same level.**

***Index Terms*— Text-to-speech synthesis, proper name pronunciation, evaluation of synthetic speech**

I. Introduction

Text-to-speech (TTS) synthesis technology, which has developed enormously over the past decade, enables a voice interface for many commercial applications, especially in telecommunication services. These applications often require the pronunciation of proper names, i.e., the names of people, streets, cities, places, companies, etc. The automatic pronunciation of proper names remains a real challenge for text-to-phoneme conversion, a component part of the pronunciation module in TTS synthesis, because the geographical and language origin of the names can be varied, and the number of distinct names is very large [1], [2]. Additionally, different people will have different opinions about the extent to which a 'foreign' name should conform to the pronunciation conventions of the local speech community, and the preferred pronunciation may vary between 'name owners' (as in the case of *Menzies*, which can be pronounced /mɛnzɪz/ or /mɪŋɪs/). An obvious way to address the problem is to compose manually a dictionary of common names and acceptable pronunciations which is then consulted during synthesis. In general, however, this 'solution' can never be complete as the possibility always exists of encountering (unknown) names in the system input that are not listed in the dictionary. Hence, a secondary or back-up strategy to the primary strategy of dictionary look-up must be devised.

Numerous studies have proposed various approaches to the problem of pronunciation of unknown proper names and have reported the results in different ways. Most studies evaluate the performance objectively in terms of word accuracy, by comparing the pronunciations generated by their models with those in some data set taken as a 'gold standard'. However, the pronunciation should also be acceptable to users of the TTS system, which obviously includes members of the general public with minimal exposure to synthetic speech. Thus, subjective evaluation is also important, particularly for proper names, which can be pronounced differently depending on the linguistic background of the speaker and other cultural factors. Because subjective evaluation is difficult compared to objective evaluation, and cannot deal with large numbers of names, there have been few attempts to do this (but see [3] who have conducted informal subjective evaluation of proper name pronunciations over the internet).

In this work, we compare four different methods of proper name pronunciation for English text-to-speech (TTS) synthesis. The first uses a dictionary of manually-supplied pronunciations; as mentioned above, this is generally the primary strategy in any practical TTS system. The remaining three methods are different data-driven approaches that use the dictionary of (known) proper names to infer pronunciations for unknown names; these are candidates as secondary or back-up strategies for those cases where dictionary matching fails. The three candidate techniques studied are: pronunciation by analogy (PbA), the decision tree method (CART) described by Black, Lenzo and Pagel [4], and the table look-up method (TLU) by van den Bosch and Daelemans [5].

These methods were chosen for several reasons. First, we believe PbA to be one of the best currently-available techniques for inferring pronunciation of common words from

The authors are with the Information: Signals, Images, Systems (ISIS) Research Group at the School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

examples in a dictionary [6] and, second, CART and TLU are very representative of the alternative data-driven techniques to PbA. The TLU method studied here is slightly old compared to the somewhat more up-to-date CART and PbA. So, although better performance would certainly be expected from methods subsequently developed by van den Bosch, Daelemans and colleagues (e.g., those included in the TiMBL 5.1 software package, see `http://ilk.uvt.nl/software.html`), TLU is included as a relatively 'pure' form of table look-up. Since our purpose is to assess the acceptability of the differently-produced pronunciations to potential users of a TTS system, the comparison is primarily subjective rather than objective (in the terms introduced in the previous paragraph), although we do include an objective evaluation for completeness.

The remainder of this paper is structured as follows. In Section II, we briefly describe the three data-driven methods used to derive automatic pronunciations of proper names. Section III describes how the dictionary of manually-supplied pronunciations was constructed. Section IV details the listening tests that were performed to evaluate the acceptability of the pronunciations produced by the dictionary and by the three data-driven techniques, including the means of selecting a reasonable number of names to test from among the large number ($> 50,000$) available in the dictionary. Section V presents the results of the subjective evaluation. For completeness, we also include the results of objective evaluation in Section VI. Section VII concludes.

II. Overview of the data-driven techniques

Data-driven approaches to letter-to-phoneme conversion generally require the letters of each word in the dictionary to be aligned with the corresponding phonemes, so converting the problem of string transduction into one of classification (i.e., the task is to classify any particular letter as belonging to a particular phoneme class.) In this work, the algorithm described in [7] was used for alignment.

A. Pronunciation by analogy

The assumption underlying PbA is that the dictionary contains implicit phonological knowledge which can be exploited to generate a pronunciation for an unknown word. An early and influential PbA system was PRONOUNCE [8], and many variants of PbA have since been based on it (e.g., [9], [10], [11], [12], [13]). The particular variant used here is now described.

When an unknown word is presented as an input to the system, so-called full pattern matching (see [12]) between the input letter string and dictionary entries is performed, starting with the initial letter of the input string aligned with the end letter of the dictionary entry. If common letters are found in matching positions in the two strings, 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.

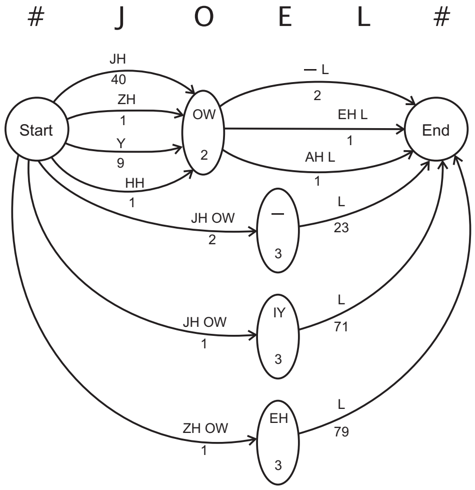

Fig. 1. Example of simplified pronunciation lattice for the name *Joel*. For clarity, only a subset of the nodes/arcs is shown. The phoneme symbols used are those of CMUDICT—see later.

