New Word Acquisition Using Subword Modeling

Ghinwa F. Choueiter, Stephanie Seneff, and James R. Glass

MIT CSAIL Laboratory
32 Vassar Street, Cambridge, MA 02139
ghinwa@mit.edu, seneff@csail.mit.edu, glass@mit.edu

Abstract

In this paper, we use subword modeling to learn the pronunciations and spellings of new words. The subwords are generated with a context-free grammar, and are intermediate units between phonemes and syllables. We first evaluate the effectiveness of the subword model in automatically generating the spelling and pronunciation of new words. Then the subword model is embedded in a multi-stage recognizer which consists of word, subword, and letter recognizers. In a preliminary set of experiments, the hybrid system outperforms a large-vocabulary isolated word recognizer. The subword model is also used to improve the performance of the letter recognizer by generating a spelling cohort which is used to train a small letter *n*-gram. The small letter *n*-gram has a reduced perplexity compared to a much larger *n*-gram, and can be used by the letter recognizer for the spoken spelling mode. This could translate to an improved letter error rate in future letter recognition experiments.

Index Terms: subword modeling, new word acquisition

1. Introduction

The need for more flexible and adaptive automatic speech recognition (ASR) has never been greater due to the widespread emergence of speech-enabled applications and devices [11, 4] as well as spoken dialogue systems for information retrieval [9, 15]. One of the factors impeding the broad acceptance of ASR is the frustration experienced by users when the system breaks down when an unknown word occurs. For open-ended word recognizers with fixed vocabularies, this problem is inevitable since the recognizer does not have immediate access to either the baseforms or spellings of the unknown words. This issue of unknown words has motivated considerable research in the area where different approaches have been adopted.

Statistical grapheme-to-phoneme models have been proposed and shown to perform well in the task of spelling estimation in [3, 7], and pronunciation generation of new words in [5, 7]. Speak-and-spell models have been implemented for the acquisition of city names within dialogue systems in [1, 6]. Furthermore, out-of-vocabulary (OOV) word detection models, which involve no spelling estimation, have been embedded within word-based speech recognizers, and shown to reduce word error rate (WER) in [2, 14].

In this paper, linguistically motivated subword modeling is used for learning new words. The subwords are generated using a context-free grammar (CFG) that encodes positional and phonological constraints. The proposed approach has the advantage of automatically acquiring the spelling and pronunciation of new words. We envision the subword model implemented within a dialogue system, thereby taking advantage of user interactions and augmenting the system with a learning capability. The subword model would be activated upon the detection of an OOV word, and any newly acquired word could then be added to the lexical baseforms.

Apart from providing spoken dialogue systems with error recovery strategies, there are several potential uses of a new word acquisition capability. The subword model can be implemented within a word recognizer as in [2], for example, with the additional feature that it can learn the spelling of the detected OOV word. Furthermore, the new word acquisition mechanism can also be used to build a word recognizer bottom up. Given a set of words whose baseforms are unavailable, the subword model can dynamically generate the lexical baseforms and update the lexicon.

In this research, we focus on the effectiveness of the subword model in automatically generating the spelling and pronunciation of new words. We also implement the subword model within a very simple dialogue system where a person speaks a word, and an isolated word recognizer (Stage I) proposes and displays a list of top candidate words. If the person rejects all the words, the system enters the second stage (Stage II), which uses a subword model. The subword model generates hypothesized word spellings via a sound-to-letter model, and filters invalid spellings using a very large lexicon. If the person rejects the list of words presented by the second stage, a third stage (Stage III) prompts for a spoken spelling of the word.

For this particular dialogue system, we are interested in addressing two questions: (1) How does the isolated word recognizer augmented with the subword model compare against a large-vocabulary isolated word recognizer ? and (2) How can the subword model be used to improve the performance of the spelling mode?

