On Generating Combilex Pronunciations via Morphological Analysis

Korin Richmond, Robert Clark, Sue Fitt

Centre for Speech Technology Research, Edinburgh University, United Kingdom

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

Abstract

Combilex is a high quality lexicon that has been developed specifically for speech technology purposes and recently released by CSTR. Combilex benefits from many advanced features. This paper explores one of these: the ability to generate fully-specified transcriptions for morphologically derived words automatically. This functionality was originally implemented to encode the pronunciations of derived words in terms of their constituent morphemes, thus accelerating lexicon development and ensuring a high level of consistency. In this paper, we propose this method of modelling pronunciations can be exploited further by combining it with a morphological parser, thus yielding a method to generate full transcriptions for unknown derived words. Not only could this accelerate adding new derived words to Combilex, but it could also serve as an alternative to conventional letter-to-sound rules. This paper presents preliminary work indicating this is a promising direction.

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

1. Introduction

Speech technology very much relies on the availability of a mapping from words to pronunciations. As the final stage of front-end linguistic processing in text-to-speech synthesis (TTS), a specification for the pronunciation of the words in the sentence must be found. In automatic speech recognition (ASR), this mapping is typically relied upon to form the hypothesised word sequences in terms of phone-based statistical acoustic models. Two simple ways to realise this mapping are commonplace: a lexicon simply listing words and their pronunciation; and letter-to-sound (LTS) rules, either hand-written or learned automatically from data, to predict a pronunciation from a word's orthographic form. For many languages, either one of these, or a combination of the two, gives acceptable results. For some languages, however, they are insufficient and word morphology must be taken into account. For example, agglutinating languages such as Finnish have a massive number of potential words that can be formed using a large set of affixes. It would simply not be practical to list all these in a lexicon for ASR [1]. Meanwhile for German, TTS systems benefit from morphological analysis to deal with the productive process of word-compounding [2].

In the case of English, morphological analysis has not traditionally played a significant role in TTS or ASR. English has a rich and fairly complex system of derivational morphology, but only a tiny number of inflections[1]. Consequently, for ASR it is regarded as feasible just to have a large list of words as "atomic" entities, without considering their underlying morphology. Meanwhile, English spelling is irregular enough that a lexicon must be used for TTS, but it is still regular enough that LTS rules for out-of-vocabulary (OOV) words are worthwhile. As a result of this situation, English speech technology lexicons, such as CMU [3], typically have not taken morphology into account at all. In stark contrast, the Combilex lexicon [4, 5] aims not only to include information pertaining to morphological derivation, but also to exploit and benefit from it.

Combilex is a new lexicon that has been developed from scratch and recently released by CSTR. Combilex benefits from a number of advanced technical features that set it apart from other speech technology lexicons. One of these, of primary interest here, is its use of morphological derivation. Whereas standard lexicons list morphologically related words separately, the entries for derived words in Combilex are explicitly linked to their root words and *generated from them by rule*. For example, the word "run" is classified as a *core* word (free root), and a pronunciation has been explicitly entered for it. However, entries for words derived from this, such as "running" or "rerun", are encoded only in terms of the root "run" and the appropriate affixes; pronunciations and part-of-speech (POS) tags for such derived words are generated automatically. The motivation for structuring Combilex in this way was twofold. First, we aimed to accelerate lexicon development itself. It proved significantly faster to enter derived words in terms of their morphology, than to enter pronunciations and POS tags for all words by hand individually. Second, this structure helps to ensure consistency, since the pronunciations for all morphologically related words are explicitly linked. If we should modify or update the pronunciation for a core word, that change would be automatically propagated to all related words.