The pronunciation lattice is a directed graph that defines possible pronunciations for the input string, built from the matching substring information. A lattice node represents a matched letter, L_i, at some position, i, in the input. The node is labelled with its position i and the phoneme corresponding to L_i in the matched substring, P_{im} say, for the mth matched substring. An arc is labelled with the phonemes intermediate between P_{im} and P_{jm} $(j > i)$ 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. Arcs are directed from i to j. If the arcs correspond to bigrams, the arcs are labelled only with the frequency. (The string of phonemes intermediate between P_{im} and P_{jm} is empty.) Phonemes P_{im} and P_{jm} label the nodes at each end of the arc, i.e., i and j respectively. 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.

Figure 1 shows an example pronunciation lattice for the name *Joel*. For clarity, only nodes and arcs contributing to the shortest path are shown, since only these affect the decision on pronunciation (see next paragraph). The lattice was built by removing *Joel* from the dictionary and then matching it against the remaining entries. This *leave-one-out* strategy is a very convenient way to derive pronunciations for all the words in a given dictionary, and thereby to assess performance; it is also known as k-fold cross validation, where k is the size of the dictionary. The candidate pronunciations are not necessarily distinct. Different shortest paths can obviously correspond to the same phoneme string, as in this example.

Finally, the decision function identifies the 'best' candidate pronunciation of the input according to some criterion. Possible pronunciations for the input correspond to the string assembled by concatenating the phoneme labels on the nodes or arcs in the order that they are traversed in moving through the lattice from *Start* to *End*. If there is just one candidate corresponding to a unique shortest path, this is selected as the output. If there are tied shortest paths, then the combination of three of the scoring strategies in the multi-strategy approach of [12], [13] is used to select the output. These scoring strategies are:

1) the maximum product of the arc frequencies along each of the shortest paths;
2) the maximum frequency of the same pronunciation within the shortest paths;
3) the maximum 'weak link' value, where the weak link is the minimum of the arc frequencies on a given path.

These were chosen because their combination out-performed other combinations of strategy for the pronunciation by analogy of words (not proper names) in [13]. The individual scores for each candidate were ranked for each strategy, and points allocated according to this rank (a form of Borda counting [14]). The points were then multiplied to produce a final score for each candidate and the best-scoring one selected as the output.

Prior to selecting a subset of pronunciations for subjective testing (see Section IV-A below), pronunciations were produced for *all* names in the dictionary. (This also served as the objective evaluation, see Section VI.) It is, of course, essential that these pronunciations are for 'unknown' words (i.e., 'unseen' in the sense of not being available to the analogical inferencing process). For PbA, this is easily achieved using the very simple leave-one-out strategy described above.

B. Decision tree (CART)

Since letter-to-phoneme conversion (of aligned strings) is a classification problem, decision trees have long been used for this purpose. Perhaps the earliest such attempt was that of Lucassen and Mercer in 1984 [15]. In this work, we have used the CART algorithm of Black et al. [4], where CART stands for *Classification And Regression Tree* (see [16]). The CART approach uses feature data to predict class membership. In this case, the feature data are the letters in a fixed-length context window and the class is the phoneme corresponding to the central, 'target' letter. Since the letter and phoneme strings are aligned, a one-to-one correspondence between the target letter and the class (i.e., the corresponding phoneme) can be assumed.

The first step in building a decision tree is to create from each word of the training data a set of learning vectors containing n-grams (context windows consisting of strings of n letters, where n is odd) and the corresponding phoneme for the middle letter of each. As an example, Table I shows all 7-gram learning vectors for the name "Aardema" whose pronunciation is /AA R D EH M AH/ (according to the CMUDICT phoneme set used in this work—see later).

Next, the learning vectors for each word in the training dataset are fed as the input into CART to create the decision tree. In the training process, predictions about the phonemes corresponding to earlier letters in the word can be used in the word to make decisions about the phoneme corresponding to the current, target letter; this is called phonemic feedback.

TABLE I

COMPLETE SET OF 7-GRAM LEARNING VECTORS FOR THE NAME "AARDEMA", PRONOUNCED /AA R D EH M AH/. THE FIRST ELEMENT OF THE VECTOR IS THE PHONEME CORRESPONDING TO THE 4TH LETTER IN THE 7-GRAM.

–	–	–	–	A	A	R	D
AA	–	–	A	A	R	D	E
R	–	A	A	R	D	E	M
D	A	A	R	D	E	M	A
EH	A	R	D	E	M	A	–
M	R	D	E	M	A	–	–
AH	D	E	M	A	–	–	–

Thus, there are two parameters in CART that we need to fine-tune, namely the width n of the context window and number of previous phonemes, P, to be used in phonemic feedback.

Again, it is obviously important that the pronunciations used in the subjective tests are 'unseen'. To ensure this, we trained and tested the trees using leave-one-out on the entire set of dictionary words (52,911 names, see Section III below) prior to selection of a small subset of names for testing. This also served as the objective evaluation of CART, see Section VI. Leave-one-out was enormously expensive computationally (approximately one month elapsed time) since the trees had to be rebuilt over 50,000 times. This was done with parameters $n = 9$ and $P = 3$, which were the best performing in a prior 10-fold cross-validation.