In the rest of this paper, Sections 2 and 3 describe the subword units and the spelling estimation model. Section 4 describes the data and Section 5 describes the experimental setup and reports on the preliminary results. Section 6 concludes and discusses future work.

2. Subword Units

The subwords used in this research are generated through a bootstrapping procedure with a CFG that encodes phonological constraints and sub-syllable structure. First a list of linguistically motivated subwords is proposed, and a set of hand-written rules are used to describe all possible ways a particular subword can be spelled. The subwords are intermediate units between phonemes and syllables which only encode pronunciation information. Next a lexicon is parsed with these rules, and the CFG is manually augmented to cover the words that failed to parse. This process is iterated until the entire lexicon parses. The total number of subwords obtained through this procedure is 677. The CFG is described in more detail in [13]. Figure 1 illustrates the parsing of the words *diction* and *facial* into subword

Syllable Structure	onset	rhyme	onset	usyl
Subword Units	d+	-ihk	sh+	-axn
Letter Clusters	d	ic	ti	on

Syllable Structure	onset	rhyme	onset	usyl
Subword Units	f+	-ey	sh+	-axl
Letter Clusters	f	a	ci	al

Figure 1: Parse analysis of the words *diction* and *facial* obtained using the CFG.

units such as onset and rhyme (denoted with + and - respectively). The illustration also shows a common characteristic of the English language where the same pronunciation is realized by different letter clusters (e.g., "ti" vs "ci") based on context. Two by-products of the CFG are a direct mapping between subwords and their spellings as well as between subwords and their phonemic representation. The former mapping is used to create statistical sound-to-letter (S-to-L) and letter-to-sound (L-to-S) models, while the latter to generate phonemic baseforms from subword representations.

3. Pronunciation and Spelling Estimation

In this section, we model the pronunciation and spelling estimation processes mathematically, and describe our current implementation using finite-state transducers (FSTs).

Given acoustic observations, A, the optimal letter spelling, $L^\star$, can be written as:

$$\begin{aligned} L^\star &= \underset{L}{argmax}\, P(L|A) = \underset{L}{argmax} \sum_{U} P(L,U|A) \\ &\approx \underset{L}{argmax}\, \underset{U}{max}\, P(L,U|A) \qquad (1) \\ &\approx \underset{L}{argmax}\, \underset{U}{max}\, P(A|U)P(U)P(L|U) \end{aligned}$$

Where L is a sequence of letters, and U is the set of subwords units. $P(A|U)$ is the acoustic model, and $P(U)$ is modeled as an *n*-gram on the subwords. The last line assumes that the acoustic events, A, are conditionally independent of the letters, L, given the subwords, U, i.e. $P(A|U, L) = P(A|U)$.

The product $P(A|U)P(U)$ models the subword search space, which can be implemented as a weighted FST, R [10]:

$$R = C\ o\ P\ o\ Lex\ o\ G \qquad (2)$$

Where C denotes the mapping from context-dependent model labels to context-independent phone labels, P the phonological rules that map phone labels to phoneme sequences, Lex the subword lexicon, which is a mapping from subword to phonemic units obtained from the grammar, and G the subword language model (LM). The architecture of R follows a typical word-based speech recognizer where the input is a set of context-dependent phone models and the output is an *N*-best list of subwords which encodes possible pronunciations of the utterance. A search through R produces an *N*-best list of subword sequences, which is denoted R_{N-best}.

Finally, the spelling search space, as represented in Equation 1, can be modeled as:

$$L = R_{N-best}\ o\ T_{U2L}\ o\ D \qquad (3)$$

T_{U2L} is a statistical sound-to-letter mapping which encodes the conditional probability of letter sequences, L, given the subwords, U. D is a deterministic word filter or acceptor, and is used to inforce hard spell-checking, such that if the generated spelling is not in a very large lexicon, it is rejected. Following the filtering stage, a spelling cohort of size M is generated. A letter transition weight is set to reduce the length difference between the reference words and the top hypothesis.

