IMPROVED KEYWORD-SPOTTING USING SRI'S DECIPHER™ LARGE-VOCABUARLY SPEECH-RECOGNITION SYSTEM

Mitchel Weintraub

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

ABSTRACT

The word-spotting task is analogous to text-based information retrieval tasks and message-understanding tasks in that an exhaustive accounting of the input is not required: only a useful subset of the full information need be extracted in the task. Traditional approaches have focussed on the keywords involved. We have shown that accounting for more of the data, by using a large-vocabulary recognizer for the wordspotting task, can lead to dramatic improvements relative to traditional approaches. This result may well be generalizable to the analogous text-based tasks.

The approach described makes several novel contributions, including: (1) a method for dramatic improvement in the FOM (figure of merit) for word-spotting results compared to more traditional approaches; (2) a demonstration of the benefit of language modeling in keyword spotting systems; and (3) a method that provides rapid porting of to new keyword vocabularies.

1. INTRODUCTION

Although both continuous speech recognition and keyword-spotting tasks use the very similar underlying technology, there are typically significant differences in the way in which the technology is developed and used for the two applications (e.g. acoustic model training, model topology and language modeling, filler models, search, and scoring). A number of HMM-based systems have previously been developed for keyword-spotting [1-5]. One of the most significant differences between these keyword-spotting systems and a CSR system is the type of non-keyword model that is used. It is generally thought that very simple non-keyword models (such as a single 10-state model [2], or the set of monophone models [1]) can perform as well as more complicated non-keyword models which include words or triphones.

We describe how we have applied CSR techniques to the keyword-spotting task by using a speech recognition system to generate a transcription of the incoming spontaneous speech which is searched for the keywords. For this task we have used SRI's DECIPHER™ system, a state-of-the-art large-vocabulary speaker-independent continuous-speech recognition system [6-10]. The method is evaluated on two domains: (1) the Air Travel Information System (ATIS) domain [13], and (2) the "credit card topic" subset of the Switchboard Corpus [11], a telephone speech corpus consisting of spontaneous conversation on a number of different topics.

In the ATIS domain, for 78 keywords in a vocabulary of 1200, we show that the CSR approach significantly outperforms the traditional wordspotting approach for all false alarm rates per hour per word: the figure of merit (FOM) for the CSR recognizer is 75.9 compared to only 48.8 for the spotting recognizer. In the Credit Card task, the spotting of 20 keywords and their 58 variants on a subset of the Switchboard corpus, the system's performance levels off at a 66% detection rate, limited by the system's ability to increase the false alarm rate. Additional experiments show that varying the vocabulary size from medium- to large-vocabulary recognition systems (700 to 7000) does not affect the FOM performance.

A set of experiments compares two topologies: (1) a topology for a fixed vocabulary for the keywords and the N most common words in that task (N varies from Zero to Vocabulary Size), forcing the recognition hypothesis to choose among the allowable words (traditional CSR), and (2) a second topology in which a background word model is added to the word list, thereby allowing the recognition system to transcribe parts of the incoming speech signal as background. While including the background word model does increase the overall likelihood of the recognized transcription, the probability of using the background model is highly likely (due to the language model probabilities of out of vocabulary words) and tended to replace a number of keywords that had poor acoustic matches.

Finally, we introduce an algorithm for smoothing language model probabilities. This algorithm combines small task-specific language model training data with large task-independent language training data, and provided a 14% reduction in test set perplexity.

2. TRAINING

2.1. Acoustic Modeling

DECIPHER™ uses a hierarchy of phonetic context-dependent models, including word-specific, triphone, generalized-triphone, biphone, generalized-biphone, and context independent models. Six spectral features are used to model the speech signal: the cepstral vector (C1-CN) and its first and second derivatives, and cepstral energy (C0) and its first and second derivatives. These features are computed from an FFT filterbank and subsequent high-pass RASTA filtering of the filterbank log

energies, and are modeled either with VQ and scalar codebooks or with tied-mixture Gaussian models. The acoustic models used for the Switchboard task use no cross word acoustic constraints.