C. Table look-up (TLU)

Daelemans and van den Bosch [5] describe a simple table look-up procedure with some default tables that are invoked in the case of matching failure so as to achieve generalisation ability. During table construction, all unambiguous one-to-one letter-to-phoneme mappings are found and stored in the 0-1-0 subtable. Then, the width of the letter window is expanded on the right by one character, and all unambiguous 0-1-1 patterns found and stored in the 0-1-1 subtable, excluding those patterns already in the 0-1-0 subtable. Then, the window width is expanded on the left by one character and the procedure repeated. The process of expanding the window on right or left and storing all the patterns that have not been stored in the earlier table continues until all patterns in the training set are compressed in the look-up table. In this work, this occurs with a 10-1-10 window. Additionally, two default tables are assembled to provide generalisation ability. The first default table, referred to as the *best-guess* table, contains all occurring 1-1-1 patterns and their most frequently occurring phonemic mapping. The second table, referred to as the *final-guess* table, contains all letters and their most frequently occurring phonemic mappings.

The conversion algorithm starts by searching for a matching letter pattern for each letter of an input word in the 0-1-0 subtable. Note that, if found, this is guaranteed to be unambiguous. If no pattern is matched, each (single-letter) pattern is extended to a 0-1-1 pattern and the 0-1-1 subtable is then searched. This is repeated until a matching pattern with a minimal extension is found and the corresponding letter-phoneme mapping is returned. If no match can be found, the best-guess table is scanned to return the 'best' mapping. If table look-up fails again, the default phoneme of that letter from the final-guess table is returned. Finally, all phonemes are concatenated to create the pronunciation of the word.

As with the other data-driven methods, it was important that pronunciations corresponded to 'unseen' words. This was achieved as follows. During table creation, we noted the specific words used to derive each pattern (i.e., table entry). If the pattern was unique, in the sense that only one word exhibited that pattern, this pattern was ignored in deriving a pronunciation for that unique word. This was the only adjustment made. Strictly, the frequency of the mappings ought also to be reassessed, but for simplicity (and because we believe it had little effect on results) this was omitted. Again, this was done for all 52,911 names in the dictionary, and also served as the objective evaluation of TLU, see Section VI.

III. Dictionary of Name Pronunciations

Unfortunately, no single, easily-available list of proper names and their pronunciations (as judged suitable for an English-speaking context) exists. A file of proper names without pronunciations can be downloaded from `http://www.festvox.org/cmunames.lex.gz`. However, the pronunciation of many of these names can be found in CMUDICT, obtainable from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, where the phoneme set, consisting of 39 phoneme symbols, can also be found. The phoneme set While CMUDICT contains large numbers of 'common' words (as opposed to proper names), it also includes the most frequent names in the United States (as of some 20 years ago) and their pronunciations, from a wide variety of origins [17]. We were thus able to produce a dictionary of 52,911 names present in `cmunames.lex` and having a pronunciation specified in CMUDICT. The listed pronunciations were taken to be the correct 'gold standard' for automatic inference of pronunciations using the three data-driven methods. For this reason, we will use these symbols in preference to those of the International Phonetic Alphabet [18] from this point.

For TLU, no unambiguous 0-1-0 patterns were actually found in CMUDICT. Nonetheless, the general description of the method in Section II-C above remains valid.

IV. Listening tests

The listening tests were designed to assess the acceptability of the proper name pronunciations produced by the four methods (manually-specified via CMUDICT, hereafter 'CMU', and the three different data-driven methods) to potential users of a text-to-speech system for English. Although separately evaluated as such, we do not intend the dictionary pronunciations to be considered as exactly equivalent to the data-driven pronunciations for two reasons. First, dictionary look-up would be used in a practical system in conjunction with one or other data-driven method (in the way of a primary and a back-up strategy). Second, the data-driven methods use the dictionary as a knowledge base from which to make inferences about an acceptable pronunciation of 'unknown' names, so the two (dictionary and automatic inference) are clearly neither equivalent nor independent.

TABLE II
Possible patterns of disagreement/agreement in pronunciations produced across the four methods. The number of "two pronunciations same, remaining 2 same" cases is reduced by a factor of 2 because of symmetry.

Pattern	Number of cases
All pronunciations same, zero disagreement	$C(4,0)=1$
Three pron. same, remaining 1 different	$C(4,1)=4$
Two pron. same, remaining 2 different	$C(4,2)=6$
Two pron. same, remaining 2 same	$\frac{C(4,2)}{2}=3$
Zero pron. same, zero agreement	$C(4,4)=1$
TOTAL	15

A. Selection of test stimuli

It is obviously not practical to expect listeners to rate all 52,911 pronunciations in our dictionary, so some principled way of selecting among the large number of names to produce a much smaller subset suitable as test stimuli is required. This is not especially easy to do for two main reasons:

1) In spite of the CMUDICT pronunciations being taken as the 'gold standard' for the automatic inference of pronunciations using the three data-driven methods, we did not wish to assume that the dictionary pronunciations are actually correct. (There are, in fact, some very obvious errors in the dictionary.) The CMU method is to be assessed via the listening tests on the same basis as PbA, CART and TLU. This means that we cannot identify 'errors' in pronunciation as such. Rather, pronunciations are to be treated as more or less acceptable to potential users. So we cannot use any notion of error in selecting stimuli for presentation to listeners.
2) Since we have four methods of arriving at pronunciations, the number of possible patterns of disagreement/agreement across the four is quite large. In fact, consideration of the combinatorics shows that there are 15 different possible patterns (Table II). Figure 2 shows the proportion of pronunciations exhibiting each of these different possible patterns of (dis)agreement. If we were to select names as stimuli at random from within the total of 52,911, there would be no control of the pattern of disparity across methods, rendering meaningful analysis of the results virtually impossible. It seems preferable to focus on just one of the patterns of (dis)agreement, i.e., just one of the rows of Table II.