In the rest of this paper, we refer to the output of R as a subword *N*-best list and the output of L as a spellings cohort.

4. Data Sets

Our evaluations are performed on 4682 nouns drawn from the development set of the Phonebook telephone-quality isolated words corpus [12]. The original lexicon for the isolated word recognizer consists of 55k nouns extracted from the LDC Pronlex dictionary. In our experiments, we refer to the Phonebook nouns that are in the 55k lexicon as IV$_{55k}$ (in-vocabulary), and to the words that are not as OOV$_{55k}$. There are 3228 IV$_{55k}$ and 1454 OOV$_{55k}$ words in the Phonebook nouns.

The word acceptor, D, is built with a ∼300k lexicon, which is mostly a subset of the Google *n*-gram corpus. The Google corpus originally contains ∼13 million unique words, and is very noisy. It is reduced to ∼2.5 million words by only keeping lower-cased words with alphabetic symbols. The corpus is then intersected with a carefully cleaned ∼500k lexicon and is augmented with nouns from the Phonebook development set and Pronlex. The result is a ∼300k clean corpus of commonly used English words. The 300k lexicon is also used to build a large isolated-word recognizer.

5. Experiments and Results

This section describes several experiments conducted on the Phonebook data. We utilize the L-to-S system to automatically generate subword and phonemic baseforms for words in the 300k lexicon. We then investigate how effective this baseforms file is in both traditional isolated word recognition and as a training corpus for the subword language model.

The SUMMIT segment-based speech recognition system is used in all our experiments [8]. Context-dependent diphone acoustic models are used and their feature representation is based on 14 MFCCs (Mel-Frequency Cepstral Coefficients) averaged over 8 regions at hypothesized phonetic boundaries. The diphones are modeled with diagonal Gaussian mixture models with a maximum of 75 mixtures per model, and are trained on telephone speech.

A maximum likelihood (ML) estimate of T_{U2L} is obtained using the 300k lexicon. The lexicon is parsed using the L-to-S system into subword units and their corresponding spellings. The ML estimate of T_{U2L} is then obtained simply using counts over the parsed lexicon.

A more detailed look at the subword model is illustrated in Figure 2. When an utterance is presented to the subword model, a subword *N*-best list with corresponding acoustic and LM scores is produced by the subword recognizer. The subword list is transformed into an exhaustive spellings cohort by using T_{U2L}, and invalid words are filtered out with D.

The subword LM weight and the letter transition weight are empirically tuned on a development set.

Isolated Word Recognizers: First, we investigate the ability of the L-to-S system to automatically generate the baseforms of the 300k lexicon. A 300k isolated word recognizer is then built with the automatically generated baseforms, and evaluated in terms of top 10 and top 20 accuracies, meaning that success occurs if the correct word is in the top 10 and top 20 candi-

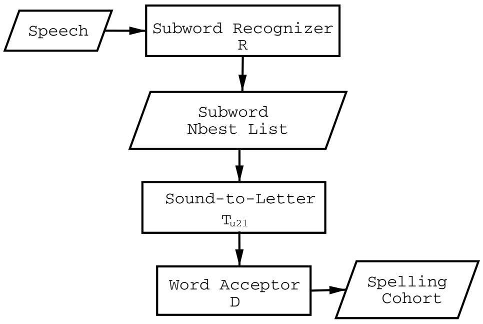

Figure 2: A diagram depicting the pronunciation and spelling estimation by the subword model.