2.2. Language Modeling

The DECIPHER™ system uses a probabilistic finite state grammar (PFSG) to constrain allowable word sequences. In the ATIS, WSJ, and Credit Card tasks, we use a word-based bigram grammar, with the language model probabilities estimated using Katz's back-off bigram algorithm [12]. All words that are not in the specified vocabulary that are in the language model training data are mapped to the *background* word model. The *background* word model is treated like all the other words in the recognizer, with bigram language model probabilities on the grammar transitions between words.

Two topologies are used for the experiments described in this paper. One topology is to use a fixed vocabulary with the keywords and the N most common words in that task (N varies from Zero to VocabSize), forcing the recognition hypothesis to choose among the allowable words. A second topology is to add the *background* word model to the above word list, thereby allowing the recognition system to transcribe parts of the incoming speech signal as *background*. A sample *background* word with 60 context-independent phones is shown below in Figure 1.

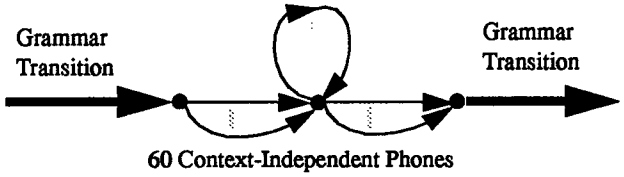

Figure 1: A sample topology for the *background* word model. The minimum duration is 2 phones and the self loop allows for an infinite duration.

2.3. Task-Specific Language Model Estimation

The Switchboard Corpus [11] is a telephone database consisting of spontaneous conversation on a number of different topics. The Credit Card task is to spot 20 keywords and their variants where both the keywords and the test set focus on a subset of the Switchboard conversations pertaining to credit cards. To estimate the language model for this task, we could (1) use a small amount of task-specific training data that focuses only on the credit card topic, (2) use a large amount of task-independent training data, or (3) combine the task-specific training with the task-independent training data.

For combining a small amount of task-specific (TS) training with a very large amount of task-independent (TI) training data, we modified the Katz back-off bigram estimation algorithm [12]. A weight was added to reduce the effective size of the task-independent training database as shown in Equation 1:

$$C(w2\ ,\ w1) = C_{TS}(w2\ ,\ w1) + \gamma \bullet C_{TI}(w2\ ,\ w1)$$

where $C(w2\ ,\ w1)$ is the counts of the number of occurrences of word w1 followed by w2, $C_{TS}(w2\ ,\ w1)$ are the counts from the task-specific database and $C_{TI}(w2\ ,\ w1)$ are the counts from the task-independent database. The weight γ reduces the effective size of the task-independent database so that these counts don't overwhelm the counts of the task-specific database.

Table 1 shows both the training set and test set perplexity for the credit card task as a function of γ. The task-specific training consisted of 18 credit card conversations (59 K words) while the task-independent training consisted of 1123 general conversations (17 M words).

Table 1: Perplexity of Credit Card Task as a Function of Task Independent-Specific Smoothing

γ	Effective Task Indep. Training Size	Training Set Perplexity	Test Set Perplexity
1.0	17,611,159	174.7	380.0
0.5	8,805,579	154.5	358.3
0.2	3,352,223	131.0	332.0
0.1	1,761,116	117.5	321.8
0.05	880,558	109.7	328.8
0.02	352,223	102.6	360.4
0.01	176,111	98.8	396.9
0.005	88,055	96.2	443.4
0.002	35,222	94.5	521.5
0.001	17,611	94.0	592.3

3. SEARCH

The DECIPHER™ system uses a time-synchronous beam search. A partial Viterbi backtrace [6] is used to locate the most-likely Viterbi path in a continuous running utterance. The Viterbi backtrace contains both language model information (grammar transition probabilities into and out of the keyword), acoustic log likelihood probabilities for the keyword, and the duration of the keyword hypothesis.