The ability to process pronunciation strings to achieve morphological derivation has certainly proved beneficial for developing and maintaining Combilex. However, it also seems possible to exploit this functionality further. In principle, a morphological analyser, or parser, could be used in conjunction with the morphological derivation component of Combilex to generate pronunciation strings and POS tags for new words automatically. This would have two uses. In the first instance, it would make entering new derived words even more convenient and efficient, since morphological analysis could be applied automatically and the set of resulting parses presented to the user. The user could then simply select the appropriate parse, and, if desired, verify the associated automatically generated pronunciation before committing it to the lexicon. The second potential use would be generating pronunciations for OOV words "on the fly", for TTS for example. This would be similar to using LTS rules. However, the advantage of the proposed approach would be that a full specification for OOV words would be made available. Combilex pronunciation strings provide not only the phones themselves, but also different levels of lexical stress, syllabification, morpheme boundaries, the link between the phones and the associated graphemes and POS tags. All these could be generated for a derived OOV word too, which would be *as good*

[1] We shall hereafter not distinguish between derivational and inflectional morphology, simply referring to both as derivation.

26 – 30 September 2010, Makuhari, Chiba, Japan

as having the word in the lexicon. This is especially useful considering TTS methods such as unit selection and HMM-based synthesis typically use extensive context features as part of their approach to waveform generation.

As far as we are aware, no work has previously been presented on this topic, at least for English. The purpose of this paper is therefore primarily to introduce and discuss this idea, as well as preliminary work to indicate the extent to which this might be a useful approach. At this preliminary stage, there are two obvious questions. The first concerns whether we can estimate how useful this approach might be in terms of how many words currently outside our lexicon are likely to be derived words. The second concerns the extent to which we might expect this approach will work for that set of words.

For the remainder of this paper, Section 2 begins by giving a brief introduction to Combilex, describing its morphological derivations component in particular. In Section 3, we look closer at the proposed idea of generating Combilex pronunciations for unseen words via morphological analysis. Instrumental to this, we briefly discuss the tool we have used for morphological analysis, and present some of our preliminary findings, before finally presenting our conclusions in Section 4.

2. Main features of Combilex

Combilex is a relatively large 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 licensing options. Combilex has many advanced features compared to other available lexicons. A fuller discussion of some of these, and the underlying design decisions, may be found in [5, 4]. More information about obtaining and using Combilex may also be obtained from the project web page [6].

Combilex is an accent-independent lexicon. This means we can use it to automatically generate surface lexicons specifically tailored to any accent group, or indeed to the accent of any individual speaker. Unlike other lexicons, which may have been created from multiple sources or authors, the pronunciations contained in Combilex have been supervised by a single lexicographer. In addition, Combilex has been created as an SQL database, and a system of phonotactic constraints and automatic consistency-checking rules are applied before any pronunciation is added to Combilex. This all helps guard against the introduction of mistakes and inconsistency due to human error. Furthermore, as mentioned in Section 1, Combilex has been implemented so that morphologically-interdependent words are explicitly linked. Specifically, only the minimum possible core set of basic words and other morphemes have pronunciations which have been entered. All other words and terms which are predictable are then generated automatically. Not only has this facilitated rapid development, but it also helps to ensure that the pronunciations of morphemes in related words remain consistent, which is a powerful aid in the task of maintaining the lexicon in the long term. In short, the method of Combilex's construction implies a high level of consistency and accuracy in the pronunciation strings it contains.

Combilex offers rich information in addition to the phones contained in a word. This includes POS tags, lexical stress, syllabification, morpheme boundaries, free variant and headword ordering, source domain and gender tags for names, source language tags for loanwords, and an explicit alignment of the phones contained in a word to their corresponding graphemes. This last feature is useful for at least three reasons. First, it can be used when generating a lexicon for a non-native accent, for which pronunciation may often be influenced by the written form of a word. Second, it can be used to build LTS rules [4]. Almost all data-driven methods for building LTS rules (e.g. decision trees [7] or Pronunciation by Analogy [8, 9]) require a training set consisting of words whose letters are aligned with the corresponding phones in their pronunciation. Finally, as we shall see in Section 2.2, this alignment is indispensable when processing pronunciation strings to effect many of the morphological derivations in English.

```
< d % i_e .{ @_a . s_c " I_i d }. I#9_i . f ae_y >
```

Figure 1: Combilex transcription for the word "deacidify".