	Top 10 accuracy			Top 20 accuracy		
	IV$_{55k}$	OOV$_{55k}$	All	IV$_{55k}$	OOV$_{55k}$	All
55k	83%	0%	57%	86%	0%	59%
300k	72%	72%	72%	77%	77%	77%

Table 1: Comparison of the 55k and 300k isolated word recognizers, in terms of IV$_{55k}$, OOV$_{55k}$, and overall accuracy. Both recognizers are evaluated based on the top ten and twenty word candidates.

dates respectively. The results are reported in Table 1 for the 3228 IV$_{55k}$ and the 1454 OOV$_{55k}$ words. We note here that all the evaluated words including the OOV$_{55k}$ words are in the 300k lexicon. We report the results for the two subsets (IV$_{55k}$, OOV$_{55k}$) separately in order to compare against the 55k word recognizer. As shown in Table 1, the performance of the IV$_{55k}$ and OOV$_{55k}$ subsets is the same for the 300k system. This illustrates that the automatically generated pronunciations are performing comparably to the manually transcribed ones. Furthermore, the IV$_{55k}$ words suffer significant degradation with the 300k system compared to the 55k word recognizer (i.e. 86% to 77% for top 20 accuracy) due to the larger vocabulary. As expected, the accuracy of the 55k word recognizer on the OOV$_{55k}$ subset is 0%.

Subword Language Models: The subword model produces an *N*-best list of subword sequences, guided by a subword trigram LM, $P(U)$, that is trained on a large corpus. A critical issue is the quality of this LM. In this section, we evaluate the performance of several subword language models. We trained the subword LMs from three training corpora: (1) the 55k lexicon, (2) the 55k lexicon, augmented with just the OOV$_{55k}$ words in Phonebook, and (3) the 300k lexicon. Figure 3 assesses the performance of the three subword recognizers on the OOV$_{55k}$ words. Each of the recognizers produces 1000 *N*-best subword lists which are then converted into a cohort of all possible valid spellings. A match occurs if the correct word is in the spelling cohort, and we report accuracies on cohorts of sizes 10, 20, and 100, as well as on the whole spelling cohorts. As illustrated in Figure 3, the inclusion of only the OOV$_{55k}$ words in the subword LM training data results in a substantial improvement in performance (i.e. 60% to 69% for top 10 accuracy). Only a slight degradation is incurred with the full 300k lexicon (i.e. 69% to 68% for top 10 accuracy).

Subword *N*-best length: The computational requirements of the subword model can be significantly reduced with a smaller subword *N*-best list, so it is of interest to measure degradation in performance as a function of *N*-best length, *N*. As illustrated in Figure 4, modest degradation is incurred in the top 10 accuracy as *N* is decreased from 1000 to 100 (69% to 66%).

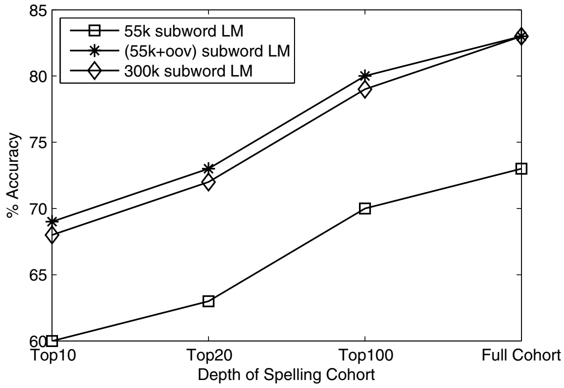

Figure 3: Accuracy of the three subword recognizers for different depths of the spelling cohort evaluated on the 1454 OOV$_{55k}$ words. The spellings are generated with a subword 1000-best list.

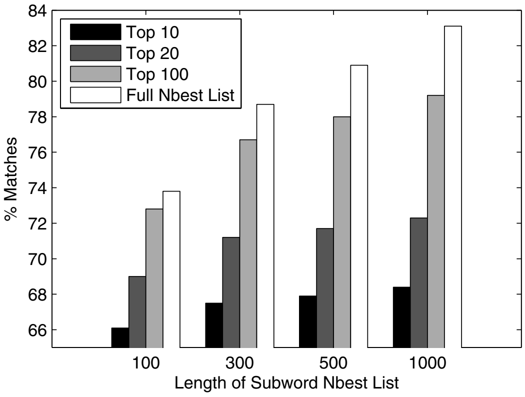

Figure 4: The subword model accuracy as a function of the length of the *N*-best list. Accuracy is reported on spelling cohorts of size 10, 20, and 100, as well as on the full spelling cohort. The 300k LM subword recognizer is used.

