INTERSPEECH
2009 BRIGHTON

Robust LTS rules with the Combilex speech technology lexicon

Korin Richmond, Robert A. J. Clark, Sue Fitt

Centre for Speech Technology Research,
University of Edinburgh, UK

korin@cstr.ed.ac.uk, robert@cstr.ed.ac.uk

10.21437/Interspeech.2009-405

Abstract

Combilex is a high quality pronunciation lexicon, aimed at speech technology applications, that has recently been released by CSTR. Combilex benefits from several advanced features. This paper evaluates one of these: the explicit alignment of phones to graphemes in a word. This alignment can help to rapidly develop robust and accurate letter-to-sound (LTS) rules, without needing to rely on automatic alignment methods. To evaluate this, we used Festival's LTS module, comparing its standard automatic alignment with Combilex's explicit alignment. Our results show using Combilex's alignment improves LTS accuracy: 86.50% words correct as opposed to 84.49%, with our most general form of lexicon. In addition, building LTS models is greatly accelerated, as the need to list allowed alignments is removed. Finally, loose comparison with other studies indicates Combilex is a superior quality lexicon in terms of consistency and size.

Index Terms: combilex, letter-to-sound rules, grapheme-to-phoneme conversion

1. Introduction

For a language such as English a speech synthesizer typically requires a lexicon to predict the pronunciation of the desired utterance from the words of the input text string. Several pronunciation lexica are available for English, though typically only for the major accent groups, such as Received Pronunciation (RP) or General American (GAM). The CMU lexicon [1], being fairly large, free of charge and released under a liberal license, is relatively widely used. However, it is commonly regarded as having variable quality; it has been compiled from several different sources, reportedly including letter-to-sound rules. Furthermore, it does not contain rich additional information other than a word's pronunciation, which is specific to the General American accent. To take another example, the older Oxford Advanced Learner Dictionary (OALD) [2], is more consistent and contains some supplementary information about words, but is smaller in size, with only approx. 63,399 entries, and is restricted to British English pronunciations. As a third example, UniSyn [3] contains a good number of entries, includes richer information, and provides the benefit of being able to produce pronunciations for a wide range of accents of English. Unfortunately, however, its license restricts its use to purely academic research.

Regardless of the specific pros and cons of pre-existing pronunciation lexica, they all share the same problem that they are finite, and typically very slow to change, whereas human language is very fluid. Thousands and thousands of English words exist, or may be invented spontaneously, which do not feature in any pronunciation lexicon. Therefore, in addition to the lexicon, we also need to build letter-to-sound rules, typically trained on the lexicon, to predict pronunciations for words which are not found in the lexicon.

This is problematic because it can be time-consuming, not to mention that no method for predicting English pronunciations has so far even approached 100% accuracy. It might be the case that LTS rules may never achieve 100% accuracy for a language such as English, simply due to the often contradictory nature of its spelling system[1]! Nevertheless, additional errors can be introduced both by inadequacies in the model at the heart of an LTS system, as well as mistakes and inconsistencies present in the lexicon upon which an LTS model is trained. Both these, then, represent areas in which further work may be applied to achieve better LTS systems. This paper considers the second of these, and specifically evaluates the usefulness and quality of a recently developed pronunciation lexicon: Combilex.

Combilex is a relatively large, high-quality pronunciation lexicon that has been developed specifically for use in speech technology applications. It has been created entirely from scratch at CSTR, and has recently been released under wide-ranging license options. Combilex benefits from several advanced features when compared with other available lexica. A discussion of some of these features and the underlying design choices may be found in [4]. More information about obtaining and using Combilex may also be obtained from the project web page [5].