For these reasons, especially 2), we decided to use as stimuli a randomly-sampled subset of those names on which only one of the four methods disagreed. This corresponds to the second row of Table II for which there are 4 cases listed, one for each of the 4 methods. This is the only pattern of (dis)agreement giving clear separation of methods. Those pronunciations for names in the sampled subset which were produced by just one of the four methods (the others agreeing) we will henceforth

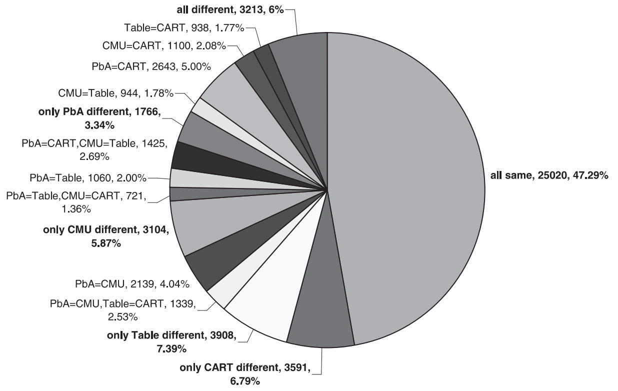

Fig. 2. Pattern of (dis)agreement across the four methods.

call *one-of-a-kind*. Those pronunciations on which the remaining 3 of the 4 methods (i.e., the rest) agree we will henceforth call *rest*. The reasoning behind this choice is that, given that we have no strong basis for deciding what an error is, name pronunciations on which 3 out of 4 methods agree are likely to be acceptable to listeners, whereas the corresponding *one-of-a-kind* pronunciations are likely to be representative of the worst pronunciations made by the particular method delivering the deviant output.

Thus, we randomly selected 150 names with *one-of-a-kind* pronunciations for each method, a total of 600. For example, we selected 150 names from the 3,591 names for which CART produced a pronunciation at variance with the other three methods, which agreed (see 'only CART different' in Figure 2). As well as these 600 names with *one-of-a-kind* pronunciations, we also had the same set of 600 names with a corresponding *rest* pronunciation, a total of 1200 stimuli.

B. Synthesis

Speech output was by diphone synthesis using Festival, a public domain system intended for speech synthesis research available from `http://www.cstr.ed.ac.uk`. To achieve good pronunciations, it was necessary to syllabify the words and add stress. Syllabification was done using the Festival function `lex.syllabify.phstress(PHONELIST)`. To obtain stress patterns, we passed the spelling patterns through Festival's letter-to-sound rules [4] and then manually transferred the stress pattern of the output to the 1200 stimuli. The voice used was male KAL.

C. Test conditions

Subjects listened to the 1200 synthesised names over headphones in the soundproof room in our laboratory at Southampton. We decided firmly against online internet-based evaluation, as used by Font Llitjós and Black [3], because of the inherent lack of experimental control with that approach. Subjects saw the names printed on a sheet as well as hearing the synthesised version. They were instructed to rate their opinions on the quality of the pronunciations on a five-point scale, according to the mean opinion score (MOS) [19] as follows:

1 Unacceptable, no one would ever pronounce this name like this
2 Poor, very few people would recognize the name from this pronunciation
3 Acceptable, some people would recognize this name
4 Almost correct, most people would recognize this name
5 Correct, almost everybody would be able to recognize this name

Listeners wrote their ratings next to the printed names. They were given printed instructions at the outset. Subjects were informed that apostrophes in names (e.g., "O'Connell") had been removed. They were further told that they would hear one name every two seconds, and be given a short break at the end of each page of (30) printed names, after which they should press a button to signal that they were ready to continue. They were asked to assess the pronunciations as if they had been produced by a telephone directory assistance service.

The complete experiment took some 70–80 minutes per subject. It was quite demanding of listeners both in time and concentration. We therefore feel that the number of names selected for the experiment (600 *one-of-a-kind* and 600 *rest*) is effectively a practical maximum.

D. Subject profiles

Since MOS values generally become more stable as the number of listeners increases, a reasonably large pool of subjects should be used, balanced by the time and cost of testing; Clark [20] suggests using at least 16 subjects. Accordingly, 24 listeners (15 male, 9 female) took part in this study. All were students at the University of Southampton, and native speakers of British English, aged between 18 and 31 years. Of the 23 subjects who responded to a prior question about their familiarity with synthetic speech, 2 were "very familiar", 14 were "somewhat familiar", and 7 were "not at all familiar". (Note that one of the 24 subjects omitted to answer this question.) Listeners were paid a small sum (£20 each) for their participation.