Next, we evaluate the subword model in a simple dialogue system, where the user speaks a single word and a word recognizer generates a 10-best list of words. If the correct word is not in the 10-best list, the subword model is automatically triggered, and a spelling cohort of size 10 is generated. If the correct word is not in the cohort, the user is asked to spell the word.

Currently, we have designed an online user interface that implements this multi-stage isolated word recognizer. However, we have not yet collected user data to evaluate the system. For this reason, we use the 4682 Phonebook nouns to simulate words spoken by users. A 55k word recognizer is used in Stage I, and all words that fail to appear in the 10-best list are passed to the subword model in Stage II. In this research we focus on the estimation of the spelling and pronunciation of an OOV, not on the detection of an OOV word. Thus, we rely on direct user feedback to achieve perfect OOV detection. In our experiments, this is simulated by automatically passing all words that failed Stage I to Stage II.

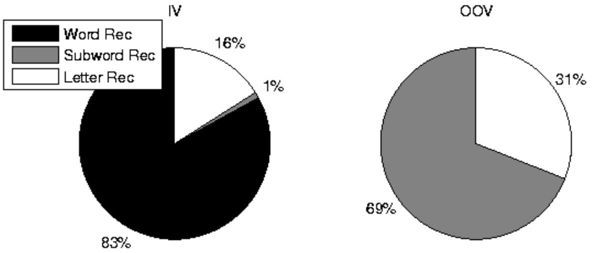

Figure 5: Accuracy of the word and subword recognition stages for a spelling cohort of size ten evaluated on IV$_{55k}$ and OOV$_{55k}$ words.

Multi-Stage System In this section, we evaluate the overall performance of the multi-stage recognizer for IV$_{55k}$ and OOV$_{55k}$ and OOV words. The 55k word recognizer is used in Stage I, and the 300k LM subword recognizer with a 1000-best list of subwords is used in Stage II. The pie charts in Figure 5 describe the percentage of matching words in a spelling cohort of size ten for the word and subword recognition stages versus words that require a spoken spelling mode (letter recognition). For the IV$_{55k}$ words, Stage I proposes the correct word among the top 10 word candidates 83% of the time. If the correct word is not in the top 10, the system reverts to the subword model in Stage II. Stage II recovers an additional 1% of the IV$_{55k}$ words, which now make the top-10 cut due to the availability of alternative pronunciations beyond the baseform supplied in the lexicon. The top 10 accuracy of Stage II on the OOV$_{55k}$ words is 69%. We note that the top 10 list of Stage II excludes any results from Stage I. Hence, we can compare the overall accuracy of Stages I and II to the top 20 accuracy of the 300k isolated word recognizer shown in Table 1. The overall accuracy of the first two stages is 79%, which outperforms the top 20 accuracy of the 300k isolated word recognizer (77%), most probably due to the more focused 55k word recognizer in the first stage.

Perplexity Experiments The analysis of results for the spoken spelling mode is being deferred until we have collected user data. However, in this section, we address the question of whether the subword model can improve the performance of the spoken spelling mode. The spelling mode is based on a letter recognizer and requires a letter LM. We propose and evaluate two letter trigrams: (1) a small letter trigram built from the 100 top candidates in the spelling cohort produced by the subword model, and (2) a large letter trigram built with the 300k lexicon. The two trigrams are evaluated in terms of mean perplexity on the OOV$_{55k}$ words that failed to be recognized by the subword model. Mean perplexities of 11.7 and 16 are obtained for the 100 and 300k letter trigrams respectively. The 27% relative reduction in perplexity achieved by the small spelling cohort should hopefully propagate to an improved letter error rate. An added benefit is a dramatically reduced LM size.