2.1. Combilex transcriptions

Combilex transcriptions are written using a set of "metaphones", which are 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 generalisation of how a word is pronounced in all accents of English. Base-form transcriptions may then be processed *automatically* to yield numerous lexicons 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.

The best way to introduce the main features of Combilex transcriptions is with an example. Fig. 1 gives the base-form transcription for the word "deacidify". The braces "{...}" indicate free root morpheme boundaries, while "<" and ">" indicate prefix and suffix morpheme boundaries respectively. Thus, this word is encoded as a derivation of the free root morpheme "acid", using suffix "-ify>" and prefix "<de-". The symbols "ˈ" and "ˌ" denote primary and secondary stress respectively, while "." marks a syllable boundary.

Every Combilex base-form transcription contains an indication of the alignment of the constituent metaphones to their corresponding graphemes. This is denoted as pairs of metaphones and graphemes tied together with an underscore "_". Metaphones appear to the left of the underscore, with graphemes to the right. For example, the symbol "@_a" represents a schwa vowel (IPA symbol /ə/) that is tied to the grapheme "a". As a notational economy, the grapheme (and underscore) is omitted wherever it is identical to the metaphone string. For example in Fig. 1 the symbol "f" represents a voiceless labiodental fricative which is aligned to the grapheme "f". Though not shown in this example, where more than one metaphone symbol is associated with a given grapheme, they are concatenated with a "," symbol. Furthermore, a "0" is used to represent a null metaphone, i.e. one that has no acoustic realisation, while "+" indicates a grapheme (e.g. "e") which, although not pronounced itself, does have an effect on the phonemic realisation of surrounding graphemes. Finally, it is important to note that this alignment is retained through the conversion from base-form to surface-form transcriptions. Consequently, we can easily obtain phone-grapheme alignments for all surface-form lexicons generated from Combilex.

2.2. Morphological derivations

Combilex currently is capable of processing base-form transcriptions with more than 90 suffixes and 30 prefixes. Example suffixes include -ed>, -ic>, -ify> and -ness>, while example prefixes include <de-, <semi-, <re-, <over- and so on. Affixes were added to Combilex in order of their productivity. Very productive inflectional suffixes,

domestic	{ d @_o . m " E_e . s t I_i k_c }
ity>	@_i . t iy_y >
{domestic}ity>	{ d % Q@U_o . m E_e . s t " I_i s_c }. @_i . t iy_y >

Figure 4: An example of a complex process of derivation; the ity> suffix, in common with many affixes, causes significant changes in the stem. Not only can it cause changes in compliance with spelling rules, but it also causes the primary stress to move to the final syllable of the stem, as well as changing the final [k] phone to [s]. Secondary stress is then introduced to the first syllable of the stem, which triggers a vowel change (non-reduction).

free	{ f r " i_ee }
dom>	. d @_o m >
{free}dom>	{ f r " i_ee }. d @_o m >

Figure 2: An example of simple derivation: suffix -dom> is comparatively straightforward: the pronunciation for the suffix may be simply appended to the root.

a)

bug	{ b " V_u g }
ed>	d_ed >
{bug}ed>	{ b " V_u g_gg } d_ed >

b)

buck	{ b " V_u k_ck }
ed>	d_ed >
{buck}ed>	{ b " V_u k_ck } t_ed >

Figure 3: An example of moderate complexity; the -ed> suffix cause changes in the stem in compliance with spelling rules, and also varies in form itself according to the final phone of the stem (e.g. whether an alveolar stop, voiced or voiceless).

such as -s> (e.g. to make a plural noun), were added first, whereas a less productive suffix such as -dom> was a lower priority. Consequently, although there may be affixes which are not currently implemented, the major and most productive ones are very likely to have been accounted for.

