A SEGMENT-BASED WORDSPOTTER USING PHONETIC FILLER MODELS

Alexandros S. Manos and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

A common approach to wordspotting is to augment the keyword models with "filler" models to account for non-keyword intervals. An alternative approach is to use a large vocabulary continuous speech recognition system (LVCSR) to produce a word string, and then search for the keywords in that string. While the latter approach typically yields higher performance, it requires costly computation and extensive training data. In this study, we develop several segment-based wordspotters in an effort to achieve performance comparable to that of the LVCSR spotter, but with only a fraction of the vocabulary. We investigate several methods to model the background, ranging from a few general models to refined phone representations. The task is to detect sixty-one keywords from continuous speech in the ATIS domain. The best performance we achieve is 91.4% Figure of Merit for the LVCSR spotter and 86.7% for a spotter using 57 phone-based filler models.

1. INTRODUCTION

The task of wordspotting systems is to detect (a small set) of keywords from a speech stream. The research challenge for wordspotting is one of achieving the highest possible keyword detection rate while minimizing the number of keyword insertions. Therefore, it is not sufficient to model only the keywords very explicitly; models of the background are also required. Most of the wordspotters developed in recent years were variants of HMM-based, continuous speech recognition systems [1, 2, 3, 4]. In these systems, the non-keyword intervals were represented by a variety of "filler" models, ranging from a few phonetic or syllabic fillers to whole words. It was shown that more explicit modeling of the non-keyword speech stream improves wordspotting performance. The benefits of incorporating a language model for the transitions between the keywords and the filler models were also evaluated for some of the systems [1, 2, 4], and were found to be substantial. As a general result, the large vocabulary continuous speech recogniton (LVCSR) systems with a language model component significantly outperformed any other configuration. However, the LVCSR approach to wordspotting, even though providing the best performance, has two important disadvantages, (1) it is computationally very expensive, and (2) it tends to be domain dependent, requiring knowledge of the full vocabulary, and a large body of training data.

In this paper, we describe our investigation into the use of different background models in an effort to achieve computational efficiency and maintain domain independence, while establishing acceptable wordspotting performance compared to the LVCSR wordspotters. Due to space limitations, readers are referred to [5] for further details.

This research is supported by BellSouth Intelliventures, and by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control, and Ocean Surveillance Center.

2. EXPERIMENTAL FRAMEWORK

2.1. System Description

The wordspotters described in this paper are segment-based; they are derived from the SUMMIT continuous speech recognition system [6]. The recognition network for the wordspotters is shown in Figure 1 for N keywords and M filler models.

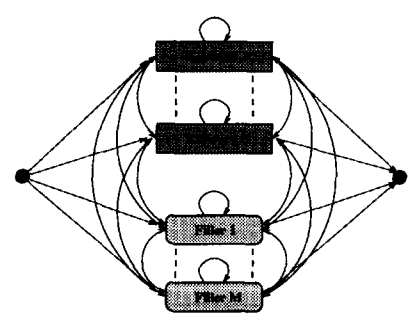

Figure 1: Recognition network for the wordspotting systems.

Any transition between keywords and fillers is allowed, as well as self transitions for both keywords and fillers. This configuration allows multiple keywords to exist in any one utterance, as well as multiple instances of a keyword within the same utterance. For the experiments described in the next section we used 1, 12, 18, 57 and 2462 filler models combined with 61 keywords in this configuration.

2.2. Signal Representation and Features

The input signal is transformed into a sequence of 5 ms frames, and each frame is characterized by 14 Mel-Frequency Cepstral Coefficients (MFCCs). The signal is then segmented into acoustically homogeneous intervals using a hierarchical algorithm, creating a segment network. A feature vector is then computed for each segment in the network. The vector consists of a segment duration measurement and 35 MFCC averages computed within and across segment boundaries.

2.3. Keyword Models

The keywords were represented by concatenations of phonetic units. They were expanded into a pronunciation network based on a set of phonological rules. Two sets of phonetic units were used in the description of the keywords, context-independent phones and word-dependent phones, in distinct experiments. The models for these units consisted of mixtures of up to 25 diagonal Gaussians in the 36-dimensional space defined by the measurements.

2.4. Filler Models

We examined the tradeoff between performance and computation time for five sets of filler models. In the LVCSR approach we explicitly modeled all 2462 non-keyword words as fillers. In the *CI-filler* approach we represented the background with 57 context-independent phone-like words. The remaining three filler sets consisted of 18, 12 and 1 models that were derived through clustering of the context-independent phones.

2.5. Language Modeling

We propose a new approach to the construction of the language model component. In previous research, when context-independent phones or more general acoustic models were used for background representation, they were all grouped into a single filler model. Thus, only a single grammar transition probability into and out of the filler was computed. In our approach, every acoustic model corresponds to a unique filler model. Using the LVCSR system and the orthographic transcriptions available from the training data, we performed forced alignments that produced transcriptions consisting of phones for the non-keyword words, and whole words for the keywords. These transcriptions were used to train the bigram language model for the keywords and the acoustic filler models. The LVCSR also used a bigram language model. Training was performed on 10,000 utterances for all wordspotting systems.