Similar to UniSyn, Combilex is accent-independent. This means we can automatically generate surface lexica specifically tailored to any accent group, or indeed to the accent of any individual speaker. In comparison with other lexica, which may have been created from multiple sources or authors, the pronunciations contained in Combilex have been supervised by a single lexicographer. In addition, a system of phonotactic constraints and automatic consistency-checking rules are applied before any pronunciation is added to Combilex. This helps to guard against human error and the introduction of inconsistency and mistakes[2]. Moreover, Combilex has been implemented in such a way that morphologically-interdependent words are explicitly related to each other. Specifically, only the minimum possible core set of basic words and other morphemes have pronunciations which have been explicitly coded in the Combilex data structures. All other words and terms which are predictable in terms of their derivation (the large majority) may then be generated automatically. Not only has this facilitated rapid development of the complete lexicon, but it also helps to ensure that the

[1] This is in part due to the many languages which have exerted an influence throughout the development of modern English, as well as the diversity of accents with which English is spoken which makes coordinated spelling reform unfeasible

[2] In fact, Combilex's user interface can furthermore make suggestions to the lexicographer during data entry in order to make the process more convenient and efficient.

base	{ n " ju_ew }.{ b % O`_o r n }
RP	{ n j,",u_ew }.{ b % O_o 0_r n }
GAM 1	{ n " u_ew }.{ b % O`_o r n }
GAM 2	{ n j,",u_ew }.{ b % O`_o r n }

Table 1: Combilex pronunciations for the word “newborn”; transcriptions are given for the base form, as well as RP and GAM surface forms.

pronunciations of morphemes in related words are consistent (and remain consistent with any changes, which is a powerful aid in the task of maintaining the lexicon in the long term). In short, the method of Combilex's construction suggests there should be a high level of consistency and accuracy in the pronunciation strings it contains. This should be reflected in the accuracy of the LTS rules that may be built with Combilex.

In addition to a high level of consistency, Combilex also offers another significant feature to help in constructing LTS rules: an explicit, expert manual alignment of the phones contained in a word to their corresponding orthographic units. This is useful since almost all data-driven methods for building LTS rules require an initial alignment between the letters contained in a word and the phones in its pronunciation. For example, systems previously presented using decision trees [6] or Pronunciation by Analogy [7, 8, 9] require a training set consisting of words whose letters are aligned with the corresponding phones in their pronunciation. In fact, the HMM-based system described by Taylor [10], is a rare example of a system which does not require aligned training data (although they still might nevertheless benefit from aligned data were it available).

This paper aims to test the consistency, usefulness and overall quality of Combilex. To do this, we have trained and tested models using Festival's LTS module with the standard procedure of automatic alignment and have compared these with models built with the same procedure, but using Combilex's expert manual alignment instead. Finally, we have compared the accuracy of the LTS rules built from Combilex with comparable rules built with other lexica.

2. Combilex pronunciations

2.1. Accent-independent transcriptions

Combilex transcriptions are encoded using a set of “metaphones” which comprise a superset of the phones found in the different accents of English. This symbol set is based on the SAMPA set, but has been necessarily modified and extended. These transcriptions are termed “base-form” pronunciations, and can be thought of as a generalization of how a word is pronounced in all accents of English. Base-form transcriptions may then be processed *automatically* to yield numerous lexica of accent-specific “surface-form” transcriptions (termed a “surface-form lexicon”), such as generic RP or GAM, or even transcriptions tailored to a specific speaker.

2.2. Phone-grapheme linking

As well at the metaphones themselves, every Combilex base-form transcription also contains an indication of the alignment of the constituent metaphones to the corresponding graphemes in the orthographic representation of a given word. This alignment is maintained (and in certain instances modified) during the conversion of base-form to surface-form transcriptions. Consequently, we can easily obtain phone-grapheme alignments for all surface-form lexica generated from Combilex.

Base	{ ae_i . d " i@_ea l }. I_i s t >
RP	{ aI_i . d " I@_ea l }. I_i s t >
GAM	{ aI_i . d " i_e . @_a l }. I_i s t >

Table 2: Combilex pronunciations for the word “idealist”; transcriptions are given for the base form, as well as RP and GAM surface forms.

