# Reversible Sound-to-letter/Letter-to-sound Modeling based on Syllable Structure *

**Stephanie Seneff**
Spoken Language Systems Group
MIT Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Cambridge, MA 02139
seneff@csail.mit.edu

## Abstract

This paper describes a new grapheme-to-phoneme framework, based on a combination of formal linguistic and statistical methods. A context-free grammar is used to parse words into their underlying syllable structure, and a set of subword "spellneme" units encoding both phonemic and graphemic information can be automatically derived from the parsed words. A statistical $n$-gram model can then be trained on a large lexicon of words represented in terms of these linguistically motivated subword units. The framework has potential applications in modeling unknown words and in linking spoken spellings with spoken pronunciations for fully automatic new-word acquisition via dialogue interaction. Results are reported on sound-to-letter experiments for the nouns in the Phonebook corpus.

## 1 Introduction

Spoken dialogue systems are emerging as an effective means for humans to access information spaces through natural spoken interaction with computers. A significant enhancement to the usability of such systems would be the automatic acquisition of new knowledge through spoken interaction with its end users. Such knowledge would include both the spelling and pronunciation of a new word, ideally leading to a successful match to an entry in a large external database. To take advantage of an integrated approach to recognizing the spoken and spelled forms of a new word, there is a need for a high-quality reversible phoneme-grapheme mapping system. This is a difficult task for English due to the many inconsistencies in letter-to-sound rules as a consequence of borrowings from multiple language groups.

It is also increasingly the case that dialogue systems must dynamically adjust the recognizer vocabulary to handle changing database contents. If a system can reliably predict the pronunciation of a new word algorithmically, especially if substantiated by a spoken pronunciation of the word during active usage, it will be far more effective in satisfying changing user needs.

In this paper, we describe a new reversible grapheme-to-phoneme framework based on combining formal linguistic knowledge with statistical data-driven techniques. We first describe and motivate our choice for the linguistic model. Section 3 describes the iterative process for obtaining a subword baseforms lexicon used to train the statistical model. Sections 4 and 5 present experiments and results for sound-to-letter modeling on 5000 nouns. We conclude after a brief section on related work.

## 2 Linguistic Model

Our linguistic model is based on syllable structure, but we felt that whole-syllable units would be too large to adequately generalize to unseen data. We thus decided to decompose syllables into onsets and

* This research was supported by the Industrial Technology Research Institute (ITRI) in Taiwan.

*Proceedings of NAACL HLT 2007, Companion Volume*, pages 153–156,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

| rhyme1 | onset | rhyme | usyl | rhyme | usyl | ambi | rhyme |
|---|---|---|---|---|---|---|---|
| -aek | s+ | -ehl | -axr | -aam | -ax+ | tf | -er+ |
| a c | c | e l | e r | o m | e | t | e r |

Figure 1: Linguistic representation for the word "accelerometer," illustrating the structure of our model.

rhymes, which would then become subword pronunciation units in a lexical baseforms file. These subword units would, in turn, be specified in terms of phonemic baseforms in a separate subword lexicon. Thus the words in our training set are represented in terms of subword units, which are converted into phonemic baseforms by simple lookup of the subword pronunciations.

A difficult aspect for English is to decide where to place the syllable boundary within a sequence of intersyllabic consonants. To guide this decision, we made use of sonority constraints combined with maximal stress and maximal onset principles. For a select subset of intersyllable consonants, we invoke the special category "ambi" for "ambisyllabic," to allow the consonant to be ambiguously assigned. In addition to onset and rhyme, we also include the category "affix," to account for those instances of (usually coronal) consonants that would lead to a violation of sonority principles in the coda position (e.g., "fif*ths*," "kep*t*", etc.), following linguistic theory (Blevins, 1995).

We decided to distinguish the first stressed and the first unstressed syllable from all other stressed and unstressed syllables in the word, in order to encode separate statistics for the privileged first position. We also combined onset and rhyme into a single whole syllable unit for a selected subset of relatively frequent unstressed syllables. In total, our current inventory consists of 678 unique symbols.

An example hierarchical representation in our formalism is illustrated in Figure 1, for the word "accelerometer."