2.6. Search

The Viterbi algorithm is used to find the best path through the labeled segment network, with the pronunciation network and the language model serving as constraints. The output is a continuous stream of fillers and keywords. The score for each hypothesized keyword is calculated as the sum, over all segments composing the keyword, of (1) the segment's phonetic match score, (2) the score based on the probability of the particular segmentation, (3) a lexical weight associated with the likelihood of the pronunciation, (4) a duration score based on the phone duration statistics, and (5) a bigram transition score.

3. EXPERIMENTS

3.1. Task

All experiments were performed in the Air Travel Information Service, or ATIS, domain [7]. The task was the detection of 61 keywords in unconstrained speech. The keywords consisted of city names, airlines, days of the week, fare types, etc. They were chosen out of the ATIS vocabulary based on their high frequency of occurrence, and the observation that they may constitute a sufficient set for a hypothetical spoken language system that will enable a client to fill out an electronic form, using speech, with information such as desired origin and destination point, fare basis, and day of departure. The sets for training and testing (see Table 1) were derived from all available data for the ATIS task. They were specifically designed to contain all keywords in balanced proportions.

	# keywords	# utterances	# speakers
Training set	15076	10000	584
Test set	2222	1397	36

Table 1: Training and test sets in the ATIS domain.

3.2. Performance Measures

The performance of the proposed wordspotting systems was measured using conventional Receiver Operating Characteristic (ROC) curves and Figure of Merit (FOM) calculations. A keyword was considered successfully detected if the midpoint of the hypothesized word fell within the reference time interval. The hypothesized keywords were sorted with respect to their scores, and the probability of detection at each false alarm rate was computed. The FOM was calculated as the average probability of detection between 0 and 10 false alarms per keyword per hour. The average computation time per utterance was also measured. We used the actual computation time when comparing between the systems since it demonstrated less fluctuation than the elapsed time. All timing experiments were performed on a Sun SPARCstation-20 with two 50MHz processors and 128MB of RAM.

3.3. LVCSR and CI-Filler Wordspotters

The LVCSR wordspotter was developed first in order to serve as a benchmark against which the performance of all other spotters would be evaluated. The background representation consisted of 2462 words. Both keywords and background words were modeled as concatenations of context-independent phones, and were expanded into a pronunciation network. The LVCSR system achieved 89.8% FOM on this set of keywords. The tradeoff for this outstanding

wordspotting performance was the rather long computation time required due to the size of the vocabulary.

The vocabulary for the CI-filler system consisted of the 61 keywords and the 57 context-independent phone-words. The output of this continuous speech recognition system is a sequence of phone-like words and keywords. There are three factors that control the decision of hypothesizing a keyword versus hypothesizing the underlying string of phones. The first one is the combined effect of two trainable parameters, the word and segment transition weights (*wtw* and *stw*). The *wtw* corresponds to a penalty for the transition into a new word, while the *stw* is a bonus for entering a new segment. These parameters acquire appropriate values during a corrective training process that attempts to equalize the number of words in the reference string and the hypothesized string. The second factor is the bigram transition score, which consists only of the transition score into the keyword in the first case, versus the sum of the bigram transition scores for the underlying string of phones in the second case. Finally, the arcs representing transitions between phones within the keywords carry weights that are added to the keyword score. Since these arc-weights can be either positive or negative, depending on the likelihood of the pronunciation path to which they belong, they can influence the keyword hypothesis either way.

The CI-filler system achieved 81.8% FOM, approximately 8% lower in absolute value than that of the LVCSR system. However, the computation time required for the Viterbi stage of this system was approximately seven times less than that of the LVCSR. These results encouraged us to search for an even smaller set of filler models for background representation. The advantages of a smaller set are less computation time and more flexibility, in the sense that wordspotting in a new domain would require less training data for language and acoustic modeling.

3.4. General Filler Models

We designed three sets of general fillers consisting of 18, 12 and 1 acoustic models. The general fillers were derived by (supervised) clustering of the 57 context-independent phones, based on their acoustic feature vectors. These classes mostly correspond to broad phonetic classes (i.e., nasals, closures, stops, etc.), thus agreeing with our acoustic phonetic intuitions.

A bigram language model was computed for each one of the systems using the general filler models. It was trained by replacing, for each sentence, the context-independent phones used to represent the non-keyword intervals with the corresponding cluster label, while keeping the keywords intact. The wordspotter with 18 filler models achieved 79.2% FOM performance, compared to 76.5% for the 12-filler system and 61.4% for the 1-filler system. The ROC curves for these systems, as well as for the LVCSR and CI-filler spotters, are shown in Figure 2.