Subjects were warned that they would hear 'Americanized' names. We were well aware of the disparity between our subjects' version of English (British) and that of the materials to which they were exposed (American English) but this was unavoidable. There are no corresponding materials easily available for British English, and we did not have access to a pool of American English subjects. We do not believe this had a serious effect; young educated British people generally have considerable exposure to popular American culture.

V. Results of Subjective Evaluation

In this section, results in terms of MOS values for *one-of-a-kind* pronunciations and *rest* pronunciations (on which 3 of the 4 methods agree) are presented. Given the justification in Section IV-A for the selection of names to test, it is reasonable to view high MOS values for *one-of-a-kind* pronunciations as indicative of the overall quality of a method, since they are likely to be representative of the worst pronunciations produced by that method. Things are not so simple, however, for the *rest* pronunciations. Here, the pronunciations are representative of the general quality of the 'opposition', i.e., the three competitor methods[1]. So high MOS values for the *rest* pronunciations can be taken to indicate that the corresponding method is problematic, or at least inferior to the opposition. In other words, if a particular method produces pronunciations which are rated highly by subjects but which are different to pronunciations produced by the remaining methods, these latter being rated poorly by listeners, then there is a sound basis for considering this method to be superior to its competitors. It is necessary to emphasize that we are referring here to relatively high/low values *across* methods. We expect the *one-of-a-kind* MOS values to be relatively lower than the *rest* values *within* methods, just because the former are likely to be representative of the worst pronunciations produced, as argued previously.

Before describing the results quantitatively, we highlight some example pronunciations for purposes of illustration. Table III shows the worst and best pronunciations, both *one-of-a-kind* and *rest*, in terms of obtained MOS values for each of the four competitor methods. (The reader should note that here the 'mean' is for a particular pronunciation across all 24 listeners.) Not only do these examples give an idea of the range of quality of pronunciations produced, there are also some interesting trends evident. The expectation that the *one-of-a-kind* MOS values should be generally lower than the *rest* values within methods is confirmed, at least for these best/worst pronunciations. The relatively high MOS values for *one-of-a-kind* coupled with relatively low values for *rest* pronunciations for CMU indicate that the dictionary pronunciations are generally superior. This is to be expected, since the dictionary forms the 'gold standard' knowledge base from which the data-driven methods infer pronunciations. There is no obvious basis on which the inferential process could *improve* on the dictionary pronunciations.

A. *Results for* one-of-a-kind *pronunciations*

General statistics of MOS values for these pronunciations are listed in Table IV. (The reader should note that here the 'mean' is for a particular method across all *one-of-a-kind* pronunciations and all 24 listeners.) MOS values exceed 3.7 (i.e., tending towards 'Almost correct') indicating that all methods gave reasonable pronunciations. As expected, dictionary pronunciations have a higher MOS than any of the automatically-inferred methods; since the dictionary acts as the knowledge base for the data-driven methods, there is no obvious basis on which they could systematically improve on CMU. From best to worst in terms of MOS of *one-of-a-kind* pronunciations, the data-driven methods are ordered PbA, then CART, then TLU.

TABLE IV

MEAN OPINION SCORES FOR SUBJECTIVE EVALUATION BY 24 LISTENERS OF 600 ONE-OF-A-KIND PRONUNCIATIONS.

	Mean	Std. Deviation
CMU	4.0372	0.5443
PbA	3.8319	0.6760
CART	3.7790	0.7285
TLU	3.7305	0.7157

Since our data are not interval data but ordinal[2] and, as a consequence, can not be normally-distributed, and because many statisticians believe that means are an inappropriate measure of central tendency for ordinal data (although the MOS measure already violates this belief), we believe non-parametric statistical analysis [21] is appropriate. Specifically, we have used the Wilcoxon signed-rank test.

Table V shows the order in which each of the 24 listeners ranked the *one-of-a-kind* pronunciations produced by each of the 4 methods according to the corresponding MOS values. It is very striking that all 24 consistently rated the CMU pronunciations first. The data-driven methods have been ranked fairly consistently in the order PbA, CART, TLU; 10 out of 24 listeners ranked the methods in this order. Table VI shows the results for the Wilcoxon signed-rank test applied to these data. The only significant differences in the opinions of the listeners are between CMU and each of the data-driven methods ($p < 0.05$, 2-tailed, df $= 149$), with CMU superior to the others.

B. *Results for* rest *pronunciations*

General statistics of MOS values for these words are listed in Table VII. (The reader should note that here the 'mean' is for a particular method across all *rest* pronunciations and all 24 listeners.) As explained and justified earlier, we expect the MOS values to be generally higher for *rest* pronunciations than for *one-of-a-kind* pronunciations (because the latter are representative of the worst pronunciations produced by a given method), whereas relatively high MOS values for *one-of-a-kind* pronunciations and relatively low values for *rest* pronunciations for a given method are taken to be indicative of the quality of that method. As the table shows, MOS values generally exceed 4 (i.e.,'Almost correct') and there is indeed a strong tendency towards higher MOS values than for *one-of-a-kind* pronunciations.

Comparing Table VII with Table IV earlier, we see (as expected) an almost inverse or 'complementary' pattern of results. The CMU *rest* pronunciations have the lowest MOS value, with the remaining three data-driven methods ordered from lowest to highest as PbA, TLU, CART with the latter two almost indistinguishable. Figure 3 shows a composite plot of the variation of MOS across the four methods for both *one-of-a-kind* and *rest* pronunciations, making the 'complementary' nature of the results very clear. In interpreting

[1]In this paper, we use the words *competitor* and *competition* to refer indiscriminately to some or all of the 4 techniques studied; the word *candidate* is used to refer to the some or all of the 3 data-driven methods as they are candidates for the back-up strategy in a practical TTS system.