## 3 Procedures

Our approach is based on a technique that exploits a context-free grammar applied to a large lexicon to aid in the preparation of a baseforms file encoding the lexicon in terms of a set of linguistically motivated subword units. The subword units, which encode syllabification and pronunciation, are initially derived automatically from a phonemic baseforms file through simple rewrite rules. The grammar is developed manually, a process that amounts to identifying all the possible ways to spell each subword unit. In an iterative procedure, parse failures are manually corrected either by modifying erroneous pronunciations or by augmenting the rules governing permissible letter sequences for the subword units. Through this process we have now converted phonemic baseforms for a lexicon of 140,000 words into the new subword units. Example entries in the baseforms file are shown in Figure 2.

| | |
|---|---|
| acrostics | -ax+ kr+ -aas t -axk +s |
| actualities | -aek ch+ -uw+ -ael -ax+ tf -iy+ +z |
| fabrications | f+ -aeb r+ -ax+ k -ey+ shaxn +z |
| preferences | pr+ -ehf rsyl -axn +s -axz |
| skepticism | sk+ -ehp t -ax+ s+ -ihz -m |
| striplings | str+ -ihp l+ -ihng +z |

Figure 2: Sample entries from the subword lexicon.

Once a grammar and a large lexicon of subword baseforms are available, the next step is to create a statistical language model encoding the letter-subword mappings. We have decided to create a new set of subword units, which we call "spellnemes," combining the letter sequence and associated pronunciation into a single symbolic tag, as illustrated in Figure 3. The sequence of spellnemes associated with each word in the lexicon can easily be obtained by parsing the word, constrained by its subword realization. The spellneme sequences for each word in the lexicon are then used to train a trigram language model. Our formalism currently has 2541 unique spellnemes, on average nearly a 4-fold expansion over the number of pronunciation-based subwords.

Derivative sound-to-letter and letter-to-sound systems are straightforward. For sound-to-letter, a provided phonemic transcript is exhaustively expanded to a graph of all possible subword realizations, and subsequently into a graph of all spellnemes asso-

```
b_r<591> oo_k<547> l<617> e_t<263>
b_r<591> oo_k<547> l<617> i_n e<281>
b_r<591> oo_k<547> l<617> y_n<250>
b_r<591> oo_k<547> m<619> o_n_t<43>
```

Figure 3: Sample entries from the tagged corpus which is used to train the statistics of the $n$-gram language model. The numeric tags encode the associated subword unit, each of which maps to a unique phonemic sequence.

ciated with each subword. The trigram language model is applied to produce an N-best list of the top-scoring hypothesized spellneme sequences. The letter-to-sound system exhaustively expands the letter sequence into all possible spellneme sequences. After applying the trigram language model, the N-best list of spellneme sequences can be mapped to the pronunciations by concatenation of the phonemic realizations of the individual subwords.

## 4 Experiments on Phonebook

We imagine a two-stage speech recognition framework for a word spoken in isolation, in which the first stage uses subword units that encode only pronunciation, and produces an N-best list of hypothesized pronunciations, represented as phonemic baseforms. The second stage is tasked with hypothesizing possible spellings from the provided phonemic baseforms, and then verifying them by a match with a lexical entry. For the purposes of this paper, we assume a perfect phonemic baseform as input, and investigate the quality of the N-best list of hypothesized spellings automatically generated by the sound-to-letter system. We quantify performance by measuring the depth of the correct word in the generated N-best list.

Our experiments were conducted on a set of nearly 5000 nouns and proper nouns, a subset of the 8000 word Phonebook vocabulary that were identified as nouns using the Web site http://www.comp.lancs.ac.uk/ucrel/claws/. We selected this set of words for two reasons: (1) they contain a substantial number of nouns not included in our original training lexicon, and (2) they will allow us to conduct speech recognition experiments from the available Phonebook corpus of words spoken in isolation over the telephone.

The trigram training corpus was restricted to a subset of 55,159 entries in our original lexicon, containing the words that were tagged as nouns in Comlex. We are interested in quantifying the gap between in-vocabulary (IV) and out-of-vocabulary (OOV) words, with respect to the training corpus. We also measure the gains that can be realized through manual repair of automatically generated baseforms for training the sound-to-letter system. Thus we conducted experiments on the following four conditions:

1. Train on 55,159 nouns, test on the 3478 word IV subset of Phonebook nouns.
2. Train on 55,159 nouns, test on the 1518 OOV words in Phonebook.
3. Augment the training set with entries for the 1518 OOV words, that are obtained automatically by processing them through the letter-to-sound system. Test on the OOV subset.
4. Augment the training lexicon with manually corrected pronunciations for the OOV subset. Test on the OOV subset.

Items (3) and (4) will show us the degree to which improvements can be gained through automatic methods, once a new list of nouns becomes available, as well as how much further gain can be realized after manual correction. Automatic methods will be feasible for a dialogue system which can extract from the Web a list of nouns appropriate for the domain, but has no phonemic baseforms available for those nouns.