2.3. Example transcriptions

The best way to introduce the major characteristics of Combilex pronunciation transcriptions and the phonographemic alignment they encode is to look at a small number of examples.

Table 1 gives the base-form transcription for the word “newborn”, together with examples of RP and GAM surface-from transcriptions which may be automatically generated from it. The braces “{...}” indicate free root morpheme boundaries, and so here we see the word “newborn” is encoded as a compound of the two free root morphemes “new” and “born”. The symbols “"” and “%” denote primary and secondary stress respectively, while “.” marks a syllable boundary. The backtick “`” indicates rhoticity, which is dropped in the RP surface form, but which is retained in the two GAM surface forms.

Orthographic alignment is denoted as pairs of metaphones and graphemes tied together with an underscore “_”. Metaphone strings appear on the left of the underscore, while graphemes appear on the right. For example, the symbol “O_o” represents the open-mid back rounded vowel (with IPA symbol /ɔ/), which is aligned, or tied, to the grapheme “o”. As a notational economy, where the metaphone symbol and grapheme are identical, the underscore and grapheme may be omitted. For example, the symbol “n” represents an alveolar nasal stop which is aligned to the grapheme “n”, whose symbol is identical and so may be omitted. Where more than one metaphone symbol is associated with a given grapheme, they are linked with the “,” symbol. The zero “0” in the RP surface form represents a null metaphone, i.e one that has no acoustic realization. As an aside, we could have represented the “O” vowel as tied to a grapheme “or”. Combilex offers the flexibility to do this, but we have chosen to retain the correspondence to the “r” phone which occurs in other accents.

Sometimes, the orthographic alignment may change during the automatic transformations from the base form to a surface form. For example, for the word “idealism”, which has the set of pronunciation transcriptions shown in Table 2, we observe that to generate the RP surface form requires only two vowel metaphone symbols to be changed to their surface realization. However, for a particular GAM surface form, we might choose to add an extra syllable to the “idea” root morpheme, and in the process to change the alignment of phones to graphemes[3].

[3] we make no assertion here whether this is or is not the correct syllabification for speakers with an American accent, but merely use this as an illustration of the automated manipulations that are possible should we wish to do so.

3. Evaluation of letter-to-sound rule improvement due to manual alignment

3.1. Evaluation I

Our aim to is test the hypothesis that a more consistent lexicon, which includes manual phone-grapheme alignments, can produce better and more robust LTS rules. To do this, a number of Classification and Regression Trees (CART) were built in line with standard practice using the `Combilex` lexicon. The standard Festival tools, including *wagon*, were used to build these models.

To train these LTS models, a `Combilex` surface form lexicon was generated for RP, or Standard British English. This lexicon contained 143,641 entries. In this standard Combilex format a single entry may contain multiple part-of-speech tags, so these entries were first expanded to give a total of 212,465 entries in the format compatible with *Festival*. However, a certain proportion of this total were then removed. This was partly to exclude data we thought unsuitable, and partly to aid comparison with other previous work. First, all entries using non-ASCII characters were removed (these are generally foreign names) along with all entries which included an apostrophe (mostly possessives and contractions) or a space (collocations). Second, words explicitly tagged as non-English were removed. This only removed a certain proportion of foreign loan words; those which have become closely assimilated into English were retained. For example, the word 'anglaise' was removed but 'baguette' was retained. Finally, for words with multiple pronunciations only the most frequent variant was kept. For example, the word "either" has pronunciation variants with an initial /i/ vowel or an /aɪ/ diphthong. Only the more frequent pronunciation with the diphthong was retained. Finally, multiple entries for a word were only admitted where associated with differing part-of-speech tags. This resulted in a lexicon which contained 155,340 words. For the models built in the experiments described in the following sections, this surface form lexicon was used either in its entirety or with further entries removed (described below). In both cases, 10% of the remaining entries (every tenth one) were set aside to be used as test data, with the remaining entries used as training data.