Each affix can be thought of as a function that takes a base transcription string and modifies it in whatever way necessary to produce the pronunciation for the derived word. The complexity of the processing required varies widely. Some affixes are very straightforward. The -dom> suffix requires merely to append a fixed string to the root transcription, as indicated in Fig. 2. Unfortunately, the majority of affixes are more complicated than this. Fig. 3 gives a moderately complex example, where the phonetic form of the -ed> suffix varies depending on the ending of the word to which it is attached, and English spelling rules such as consonant doubling must also be taken into account to maintain the alignment between metaphones and graphemes. However, the burden of maintaining this mapping is small compared to the benefit it provides for ever more complex derivations, for example those involving stress shifts and vowel changes as shown in Fig. 4. This type of derivation would be very difficult to implement were it not for the orthographic alignment of Combilex transcriptions. Many affixes cause stress pattern changes in this way, for example aire>, arian>, atic>, ation> and so on, and it is exactly this type of phone change that proves challenging for conventional letter-to-sound rules. Addressing this systematically is one of the major attractive theoretical advantages of the proposed morphological analysis approach to pronunciation prediction.

3. Generating OOV transcriptions

A key question at this preliminary stage is how useful it is likely to be to generate Combilex transcriptions for derived OOV words via morphological analysis. The answer lies partly in how many OOV words are derivations, and partly in how accurately we could generate morphological parses for such words.

3.1. How many OOV words are derivations?

Unfortunately, this is by nature a difficult question. It could be suggested a large number of the words not included in Combilex are names, for which an approach based on morphological analysis would be no use at all. However, it is true to say that names too undergo morphological transformation. For example, derived from the one name "Adam" we of course find "Adams", but other less obvious derivations such as "Adamesque", and even "Adamness" and "Adamify", may be easily found with a web search, and with apparent meanings as one would expect!

To estimate an upper bound on the total possible productivity of morphological derivation, we could try an empirical approach of taking a representative set of roots and generating all derivations by applying all possible combinations of affixes. Unfortunately, this is impractical, not least because it is unclear what stopping criterion to use (neoantidisestablishmentarianism? postneoantidisestablishmentarianism?...). Moreover, blindly and exhaustively generating derived words would not reflect human judgement about what constitutes an acceptable, meaningful derivation.

It seems clear intuitively that however many non-derived words there are, there are likely to be many multiples of that number which are derivations *of them*. The statistics from Combilex support this: approx. 22k headwords are "core", non-derived words; 12k words are names; and 100k words have been entered as derivations. Thus, it seems at least likely that the proposed approach will be applicable to a large proportion of OOV words.

3.2. Morphological analysis experiment

We have used the well-known PC-KIMMO [10, 11] morphological analysis tool (version 2.1.13), which is based on Kimmo Koskenniemi's finite-state two-level model of morphology, in an experiment to investigate what proportion of derived words we might expect to have pronunciation strings predicted correctly via the proposed approach. We have furthermore used Englex (version 2.0b5), which provides a two-level description of English morphology for PC-KIMMO consisting of a set of orthographic rules and a lexicon containing approx. 20k roots and affixes. As a straightforward way to evaluate the proposed method, we took the complete set of derived headwords in Combilex and analysed them using PC-KIMMO. One of three outcomes was possible for each word: a) the morphological parse would match that specified in Combilex; b) the parse would not match; or c) no parse would be found. The results obtained are shown in Table 1 (labelled "run 1").

Analysis of the list of words which failed to parse revealed the overwhelming majority contained a root which was either not present in PC-KIMMO's lexicon, or did not include all possible POS types. We chose to add 10,143 new words

run	# match	# non-match	# no parse
1	54,615 (55.3%)	23,497 (23.8%)	20,627 (20.9%)
2	74,474 (75.4%)	18,024 (18.3%)	6,241 (6.3%)

Table 1: Morphological analysis of Combilex derived words.

to PC-KIMMO's lexicon including: 5,171 proper nouns, 4,353 nouns, 870 verbs and 486 adjectives (some with multiple types). We have not so far addressed the issue of missing POS tags for existing words, which would further reduce parse failures.

Meanwhile, analysis of those words for which the parse found did not constitute a single, exact match revealed a number of reasons, which we broadly classify into three groups:

1) Incompatible format In some cases Englex and Combilex use a different approach to representing affix morphemes, which requires harmonisation. For example, Englex has two suffixes "-ise>" and "-ize", whereas Combilex has only "-ise>" and treats "-ize>" as a variant spelling.

