On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition

Lori Lamel and Gilles Adda
Spoken Language Processing Group
LIMSI-CNRS
91403 Orsay, FRANCE
{lamel,gadda}@limsi.fr

ABSTRACT
Creation of pronunciation lexicons for speech recognition is widely acknowledged to be an important, but labor-intensive, aspect of system development. Lexicons are often manually created and make use of knowledge and expertise that is difficult to codify. In this paper we describe our American English lexicon developed primarily for the ARPA WSJ/NAB tasks. The lexicon is phonemically represented, and contains alternate pronunciations for about 10% of the words. Tools have been developed to add new lexical items, as well as to help ensure consistency of the pronunciations. Our experience in large vocabulary, continuous speech recognition is that systematic lexical design can improve system performance. Some comparative results with commonly available lexicons are given.

1. INTRODUCTION
Creation of pronunciation lexicons for speech recognition is widely acknowledged to be an important aspect of system development, but is it rarely addressed in detail. This is probably because the lexicons are often manually created and make use of knowledge and expertise that is difficult to codify. Lexical design entails two main parts - selection of the vocabulary items and representation of the pronunciation entry using the basic units of the recognition system. For large vocabulary, continuous speech recognition systems, the unit of choice is usually phonemes or phone-like units. Vocabulary selection to maximize lexical coverage for a given size lexicon has been previously reported. On the ARPA North American Business News (NAB) task, the out-of-vocabulary (OOV) word rate with a 20k lexicon is about 2.5%. With a 20k word vocabulary and unrestricted test data, there are about 1.6 errors for each OOV word. An obvious way to reduce the error rate due to OOVs is to increase the size of the lexicon. This was found to be the case for up to 65k words, despite the potential of increased confusability of the lexical entries. By reducing the OOV rate, we recover on average 1.2 times as many errors as OOV words removed[1].

Our experience in large vocabulary, continuous speech recognition is that systematic lexical design can improve the overall system performance. The LIMSI American English lexicon developed for the ARPA WSJ/NAB task contains 65,500 words and 72,637 pronunciations[1]. It is represented phonemically, with an average of 6.5 phones/transcription. Alternate pronunciations are given for about 10% of the words, and represent frequent pronunciation variants as well as systematic variations. The 1993 LIMSI WSJ training and 20k test lexicons have been shown to perform well by other sites (CUED, ICSI, Philips and SRI), who have compared this lexicon to other publicly available lexicons.

In this paper we give an overview of how the LIMSI pronunciation lexicon is designed. This includes a description of the tools used to determine pronunciations of new lexical items and tools developed for checking the consistency of the entries.

2. LEXICAL REPRESENTATION
Our approach is to represent the lexicons with standard pronunciations using the set of 45 phonemes given in Table 1. In generating the pronunciations we have attempted to remain close to standard pronunciations and do not explicitly represent allophones. For example, in contrast to the TIMIT lexicon[2], stop allophones of /t/ and /d/ as flaps are not represented. We have chosen a phonemic representation, as most allophonic variants can be predicted by rules, and their use is optional. More importantly, there often is a continuum between different allophones of a given phoneme and the decision as to which occurred in any given utterance is subjective. By using a phonemic representation, no hard decision is imposed, and it is left to the acoustic models to represent the observed variants in the training data.

<table>
<thead>
<tr>
<th>Example alternate pronunciations. Phones in { } are optional, phones in [ ] are alternates.</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTING kawn{t}</td>
</tr>
<tr>
<td>INTEREST Intરist In{t}Xист</td>
</tr>
<tr>
<td>INDUSTRIALIZATION Indાસ્ટીલ[xY] zeSxn</td>
</tr>
<tr>
<td>COUPON k{y}упан</td>
</tr>
<tr>
<td>EXCUSE Eкску[sz]</td>
</tr>
</tbody>
</table>

For each word the baseform transcription is used to generate a pronunciation graph to which word-internal phonological rules are optionally applied during training and recognition to account for some of the phonological variations observed in fluent speech. Some example alternate pronunciations are given in Figure 1 using the phone symbol set given in Table 1. The pronunciation for “COUNTING” allow the /t/ to be optional, as a result of a word-internal phonological rule. The second word “INTEREST”, may be produced with 2 or 3 syllables, depending upon the speaker, where in the latter case the /t/ may be deleted.

There are a variety of words for which frequent alternative pronunciation variants are observed, and these variants are not due to
allophonic differences. One common example is the suffix “IZATION” which can be pronounced with a diphthong (/Y/) or a schwa (/x/). Out of 7 occurrences of the word “INDUSTRIALIZATION” in the training data, 3 are pronounced with /Y/ and 4 with /x/. Another pronunciation variant is the palatalization of the /k/ in a /u/ context, such as in the word “COUPON”. In the spectrogram on the left of Figure 2 the word was pronounced /kupan/ (406c0210), whereas on the right the pronunciation is /kyupan/ (20ac0103). The grid is 100ms by 1 kHz. In contrast, the alternate pronunciations for “EXCUSE” reflect different parts of speech (verb or noun).