3.1.1. Baseline automatically-aligned models

We first built two baseline LTS models using the same hand-seeded epsilon scattering techniques described in [11] in order to provide a direct baseline comparison. These baseline models were not built using `Combilex`'s included phone-grapheme alignments, but instead using the 'cumulate pairs' method of aligning graphemes and phonemes specified in the Festvox build tools. Model *B1* was built using the full surface-form lexicon, while model *B2* was built using a pruned surface-form lexicon which had words of less than four letters removed. The set of *allowables* specified for these baseline models meant all but 99 words (0.06%) in the training set aligned and could be used in training.

3.1.2. manually aligned models

Next, we built two equivalent models using `Combilex`'s phone-grapheme alignments to test the hypothesis that these models would outperform their respective baseline. Model *M1* was built using the full lexicon; this was the same training set as used for model B1, with the addition of the 99 words that failed to align using the epsilon scattering. Model *M2* was trained using the same pruned set of training data as *B2*.

Model	Words correct (%)	Phones correct (%)
B1	85.30	97.72
B2	84.49	97.44
M1	86.30	97.90
M2	86.50	97.99
OALD	78.13	93.97

Table 3: Initial models: percentage of words and phones correct. "B" indicates a baseline model using automatic alignment, while "M" indicates a model using `Combilex`'s manual alignment. A "1" indicates the full lexicon, while a "2" denotes the lexicon with words less than 4 letters removed.

3.1.3. Results

Table 3 shows the percentage of words and phones correct for each of the models. The total size of the test set in each case is around 15,000 words, or 130,000 phones. The results from a previous study using the OALD lexicon are provided for comparison [6].

First, we compare the data set derived from `Combilex` with that derived from OALD [6]. These are both British English lexica, with the major difference being that `Combilex` is significantly bigger, with 155,340 words as opposed to 70,646. We see across the board that the `Combilex`-trained models produce better results than the OALD models in terms of the percentage of both words and phones correct. It is, however, difficult to draw further conclusions, as the content of these lexica differs considerably.

To test the significance of the improvement in performance between the series of models presented in Table 3, we treated the proportion of words correct as a binomial distribution which was then estimated by normal distribution parameters. The difference between the output of two models could then be tested using Welch's t-test. Each subsequent improvement between models in the order as presented in Table 3 was found to be significant at the 1% level.

In general, the models using the `Combilex` expert manual alignments are better than those using the epsilon scattering technique. It is interesting to note that the removal of words with less than four letters lead to a reduction in performance for the epsilon scattered baseline model B2. This is inconsistent with results reported elsewhere, and warrants further investigation, presented in the next section.

3.2. Evaluation II

A more restricted, conservative lexicon was created to reduce duplication of pronunciations in the training and test sets. This time, the initial lexicon was further filtered so that each orthographic entity only appeared once, rather than permitting multiple entries with only differing part-of-speech tags. This removed potential duplications of many words, such as 'hog', which could appear as a noun and a verb. It also removed homographs such as the 'row'. A new set of models (B3, B4, M3 and M4) were then trained using the same procedures as models B1, B2, M1 and M2 respectively.

3.2.1. Results

Results for models B3,B4,M3 and M4 are presented in Table 4. These results are lower than the original set, but still models

Model	Words correct (%)	Phones correct (%)
B3	77.56	85.65
B4	79.61	96.69
M3	80.26	97.02
M4	80.83	97.15
OALD	78.13	93.97

Table 4: Percentage words and phones correct, for each model on the held-out test data with models trained on the restricted data set (unique word entries). Compare with Table 3. "3" denotes words of less than four letters were present, while "4" denotes they were removed from the lexicon.

B4,M3&M4 outperformed the previously reported OALD results. We now observe that pruning short words consistently improved results. The Combilex manual-alignment models still consistently outperformed the baseline models. It would be possible to continue refining training and test sets to minimize similarity between the two, but it is not clear what this would achieve. Lower performance would be expected, but the models would not be remotely comparable to other published work, or indeed particularly useful in practice.