A duration-normalized likelihood score for each keyword is computed using the following Equation 2:

$$KeyScore = \frac{AP + GP + \text{Constant}}{Duration}$$

where AP is the acoustic log-likelihood score for the keyword, and GP is the log probability of the grammar transition into the keyword, and Constant is a constant added to the score to penalize keyword hypotheses that have a short duration. None of the earlier HMM keyword systems used a bigram language in either the decoding or the scoring. Many previous systems did use weights on the keywords to adjust the operating location on the ROC curve.

A hypothesized keyword is scored as correct if its mid-point falls within the endpoints of the correct keyword. The keyword scores are used to sort the occurrences of each keyword for computing the probability of detection at different false-alarm levels. The overall figure-of-merit is computed as the average detection rate over all words and over all false alarm rates up to ten false alarms per word per hour.

4. EXPERIMENTS

4.1. ATIS Task

The ATIS task [13] was chosen for keyword-spotting experiments because (1) the template-based system that interprets the queries of the airline database focuses on certain keywords that convey the meaning of the query, and ignores many of the other filler words (e.g. "I would like...", "Can you please ..."), (2) the task uses spontaneous speech, and (3) we have worked extensively on this recognition task over the last two years. Sixty-six keywords and their variants were selected as keywords based on the importance of each of the words to the SRI template-matcher which interprets the queries.

SRI applied two different recognition systems to the ATIS keyword spotting task. The first system was SRI's large-vocabulary speaker-independent speech recognition system that we have used for the ATIS speech-recognition task [3]. The vocabulary used in this system is about 1200 words, and a back-off bigram language model was trained using the ATIS MADCOW training data [13]. Many of the words in the vocabulary use word-specific or triphone acoustic models, with biphone and context-independent models used for those words that occur infrequently.

The second system is a more traditional word-spotting system. There are 66 keywords plus 12 variants of those keywords for a total of 78 keyword models. There is a *background* model (see Figure 1) that tries to account for the rest of the observed acoustics, making a total of 79 words in this second system. This second system also uses a back-off bigram grammar, but all non-keywords are replaced with the *background* word when computing language model probabilities.

The acoustic models for the keywords and their variants were identical in the two systems. The only difference between the two systems is that the first system uses ~1100 additional words for the *background* model, while the second system uses one *background* model with 60 context-independent phones. The resulting FOM and ROC curves are show in Figure 2 for the two systems.

Table 2: ATIS Keyword Spotting Results

System Description	Number of Filler Models	FOM
ATIS Recognizer	1100	75.9
Spotting Recognizer	1	48.8

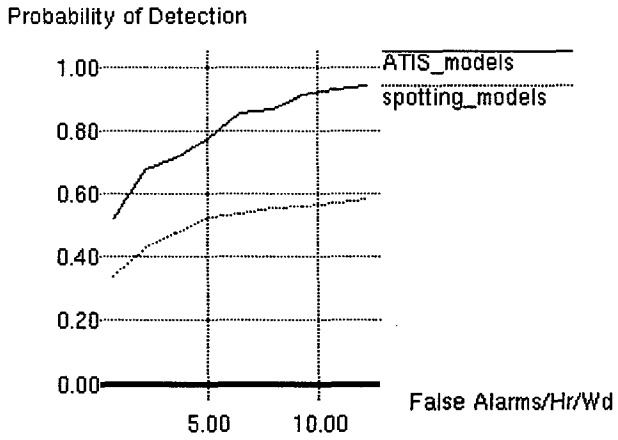

Figure 2: Probability of detection as a function of the false alarm rate for the above two CSR systems on the ATIS Task.