2) Differing choices for morphological structure For example, PC-KIMMO has a lexical entry for the word "abolition" and does not parse this word into submorphemes (although does indicate its root is "abolish"), whereas Combilex represents this word with the morphological structure "{abolish}ion>". This means all parses for derived words based on "abolition" do not match, though they may otherwise be the same.

3) PC-KIMMO returns multiple parses for many words. For example, "codirector" is parsed as "<co{direct}or>" (as in Combilex) and "<co<di{rector}". Multiple parses sometimes make sense (e.g. for homographs), but often some parses either do not seem likely or are certainly not correct.

Of these categories, the first is the most straightforward to remedy, and we have undertaken to harmonise the major differences in morpheme format between PC-KIMMO and Combilex. It is possible that the second category would not cause much trouble. For example, Combilex would generate the same pronunciation for "abolitionism" using the "-ism>" suffix irrespective of whether it used its ready-derived pronunciation for "abolition" or whether it processed the derivation for "{abolish}ion>" on the fly. Meanwhile, the production of multiple parses would not pose much of a problem when entering new words into Combilex, since the user would merely be required to select the desired morphological structure. However, it would be more problematic when generating pronunciations unsupervised for TTS. Overall, addressing incompatibilities of types 2) and 3) will require further investigation, and we have not addressed these in the work presented here.

Having added a number of missing roots to PC-KIMMO's lexicon, as well as attempting to harmonise the format of morphological parses between PC-KIMMO and Combilex, we repeated the experiment. These updated results are labelled "run 2" in Table 1. This time the morphological parse for 75% of words exactly matched that specified in Combilex. This is an encouraging result.

4. Conclusions

This paper has proposed a method to generate the pronunciations for derived OOV words via morphological analysis and the derivation functionality of Combilex. One major use for this would be to make it very efficient to increase the coverage of words in Combilex. Another major use could be for TTS, in place of letter-to-sound rules. The attraction of the latter is that for correctly parsed words we would obtain all the same information available for words that are actually present in the lexicon. Importantly, these features are very often used to define context in HMM-based and unit selection synthesis. We have posited that the number of OOV words that are derivations is potentially very large, and in fact there are likely to be many more derived words overall than non-derived ones. Furthermore, preliminary experimentation has shown that the large majority of these may be parsed successfully (75% correct so far, using very simple and conservative match criteria). We conclude therefore that this is a promising line of research and is very much worth pursuing.

In future work, we shall work on increasing the proportion of words which produce a parse that is compatible with Combilex. Nevertheless, it is unlikely we could achieve 100% correct parses, and we will inevitably need to address this issue, which we have not done at all at this stage. Complete failure to find a parse is arguably not a serious problem; for TTS, it would be simple to fall back to using LTS rules for example. However, the issue of incorrect parses will require more attention. Finally, once we have addressed these two issues and integrated morphological analysis more closely with the morphological derivation component of Combilex, it will be interesting to conduct a fair comparison against the performance of a state-of-the-art LTS system.

5. Acknowledgements

The development of Combilex was supported by a Proof-of-Concept grant from Scottish Enterprise. K. Richmond is currently supported by EPSRC grant EP/E01609x/1.

6. References

[1] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Trans. Speech Lang. Process.*, vol. 5, no. 1, pp. 1–29, 2007.

[2] B. Möbius, "The Bell Labs German text-to-speech system," *Computer Speech and Language*, vol. 13, no. 4, pp. 319 – 358, 1999.

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

[4] K. Richmond, R. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1295–1298.

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

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

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

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

[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] "PC-KIMMO: A morphological parser," http://www.sil.org/pckimmo/index.html.

[11] E. L. Antworth, *PC-KIMMO: a two-level processor for morphological analysis*, ser. Occasional Publications in Academic Computing. Dallas,TX: Summer Institute of Linguistics, 1990, no. 16.