4. Discussion

Interpreting LTS test results is problematic because determining what makes a meaningful test set is hard. On one hand, there is a paradox of motivation in that a large number of words for which a unique pronunciation can easily be specified, and so are present in the lexicon, are unlikely to need their pronunciation to be predicted by LTS. This implies whatever test set we specify is at best a poor approximation to what the LTS model would actually be required to do in practice. Furthermore, words which are not generally found in pronunciation lexica, for example foreign names, are often not pronounced consistently by human speakers anyway.

It may also be a problem that many forms of a word are present in a lexicon. For example, the word "lock" occurs four times in the Combilex lexicon with different parts of speech: nn, nnp, vb and vbp[4]. Should we include all these, it is likely one of them will appear in the test-set. This means the test data will be less independent of the training data, which could tend to increase LTS test performance. However, if only one variant for each word were retained in the lexicon used to build and test LTS models, we would also remove cases where there is a pronunciation difference, i.e. for homographs which are not homophones. This would also tend to increase LTS test performance. Finally, it could be argued that having morphologically related words in the training and test sets (such as "locked", "locker", "locking" etc.) is likely to inflate test results. It would, however, be difficult to justify their removal, since the model would then be unable to generalize morphological derivations of known roots, such as plurals, adjectival forms and so on.

Ultimately, if the test data is merely a subset of the lexicon, then arguably all we can really test is the consistency of a lexicon. This is in itself an important exercise. From our results, we conclude that Combilex is highly consistent.

5. Conclusions

We have introduced several features of the Combilex speech technology lexicon. One of these is the expert manual alignment of metaphones to graphemes in the lexicon's base form transcriptions. We have sought to evaluate these for training LTS models. From our results, we conclude that Combilex is a consistent lexicon and can be used to generate consistent letter-to-sound rules. Importantly, the presence of the expert-aligned phone-grapheme mapping makes it very easy to produce letter-to-sound rules for any surface-form lexicon that Combilex can generate.

6. Future work

There are potentially better LTS methods than Classification and Regression trees, such as pronunciation-by-analogy [9] or joint sequence models [12]. We would like to explore whether these methods could also benefit from an expert manual phone-grapheme alignment.

Finally, it is in principle possible to train a grapheme-to-phoneme model on Combilex base-form transcriptions (see Section 2), and then convert the resulting pronunciation to desired surface form using the standard Combilex automatic processing. However, this would also require prediction of stress and syllabification in many cases. This is possible and will be the subject of future work.

7. Acknowledgments

The development of Combilex was supported by a Proof-of-Concept grant from Scottish Enterprise. K. Richmond is currently supported by EPSRC grant EP/E027741/1.

8. References

[1] "The Carnegie Mellon University pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[2] R. Mitten, "Computer-usable version of Oxford Advanced Learner's Dictionary of current english," Oxford Text Archive, 1992.

[3] S. Fitt, "Unisyn multi-accent lexicon," http://www.cstr.ed.ac.uk/projects/unisyn.

[4] S. Fitt and K. Richmond, "Redundancy and productivity in the speech technology lexicon - can we do better?" in *Proc. Interspeech*, Sept. 2006, pp. 165–168.

[5] S. Fitt, K. Richmond, and R. Clark, "Combilex," http://www.cstr.ed.ac.uk/research/projects/combilex.

[6] V. Pagel, K. Lenzo, and A. Black, "Letter-to-sound rules for accented lexicon compression," in *Proc. ICSLP*, Sydney, Australia, 1998.

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

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

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

[10] P. Taylor, "Hidden Markov models for grapheme-to-phoneme conversion," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 1973–1976.

[11] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 77–80.

[12] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, 2008.

[4] Penn Treebank part-of-speech tags