6. Summary

We presented a subword model that can estimate the pronunciation and spelling of a new word. We first showed it is effective in automatically generating baseforms files required for building word recognizers. We also implemented the subword model in a multi-stage isolated-word recognizer, and reported on preliminary results, comparing it with a more traditional, isolated word recognizer. Finally, we used the spelling cohort produced by the subword model to build a letter trigram for the spoken spelling stage. The small letter trigram achieves a lower mean perplexity than a 300k trigram on a subset of the OOV$_{55k}$ words.

In the future, we plan to investigate in more detail the model's capability to dynamically generate and update a baseforms file. We will also implement a live system that includes a word, subword and back-up letter recognizer. We would also like to embed a subword-based OOV model within a continuous speech recognizer. This would involve more challenging issues such as the correct detection of an OOV word occurrence.

7. References

[1] Bauer J. G. and Junkawitsch J., "Accurate recognition of City Names with Spelling as a Fall Back Strategy", *Proc. Eurospeech*, Budapest, Hungary, 263–266, 1999.

[2] Bazzi I. and Glass J. R., "Learning Units for Domain-Independent Out-of-Vocabulary Word Modeling", *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[3] Bisani M. and Ney H., "Open Vocabulary Speech Recognition with Flat Hybrid Models", *Proc. Interspeech*, Lisbon, Portugal, 2005.

[4] Chang E., Seide F., Meng H. M., Chen Z., Shi Y., and Li Y., "A System for Spoken Query Information Retrieval on Mobile Devices", *IEEE Trans. on Speech and Audio Proc.*, 10(8):531–541, 2002.

[5] Chen S. F., "Conditional and Joint Models for Grapheme-to-Phoneme Conversion", *Proc. Interspeech*, 2033-2013, Geneva, Switzerland, 2003.

[6] Filisko E. and Seneff S., "Developing City Name Acquisition Strategies in Spoken Dialogue Systems Via User Simulation", *Proc. SIGDIAL*, Lisbon, Portugal, 2005.

[7] Galescu L. and Allen J., "Name Pronunciation with a Joint N-Gram Model for Bi-directional Grapheme-to-Phoneme Conversion", *Proc. ICSLP*, Denver, Colorado, 2002.

[8] Glass J. R., "A Probabilistic Framework for Segment-Based Speech Recognition", *Computer Speech and Language*, 113-127, 2003.

[9] Gorin A., Riccardi G., and Wright J., "How May I Help You?", *Speech Communication*, 23:113–127, 1997.

[10] Hetherington L., "The MIT Finite-State Transducer Toolkit for Speech and Language Processing", *Proc. ICSLP*, Jeju, South Korea, 2004.

[11] Muthusamy Y., Agarwal R., Gong Y., and Viswanathan V., "Speech-Enabled Information Retrieval In The Automobile Environment", *Proc. ICASSP*, 2259–2262, Phoenix, AZ, 1999.

[12] Pitrelli J., Fong C., Wong S., Spitz J., and Leung H., "Phonebook: A Phonetically-Rich Isolated-Word Telephone-Speech Database", *Proc. ICASSP*, 101–104, Detroit, MI, 1995.

[13] Seneff S., "Reversible Sound-to-letter/Letter-to-sound Modeling based on Syllable Structure", *Proc. NAACL-HLT*, Rochester, NY, 2007.

[14] Yazgan A. and Saraclar M., "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition". *Proc. ICASSP*, 745–748, Montreal, Canada, 2004.

[15] Zue V., Seneff S., Glass J., Polifroni J., Pao C., Hazen T. J., and Hetherington L., "Jupiter: A Tephone-Based Conversational Interface for Weather Information". *IEEE Trans. on Speech and Audio Proc.*, 8(1):85–96, 2000.