[2]Interval data are continuous data where differences are interpretable, but here there is a mapping from listener opinion to five discrete, arbitrary categories. Hence, it is problematic to interpret the difference between mean opinions scores of, say, 3.45 and 3.55. There is, however, a natural and logical ordering of the categories, so our data are ordinal.

TABLE III

EXAMPLE PRONUNCIATIONS, SHOWING THE WORST AND BEST PRONUNCIATIONS, BOTH *one-of-a-kind* AND *rest*, IN TERMS OF OBTAINED MOS VALUES FOR EACH OF THE FOUR COMPETITOR METHODS.

Method	name	*one-of-a-kind* pron.	MOS	rest pron.	MOS
CMU	Freda	F R EH D	1.63	F R EH D AH	2.88
PbA	Czech	CH AH HH	1.21	CH EH K	4.92
CART	Joey	JH AA IY	1.29	JH OW IY	4.75
TLU	Borchard	B ER AA R D	1.58	B ER SH AA R D	4.00

(a) Worst *one-of-a-kind*

Method	name	*one-of-a-kind* pron.	MOS	rest pron.	MOS
CMU	Even	IY V AH N	4.88	EH V AH N	3.17
PbA	Hamlet	HH AE M L EH T	4.96	HH AE M L AH T	4.92
CART	Stasney	S T AE Z N IY	4.83	S T AE S N IY	4.50
TLU	Imperato	IH M P ER R AA T OW	4.83	IH M P ER AA T OW	4.38

(b) Best *one-of-a-kind*

Method	name	*one-of-a-kind* pron.	MOS	rest pron.	MOS
CMU	Tutor	T UW T ER	4.67	T AH T ER	1.50
PbA	Dupre	D AH P R	1.96	D AH P ER	2.33
CART	Goucher	G AW CH ER	3.96	G AW K ER	2.42
TLU	Birr	B ER	2.25	B R	2.25

(c) Worst *rest*

Method	name	*one-of-a-kind* pron.	MOS	rest pron.	MOS
CMU	Fossett	F AA S AH T	4.83	F AA S EH T	4.83
PbA	Pandora	P AA N D AO R AH	4.46	P AE N D AO R AH	4.92
CART	Melanie	M EH L AA N IY	4.79	M EH L AH N IY	4.96
TLU	Terri	T ER R IY	3.67	T EH R IY	4.88

(d) Best *rest*

TABLE V

RANKINGS OF THE FOUR METHODS FOR THE ONE-OF-A-KIND PRONUNCIATION BY THE 24 SUBJECTS ACCORDING TO MOS VALUES. THE EQUALS SIGN ('=') INDICATES TIED RANKINGS.

Subject	CMU	PbA	CART	TLU
1	1	2	3	4
2	1	2	4	3
3	1	2	3	4
4	1	2	4	3
5	1=	1=	3	4
6	1	2	4	3
7	1	3	2	4
8	1	2	3	4
9	1	3	2	4
10	1	2	3	4
11	1	3	4	2
12	1	2	3	4
13	1	2=	2=	4
14	1	3	2	4
15	1	2	3	4
16	1	2	3	4
17	1	4	2	3
18	1	2	4	3
19	1	2	3	4
20	1	2	3	4
21	1	4	3	2
22	1	2	3	4
23	1	4	2=	2=
24	1	2=	4	2=

TABLE VI

RESULTS OF WILCOXON SIGNED-RANK TEST FOR ONE-OF-A-KIND PRONUNCIATIONS.

	Z	Sig. (2-tailed)
CMU–PbA	-2.535	0.011
CMU–CART	-2.737	0.006
CMU–TLU	-3.268	0.001
PbA–CART	-0.366	0.715
PbA–TLU	-0.974	0.330
CART–TLU	-0.341	0.733

TABLE VII

MEAN OPINION SCORES FOR SUBJECTIVE EVALUATION BY 24 LISTENERS OF THE 600 REST PRONUNCIATIONS.

	Mean	Std. Deviation
CMU	4.0036	0.6396
PbA	4.1975	0.4685
CART	4.2800	0.4628
TLU	4.2778	0.3937

this figure, readers should recall that *one-of-a-kind* scores are considered to be representative of the worst pronunciations produced by a method whereas *rest* scores are considered to be representative of the quality of pronunciations produced by the three competitor methods. One interesting feature of the results is that listeners rated the CMU *one-of-a-kind* and

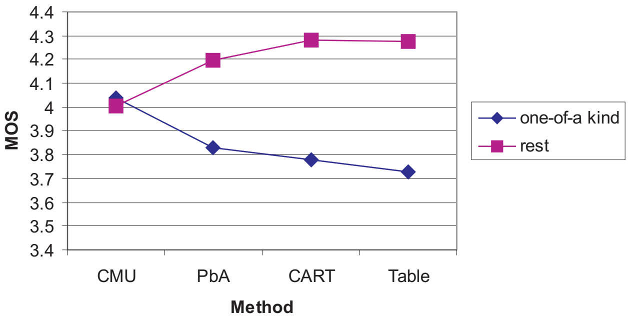

Fig. 3. Composite pattern of MOS values across the four methods for both one-of-a-kind and rest pronunciations. Scores for *one-of-a-kind* words for a method are taken to be indicative of the poorest pronunciations produced by that method. Scores for *rest* words for a method are taken to be indicative of the general quality of pronunciations produced by the three competitor approaches to that method. See text for further explanation.