There are two possible explanations for the experimental results in Figure 2 and Table 2. The first explanation is that the ATIS recognizer has a much larger vocabulary, and this larger vocabulary is potentially better able at matching the non-keyword acoustics than the simple *background* model. The second explanation is that for the larger vocabulary ATIS system, the back-off bigram grammar can provide more interword constraints to eliminate false alarms than the back-off bigram grammar that maps all non-keywords to the filler model. Additional experiments are planned to determine the extent of these effects.

4.2. Credit Card Task

The Credit Card task is to spot 20 keywords and their 58 variants on a subset of the Switchboard database. The keywords were selected to be content words relevant to the credit card topic and based on adequate frequency of occurrence of each keyword for training and testing.

Acoustic models were trained on an 11,290 hand-transcribed utterances subset of the Switchboard database. A back-off bigram language model was trained as described in Section 2.3, using the text transcriptions from 1123 non-credit-card conversations and 35 credit card conversations. The most common 5,000 words in the non-credit-card conversations were combined with the words in the credit card conversations, the keywords, and their variants to bring the recognition vocabulary size to 6914 words (including the *background* word model).

The resulting CSR system was tested on 10 credit-card conversations from the Switchboard database. Each conversation consisted of two stereo recordings (each talker was recorded separately) and was approximately 5 minutes long. Each of the two channels is processed independently. The resulting ROC curve is shown in Figure 3. The ROC curve levels out at 66% because the CSR system hypothesized 431 keywords out of a total of 498 true keyword locations. Our current CSR approach, which uses the Viterbi backtrace, does not allow us to increase the keyword false alarm rate.

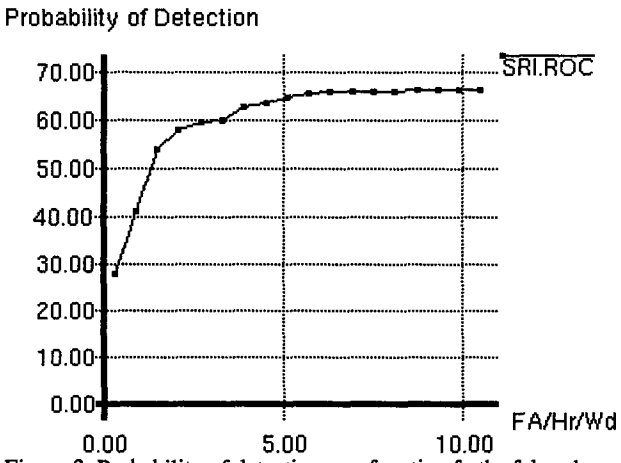

Figure 3: Probability of detection as a function fo the false alarm rate for the 6914 word CSR system on the Credit Card Task.

The effect of using different scoring formulas is shown in Table 3. If only the duration-normalized acoustic log-likelihoods are used, an average probability of detection (FOM) of 54% is achieved. When the grammar transition log-probability into this keyword is added to the score (Eqn 2), the FOM increases to 59.9%. In addition, if a constant is added to the score before normalization, the FOM increases for both cases. This has the effect of reducing the false-alarm rate for shorter-duration keyword hypotheses. We have not had a chance to experiment with the grammar transition leaving the keyword, nor with any weighting of grammar scores relative to acoustic scores.

Table 3: Credit Card FOM Scoring

	Acoustic Likelihood + Grammar Transition	Acoustic Likelihood
Keyword Score	59.9	54.0
Optimized Score	60.5	57.1

We then varied the recognition vocabulary size and determined its effect on the keyword-spotting performance. These experiments show that varying the vocabulary size from medium- to large-vocabulary recognition systems (700 to 7000) does not affect the FOM performance.

Table 4: Credit Card FOM as a Function of CSR Vocabulary Size

Vocabulary Size	FOM
725	59.3
1423	59.5
6914	59.9

Finally, we experimented with including or excluding the *background* word model in the CSR lexicon. While including the *background* word model does increase the overall likelihood of the recognized transcription, the probability of using the *background* model is highly likely (due to the language model probabilities of OOV words) and tended to replace a number of keywords that had poor acoustic matches. Table 5 shows that a slight improvement can be gained by eliminating this *background* word model.