Fast speakers tend to poorly articulate unstressed syllables (and sometimes skip them completely), particularly in long words with sequences of unstressed syllables. Although such long words are typically well recognized, often a nearby function word is deleted. In an attempt to reduce these kinds of errors, alternate pronunciations for long words such as “AUTHORIZATION” and “POSITIONING”, are included in the lexicon allowing schwa-deletion or syllabic con-sonants in unstressed syllables. Such alternative pronunciations are also provided for common 3 syllable words such as “INTEREST” (which has the pronunciations /IntrIst/, /IntXIst/ and /InXIst/). Figure 3 shows two examples of the word interest by different speakers reading the same text prompt: “In reaction to the news interest rates plunged...” (20tc0106,40lc0206). The pronunciations are those chosen by the recognizer during segmentation using forced alignment. In the spectrogram on the left of Figure 4 the word “AUTHORIZED” has 3 syllables /cTXYzd/ (407c020r) while on the right the chosen the pronunciation was /cTrYzd/ (40ko0302).

3. PRONUNCIATION GENERATION TOOL

Since generating pronunciations is time-consuming and error-prone (it is mostly manual work), several utilities were developed to facilitate the work. While these utilities can be run in an automatic mode, our experience is that human verification is required, and that interactive use is more efficient. (For example, an erroneous tran-
scription early on was obtained for the word “used”. The program
derived the pronunciation /ə/st/, from the word “us”. These types of
effects can only be detected manually.)

Figure 5: Pronunciation generation tool.

An overview of the procedure is shown in Figure 5. First, missing
pronunciations are generated by rule when possible, by automatically
adding and removing affixes. Some example affix rules are
given in Figure 6 along with example words. The rules apply to
either prefixes (P) or suffixes (S) and specify ordered actions (strip,
strip+add, ...) which apply to the words (letters) and context de-
pendent actions to modify pronunciations. For example, if the word
"r undoled" is unknown, the letter sequence "ed" is removed and the
"r" undoubled. If the word “blur” is located, the phone /ld/ is added
to the returned pronunciation.

When multiple pronunciations can be derived they are presented for
selection, along with their source. The source lexicons that we make
use of are (in order of decreasing confidence): the LIMSI “Master”
lexicon, which contains pronunciations for 80k words; the TIMIT
lexicon[2] (different phone set, fewer allophonic distinctions); a
modified version of the Moby Pronunciator v1.3[3] (different phone
set and conventions for diphthongs); and a modified version of MIT
pronunciations for words in the Merriam Webster Pocket diction-
ary of 1964 (different conventions for unstressed syllables). The
Carnegie Mellon Pronouncing Dictionary (version cmudict.0.4)[4]
(represented with a smaller phone set) and the Merriam Webster
American English Pronouncing Dictionary[5] (a book) are also used for
reference. While treating a new word list, all pronunciations for
new words are kept in a temporary dictionary so that inflected forms
can be derived. We observed that often when no rules applied, it
was because the missing word was actually a compound word (car-
pool), or an inflected form of a compound word (carpools). Thus,
the ability to easily split such words and concatenate the result of
multiple rule applications was added.

At the current time we have not developed any specific tools for
consistency checking, but make use of Unix utilities to extract and
verify all words with a given orthographic form. By using the
pronunciation generation tool, we ensure that pronunciations of new
words are consistent with respect to pronunciation variants in the
Master lexicon. For example, if the /ld/ is optional in certain /nd/
sequences (such as candidate) it is also optional in other similar
words (candidates, candidacy).

4. EXPERIMENTAL RESULTS

In this section we compare recognizer performance with different
lexicons, the use of single and alternate pronunciations for words,
and the use of lexical stress. The lexicons compared are the LIMSI
lexicon (LIM), LDC Pronlex [6], CMUDICT [4] and L2S [7]. The
20k wordlist and trigram LM are those used in the 1993 ARPA
WSJ baseline test. The acoustic training data consist of the 7240
sentences in the WSJ0-st84 corpus. From each source lexicon a
training lexicon and a 20k test lexicon were extracted. We were
not able to compare 65k lexicons for these experiments because too
many words in our 65k wordlist are missing from the CMU (10774
missing) and LDC (14890 missing) lexicons. Since the CMU and
LDC lexicons contain lexical stress markers, two versions of the
lexicons were created. The number of phones used to represent the
pronunciations are given in Table 2, where silence is included as a
phone.

The test data are the same 200 sentences, 10 from each of 20 speakers
(11f/9m), used in the SQALE evaluation[8]. This data set was chosen
because it is the only 20k test set for which the LIMSI lexicons had
not been already updated to include correct pronunciations for the
words in the test data. The out-of-vocabulary rate of the test data is
1.5% with OOVs occurring in 40 of the 200 sentences.