TABLE VIII

RANKINGS FOR THE FOUR METHODS FOR THE REST PRONUNCIATIONS BY THE 24 SUBJECTS ACCORDING TO MOS VALUES. THE EQUALS SIGN ('=') INDICATES TIED RANKINGS.

Subject	CMU	PbA	CART	TLU
1	4	3	2	1
2	4	3	2	1
3	4	3	1	2
4	4	3	2	1
5	4	3	1	1
6	4	3	2	1
7	4	1	2	3
8	4	3	2	1
9	4	1	3	2
10	4	3	1	2
11	4	3	2	1
12	4	3	2	1
13	4	3	1	2
14	4	3	2	1
15	4	3	2	1
16	4	3	1	2
17	4	3	2	1
18	4	2	3	1
19	4	3	1	2
20	4	3	1	2
21	3=	3=	2	1
22	3	4	1	2
23	4	1	2	3
24	4	1	2	2

rest pronunciations as effectively indistinguishable, meaning that the worst dictionary pronunciations are comparable in acceptability to some form of average of the three data-driven methods. This is perhaps understandable in hindsight as the latter take the dictionary as their 'gold standard' knowledge source for automatic inferencing. We also take this as indirect evidence of the quality and suitability of the dictionary as a 'gold standard'.

Table VIII shows the order in which each of the 24 listeners ranked the *rest* pronunciations produced by each of the 4 methods according to the corresponding MOS values. As expected, these rankings are almost a complete reversal of those for the *one-of-a-kind* pronunciations in Table V.

TABLE IX

RESULTS OF WILCOXON SIGNED-RANK TEST FOR REST PRONUNCIATIONS.

	Z	Sig. (2-tailed)
CMU–PbA	−2.370	0.018
CMU–CART	−4.566	0.000
CMU–TLU	−4.108	0.000
PbA–CART	−1.620	0.105
PbA–TLU	−1.605	0.109
CART–TLU	−0.099	0.921

Table IX shows the results for the Wilcoxon signed-rank test applied to these data. There are highly significant differences between CMU and each of the data-driven methods, with CMU superior to the other techniques ($p < 0.05$, 2-tailed test, df $= 149$), and marginally significant differences between PbA and the remaining two data-driven methods ($p \sim 0.1$). This is taken as suggestive evidence for the superiority of PbA as a back-up strategy.

VI. RESULTS OF OBJECTIVE EVALUATION

For completeness, and because it is relatively easy to do, we also performed objective evaluation of the automatic pronunciation methods. This could not be done for the dictionary pronunciations, as these have to be taken as 100% correct for scoring purposes.

Taking the dictionary pronunciations as correct, the performance of the three data-driven methods was evaluated scoring the automatically-derived pronunciations against the pronunciations in the name list, in terms of both words correct and phonemes correct. We were interested to know how the objective and subjective evaluations compared. It should be clear that they assess different things, so that one can not be predicted from the other, since:

1) The objective evaluation necessarily takes the dictionary pronunciations as correct whereas the subjective evaluation of the CMU method does not. However, the data-driven methods do treat the dictionary pronunciations as a 'gold standard' for the inferential process.
2) The objective evaluation counts all errors (word or phoneme) as equivalent whereas the subjective evaluation recognizes that each can have a different perceptual effect.

As can be seen in Table X, PbA achieved the highest percentage of words correct. We have previously argued that words correct is a more stringent measure of pronunciation accuracy and should be used in strong preference to phonemes correct [6]. The difference between word accuracy for PbA at 68.04% and for the two other methods (61.45% for CART, 61.91% for TLU) is apparently small, but is in fact enormously statistically significant (binomial test, $z \sim 9.8$, $p \sim 0$) when dealing with tens of thousands of words. We note that the phonemes correct measures do not align with words correct. TLU achieves comparable best phonemes correct score to PbA, but is far poorer in terms of word correct. This indicates that phoneme errors are not independently distributed across words, and is another reason for preferring words correct as our measure of effectiveness.

TABLE X

OBJECTIVE EVALUATION OF PRONUNCIATIONS OF THE 52,911 PROPER NAMES BY THE THREE DATA-DRIVEN METHODS IN TERMS OF WORDS CORRECT AND PHONEMES CORRECT.

	words (%)	phonemes (%)
PbA	68.04	92.17
CART	61.45	90.67
TLU	61.91	92.14

VII. DISCUSSION AND CONCLUSIONS

We have compared four methods for proper name pronunciation within a text-to-speech system for English. These were (1) a manually-compiled dictionary (method CMU) and (2) three automatic, data-driven methods for proper name pronunciation, namely: pronunciation by analogy (PbA), a decision tree method (CART), and a table look-up method (TLU). These four methods are not intended to be viewed as entirely free choices for deployment in a practical TTS system; rather we envisage that the dictionary method will form the primary strategy whereas one of the data-driven methods will constitute the back-up strategy used when dictionary look-up fails. The comparison was primarily subjective, on the basis of 24 listeners' opinions of the acceptability of so-called *one-of-a-kind* and *rest* pronunciations. The former, in which one particular method of the four produced a pronunciation which differed from that produced by the other three, was expected to be representative of the worst pronunciations that that particular method would produce. For each *one-of-a-kind* pronunciation, there is a corresponding so-called *rest* pronunciation, being that produced by the remaining three competitor methods. These were expected to be generally better (in terms of acceptability) than the *one-of-a-kind* pronunciations, and also representative of the quality of pronunciations that competitor methods could produce. So a pattern of relatively high *one-of-a-kind* scores and relatively low *rest* scores is taken as an indication of high performance for a given method.