## 5 Results

Results are shown in Table 1. With an N-best list of 30, the system has a very low failure rate for all conditions. However, there is a marked difference in performance in terms of the depth of the correct answer. The mean depth is 2.07 for the OOV words, as contrasted with only 1.15 for the IV words. Fully automatic methods to improve the sound-to-letter system lead to substantial gains, reducing the mean depth to 1.54. Manual correction provides significant further gains, achieving a mean depth of 1.13, comparable to that of the original IV subset. There were two cases where an incorrect match to a lexical entry was found at a higher level in the N-best list

| | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 30 | Mean Depth | Failed |
|---|---|---|---|---|---|---|---|---|
| OOV | 65.7% | 80.7% | 86.5% | 90.0% | 91.7% | 98.4% | 2.07 | 1.6% |
| plus auto | 84.0% | 91.6% | 93.4% | 94.7% | 95.7% | 99.0% | 1.54 | 1.0% |
| plus manual | 92.2% | 98.0% | 98.9% | 99.3% | 99.6% | 99.9% | 1.13 | 0.1% |
| IV | 91.8% | 97.5% | 98.8% | 99.3% | 99.5% | 100.0% | 1.15 | 0.0% |

Table 1: Percentage of words spelled correctly as a function of N-best depth for sound-to-letter experiments. See text for discussion.

than the correct match. These were the homonym pairs: carolyn/caroline and jasmine/jazzman.

Nouns that fail to appear in the top 30 can potentially still be recovered through simple spell checking methods. Using a conservative approach of allowing only a single letter insertion, substitution or deletion, and further, of requiring that the grammar could parse the corrected word under the constraints of the system's proposed subwords, we were able to recover over 60% of the failures.

## 6 Related Work

Many researchers have worked on letter-to-sound modeling for text-to-speech conversion (R. I. Damper and Gustafson, 1998). The topic of bi-directional phoneme-to-grapheme conversion is becoming important for application to unknown words and new word acquisition in speech understanding systems (Chung et al., 2003), although it is difficult to compare results due to different representations and data sets. In (Meng, 1996), a hierarchical approach was used for bi-directional sound-letter generation. (Rentzepopoulos and Kokkinakis, 1996) describes a hidden Markov model approach for phoneme-to-grapheme conversion, in seven European languages evaluated on a number of corpora. (Marchand and Damper, 2000) uses a fusion of data-driven and pronunciation-by-analogy methods, obtaining word accuracies of 57.7% and 69.1% for phoneme-to-grapheme and grapheme-to-phoneme experiments respectively, when evaluated on a general dictionary. (Llitjos and Black, 2001) report improvements on letter-to-sound performance on names by adding language origin features, yielding 61.72% word accuracy on 56,000 names. (Galescu and Allen, 2002) addresses bi-directional sound-letter generation using a data-driven joint $n$-gram method on proper nouns, yielding around 41% word accuracy for sound-to-letter and 68% word accuracy for letter-to-sound.

## 7 Summary and Conclusions

In this paper, we report on a new technique for reversible letter-to-sound sound-to-letter modeling, which is based on linguistic theory and statistical modeling. The system was evaluated on a set of nearly 5000 nouns from the Phonebook domain, separately for in-vocabulary and out-of-vocabulary subsets, with respect to the training corpus for the sound-to-letter system. In future work, we plan to evaluate the effectiveness of the model for automatic new word acquisition in spoken dialogue systems.

## References

J. Blevins. 1995. The syllable in phonological theory. *J. Goldsmith, Ed., the Handbook of Phonological Theory. Blackwell, Oxford.*

G. Chung, S. Seneff, and C. Wang. 2003. Automatic acquisition of names using speak and spell mode in spoken dialogue systems. In *Proc. of HLT-NAACL*, Edmonton, Canada.

L. Galescu and J. Allen. 2002. Name pronunciation with a joint n-gram model for bi-directional grapheme-to-phoneme conversion. In *Proc. ICSLP*, pages 109–112, Denver, CO.

A. Font Llitjos and A. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proc. Eurospeech*, Aalborg, Denmark.

Y. Marchand and R. I. Damper. 2000. A multi-strategy approach to improving pronunciation by analogy. *Computational Linguistics*, 26(2):195–219.

H. Meng. 1996. Reversible letter-to-sound / sound-to-letter generation based on parsing word morphology. *Speech Computation*, 18(1):47–64.

M. J. Adamson R. I. Damper, Y. Marchand and K. Gustafson. 1998. Comparative evaluation of letter-to-sound conversion techniques for English text-to-speech synthesis. In *Proc. IWSS*, pages 53–58, Jenolan Caves, Australia.

P. Rentzepopoulos and G. K. Kokkinakis. 1996. Efficient multilingual phoneme-to-grapheme conversion based on HMM. *Computational Linguistics*, 22(3).