All experiments were run using a trigram word graph generated by
merging the correct string with the output of bigram pass from the
SQALE evaluation run. Here we made the assumption that the use of
the corrected graph will not affect differently the different lexicons
so that we can still compare the results.2

For each condition (lexicon, number of pronunciations, stress), 3
model reestimation cycles (segmentation and acoustic model esti-
mation) were carried out. The training was initialized with sets
of speaker-independent (SI), context-independent (CI) phone mod-
els, mapped from a set of 46 CI phones trained on the WSJ0-st84
corpus. The SI acoustic model sets all contained about 900 context-
dependent (CD) phones.

In Table 2 the error rates for each condition are given. For the LIMSI
lexicon we compared the use of single and multiple pronunciations
in training and testing. Compared to the best results (mult/mult),
the use of only a single pronunciation for both training and test results in
an error increase of 7%. The use of multiple pronunciations is seen
to be more important in the test lexicon than in the training lexicon.
Training with a single pronunciation and testing with multiple ones
only increases the word error by 5%, while training with multiple
pronunciations and testing with only one increases the word error by
23%. Using 3 sets of 2390 tied-state CD models (the same models as
were used in the SQALE evaluation) with the LIM multi/mult lexicon,
the word error is 13.9%, corresponding to an error reduction of 9%.
The CMU lexicon is represented with 40 phones without differenti-
ating lexical stress, and with 55 phones if a stressed/unstressed

distinction is made. The primary and secondary stress markers

1The algorithm was inspired by a set of rules written by David Shipman
while he was at MIT.

2While we may expect that the absolute values of the results may be
optimistic since the correct solution was injected in the graph, this was not
the case. The trigram pass was carried out using the corrected graphs and
the same 3 sets of SQALE 2390 tied-state CD models (SI/multi) and there was
no difference in recognition error (13.9%).
To estimate the models, many contexts that did not have sufficient training examples to accurately predict the performance. Second, with a relatively small training corpus there are several reasons. First, we wanted to see if differentiating stress would lead to a gain in performance. To obtain such improvements and that systematic design is essential, we are able to estimate the relative frequencies of different pronunciation errors. While it is difficult to evaluate changes to the lexicon, we have found that small, but consistent performance improvements can be obtained and that systematic design is essential to obtaining such improvements.

## 5. DISCUSSION

It is difficult to compare the performance of different lexicons and of lexical modifications for several reasons. First, the set of CD acoustic models depends on the lexical representation and the phone contexts appearing in the training data. Second, it is difficult to measure performance differences on a small set of test data, as at most only a few occurrences of modifications can occur. In the 200 SQALE test sentences, there are 3415 words, of which 1464 are distinct, less than 10% of the lexicon entries. Even the ARPA Nov94 test set containing 400 sentences had only 8189 words, 2293 distinct, substantially less than 10% of the 65k lexicon. An obvious solution is to evaluate the system performance using as many different test sets as possible. However, carrying out the experiments is very time consuming, as each time the training lexicon is modified, several iterations of segmentation and model estimation need to be carried out. On the WSJ0-s184 training data we are able to complete a retraining and 20k trigram decoding pass in about 30h. When we want to evaluate WSJ0/1 training, the training cycle takes about 3 days. Evidently recognition without the use of bigram graphs would take much longer.

We evaluate the lexicon in the context of our recognizer by confronting the pronunciations with large corpora. By carrying out a forced alignment of the training data using its orthographic transcription, we are able to estimate the relative frequencies of different alternative pronunciations, as well as to determine sources of pronunciation errors. While it is difficult to evaluate changes to the lexicon, we have found that small, but consistent performance improvements can be obtained and that systematic design is essential to obtaining such improvements.

### Table 2: Word recognition with a 20k trigram language model.

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>#phones</th>
<th>pronns</th>
<th>trn/tst</th>
<th>#models</th>
<th>%WErr</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIM</td>
<td>46</td>
<td>single/single</td>
<td>896</td>
<td>16.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>single/mult.</td>
<td>896</td>
<td>16.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>mult./single</td>
<td>893</td>
<td>18.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>mult./mult.</td>
<td>893</td>
<td>15.2</td>
<td></td>
</tr>
<tr>
<td>CMU</td>
<td>40</td>
<td>mult./mult.</td>
<td>929</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>mult./mult.</td>
<td>941</td>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td>LDC</td>
<td>44</td>
<td>mult./mult.</td>
<td>924</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>mult./mult.</td>
<td>925</td>
<td>16.1</td>
<td></td>
</tr>
<tr>
<td>LDC-S2</td>
<td>59</td>
<td>mult./mult.</td>
<td>926</td>
<td>16.4</td>
<td></td>
</tr>
<tr>
<td>L2S</td>
<td>41</td>
<td>single/single</td>
<td>910</td>
<td>18.3</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 6: Some example affix rules.](image-url)

### References

6. COMLEX English Pronouncing Dictionary (PRONLEX), V0.2, available via the Linguistic Data Consortium.
Sound File References:

[a683s01.wav]
[a683s02.wav]
[a683s03.wav]
[a683s04.wav]
[a683s05.wav]
[a683s06.wav]