Table 5: FOM With and Without *Background* Model for Large Vocabulary CSR System

Vocabulary Size	FOM
6914	59.9
6913 (No Background)	61.6

5. SUMMARY

This paper describes how SRI has applied our speaker-independent large-vocabulary CSR system (DECIPHER™) to the keyword-spotting task. A transcription is generated for the incoming spontaneous speech by using a CSR system, and any keywords that occur in the transcription are hypothesized. We show that the use of improved models of non-keyword speech with a CSR system can yield significantly improved keyword spotting performance.

The algorithm for computing the score of a keyword combine information from acoustic, language, and duration. One key limitation of this approach is that keywords are only hypothesized if they are included in the Viterbi backtrace. This does not allow the system builder to operate effectively at high false alarm levels if desired. We are considering other algorithms for hypothesizing *"good score"* keywords that are on high scoring paths.

We introduced an algorithm for smoothing language model probabilities. This algorithm combines small task-specific language model training data with large task-independent language training data, and provided a 14% reduction in test set perplexity.

The use of a large-vocabulary continuous-speech recognition system allows the system designer a great deal of flexibility in choosing the keywords that they would like to select for the particular application. If the desired keyword is already in the lexicon, then searching for the keyword can be achieved by looking for the word in the transcription generated by the recognizer. If the word is not in the lexicon, the word can be easily added to the system since triphone models have already been trained.

The ability to transcribe spontaneous speech and search for relevant keywords will play an important role in the future development of simple spoken language applications. Such systems will be easily portable to new domains. Since the operating point for our speech recognizer is typically one which has a low insertion rate, there is little chance for a keyword false alarm. Future experimentation will determine the effectiveness of such understanding systems for human-computer interaction.

REFERENCES

1. R. Rose and D. Paul, "A Hidden Markov Model Based Keyword Recognition System," 1990 *IEEE ICASSP*, pp. 129-132.
2. J.G. Wilpon, L.R. Rabiner, C.H. Lee, and E.R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," 1990 *IEEE Trans. ASSP*, Vol 38. No. 11, pp. 1870-1878.
3. J.G. Wilpon, L.G. Miller, and P. Modi, "Improvements and Applications for Key Word Recognition Using Hidden Markov Modeling Techniques," 1991 *IEEE ICASSP*, pp. 309-312.
4. R. Rohlicek, W. Russell, S. Roukos, H. Gish, "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," 1989 *IEEE ICASSP*, pp. 627-630.
5. L.D. Wilcox, and M.A. Bush, "Training and Search Algorithms for an Interactive Wordsptting System," 1992 *IEEE ICASSP*, pp. II-97-II-100.
6. H. Murveit, J. Butzberger, and M. Weintraub, "Performance of SRI's DECIPHER Speech Recognition System on DARPA's CSR Task," 1992 DARPA Speech and Natural Language Workshop Proceedings, pp 410-414
7. Murveit, H., J. Butzberger, and M. Weintraub, "Reduced Channel Dependence for Speech Recognition," 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 280-284.
8. Butzberger, J., H. Murveit, E. Shriberg, and P. Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 339-343.
9. H. Murveit, J. Butzberger, and M. Weintraub, "Speech Recognition in SRI's Resource Management and ATIS Systems," 1991 DARAP Speech and Natural Language Workshop, pp. 94-100.
10. Cohen, M., H. Murveit, J. Bernstein, P. Price, and M. Weintraub, "The DECIPHER™ Speech Recognition System," *1990 IEEE ICASSP*, pp. 77-80.
11. J.J. Godfrey, E.C. Holliman, and J.McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," 1992 *IEEE ICASSP*, pp. I-517-I-520.
12. S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," 1987 *IEEE ASSP*, Vol. 35, No. 3. pp.400-401.
13. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 7-14.