Listeners rated *one-of-a-kind* pronunciations somewhere around 'Almost correct', with an MOS close to 4, or just below for all four methods, whereas *rest* pronunciations were rated somewhere around 'Almost correct' or just above (see Figure 3). Wilcoxon signed-rank tests showed highly significant differences between CMU and the data-driven methods, for both *one-of-a-kind* and *rest* words. The direction of these differences was such as to show that the CMU pronunciations were markedly superior. This was to be expected as the dictionary forms the knowledge base for automatic inference of pronunciations by the data-driven methods. Hence, there is no reason to expect the latter to improve systematically on the dictionary pronunciations. For the *rest* pronunciations, there was a marginal superiority of PbA over the CART and TLU methods. This mirrored the objective evaluation. No such superiority was seen for the *one-of-a-kind* pronunciations.

We have several plans for future work. We have described the procedural difficulties in selecting word pronunciations to use as stimuli in our listening tests. The experimental design adopted here, which used *one-of-a-kind* and *rest* pronunciations, is by no means the only one possible, nor is it obviously the best. Hence, future work should explore other ways of selecting stimuli for the listening tests. It would also be worthwhile to test a wider range of methods (perhaps including proprietary rules if we can gain access to a set), although the high costs of subjective testing and the practicalities of limiting test times to reasonable durations means that we can never hope to compare large numbers of different approaches. In this work, we have used Festival's letter-to-sound rules for stress assignment in all cases but it would perhaps be better to use a stress-assignment method particular to each pronunciation technique, e.g., manually-assigned for the dictionary, inferred by analogy for PbA, etc., not least because using a common stress assignment for the different methods will contribute to similarity between the three data-driven methods. Similar comments apply to the syllabification of the stimuli.

REFERENCES

[1] T. Vitale, "An algorithm for high accuracy name pronunciation by parametric speech synthesizer," *Computational Linguistics*, vol. 17, no. 3, pp. 257–276, 1991.

[2] M. F. Spiegel, "Proper name pronunciations for speech technology applications," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 419–427, 2003.

[3] A. Font Llitjós and A. W. Black, "Evaluation and collection of proper name pronunciations online," in *Proceedings of 3rd International Conference on Language Resources and Evaluation, LREC2002*, Las Palmas, Spain, 2002, pp. 247–254.

[4] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter-to-sound rules," in *Proceedings of 3rd European Speech Communication Association (ESCA)/COCOSDA International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 77–80.

[5] A. van den Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *Proceedings of 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands, 1993, pp. 45–53.

[6] R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson, "Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches," *Computer Speech and Language*, vol. 13, no. 2, pp. 155–176, 1999.

[7] R. I. Damper, Y. Marchand, J.-D. S. Marsters, and A. I. Bazin, "Aligning text and phonemes for speech technology applications using an EM-like algorithm," *International Journal of Speech Technology*, vol. 8, no. 2, pp. 149–162, 2005.

[8] M. J. Dedina and H. C. Nusbaum, "PRONOUNCE: A program for pronunciation by analogy," *Computer Speech and Language*, vol. 5, no. 1, pp. 55–64, 1991.

[9] K. P. H. Sullivan and R. I. Damper, "Novel-word pronunciation: A cross-language study," *Speech Communication*, vol. 13, no. 3–4, pp. 441–452, 1993.

[10] F. Yvon, "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, 1996, pp. 218–228.

[11] R. I. Damper and J. F. G. Eastmond, "Pronunciation by analogy: Impact of implementational choices on performance," *Language and Speech*, vol. 40, no. 1, pp. 1–23, 1997.

[12] Y. Marchand and R. I. Damper, "A multistrategy approach to improving pronunciation by analogy," *Computational Linguistics*, vol. 26, no. 2, pp. 195–219, 2000.

[13] R. I. Damper and Y. Marchand, "Information fusion approaches to the automatic pronunciation of print by analogy," *Information Fusion*, vol. 71, no. 2, pp. 207–220, 2006.

[14] J. R. Parker, "Rank and response combination from confusion matrix data," *Information Fusion*, vol. 2, no. 2, pp. 113–120, 2001.

[15] J. M. Lucassen and R. L. Mercer, "An information theoretic approach to the automatic determination of phonemic baseforms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'84*, San Diego, CA, 1984, pp. 42.5.1–42.5.4.

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

[17] A. Font Llitjós, "Improving pronunciation accuracy of proper names with language origin classes," Master's thesis, Carnegie Mellon University, Pittsburgh, PA, 2001.

[18] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, UK: Cambridge University Press, 1999.

[19] ITU-T Rec. 830, *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*, International Telecommunications Union, Geneva, Switzerland, Feb. 1996.

[20] A. Clark, "Tech note: Voice quality measurement," 2005, available from http://www.tmcnet.com/tmcnet/articles/2005/voice-quality-measurement-vo%ip-alan-clark-telchemy.htm.

[21] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. New York, NY: McGraw-Hill, 1988.