3.5. Word-Dependent Models for Keywords

In a final set of experiments, we studied the effects of introducing word-dependent phones for the keywords on FOM performance and computation time. The word-dependent phones were trained from keyword instances only, while the

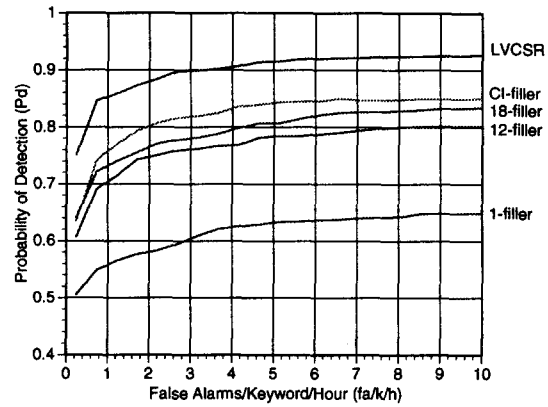

Figure 2: ROC-curves for the wordspotters with different numbers of filler models.

context-independent phones for the non-keyword words or filler models were trained from non-keyword speech only. The final score for each word-dependent phone model was linearly interpolated with the score of the corresponding context-independent phone. The interpolation weights were computed as a function of the frequency of each word-dependent model in the training set. The FOM performance for the LVCSR system increased by 1.6% in absolute value to 91.4%. An increase of 4.9% (to 86.7%) in the FOM was achieved for the CI-filler spotter with the use of word-dependent models for the keywords. The ROC curves for these systems are shown in Figure 3.

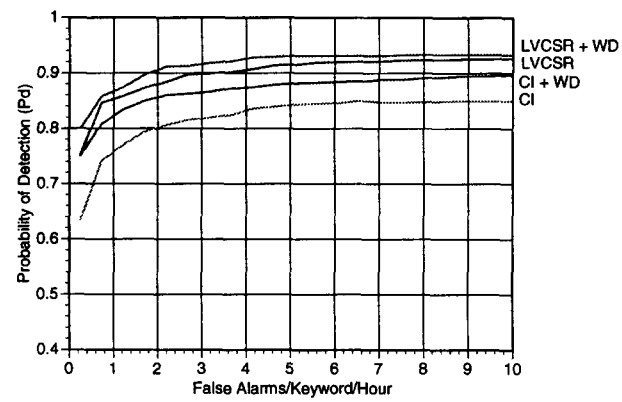

Figure 3: ROC-curves for the LVCSR and CI-filler wordspotters with and without word-dependent models.

While the Viterbi computation time remained almost unchanged for both systems, the classification time increased substantially as a result of the algorithm that we used for these experiments. This classification algorithm computes the score for all acoustic models, for all segments before the Viterbi search is initiated. An algorithm that computes acoustic scores upon demand during the search would save a

lot of computation, and would make word-dependent models more attractive.

4. DISCUSSION

There is clearly a correlation between the degree of explicitness in background modeling and wordspotting performance as measured by the FOM. The LVCSR utilizes the most detailed filler models, i.e., whole words, and achieves the highest performance of all spotters. As filler models become fewer and more general, the FOM decreases monotonically (c.f., Table 2 and Figure 4). The LVCSR system outperforms the spotter that uses only a single filler model by almost thirty percent in absolute FOM value. The largest portion of this performance gain can be attributed to the use of more refined acoustic models for the background. An increase of 20.4% in the FOM is achieved when the number of filler models is increased from one general acoustic model to fifty-seven context-independent phones. This result suggests that the use of more refined phone representations, such as context-dependent phones, could further improve the FOM. The remaining 8% gain in performance is achieved by incorporating domain specific knowledge, i.e., using models of all non-keyword words as fillers. This further improvement can be attributed to a more constrained search space and a more effective bigram component. For instance, the probability that the current word is a city name, given that the previously hypothesized word was "from," is much higher than if the previous word was the single filler model.

Wordspotter	CI models	WD models
LVCSR	89.8%	91.4%
CI fillers	81.8%	86.7%
18 fillers	79.2%	-
12 fillers	76.5%	-
1 filler	61.4%	-

Table 2: Summary of FOM performance results.

The average computation time per utterance required by each system is shown in Figure 4. As we expected, the computation required for the Viterbi stage decreased with the number of filler models. Compared to the LVCSR, the CI-filler system decreased the Viterbi computation time by approximately a factor of seven, the 18 and 12-filler systems by a factor of twelve, and the 1-filler system by a factor of 23. The classification time varied with the number of acoustic models, due to the specific algorithm that was used. As we already discussed earlier, the computation required for this stage can be significantly reduced with the use of a more sophisticated algorithm.

5. CONCLUSIONS

There is a clear tradeoff between wordspotting performance as measured by the FOM, and the Viterbi computation time required for spotting. More explicit modeling of the background results in higher performance, but also requires more computation. An acceptable compromise between FOM performance and computation time seems to be the CI-filler system. It achieves over 80% FOM, and provides significant savings in computation compared to the LVCSR spotter.

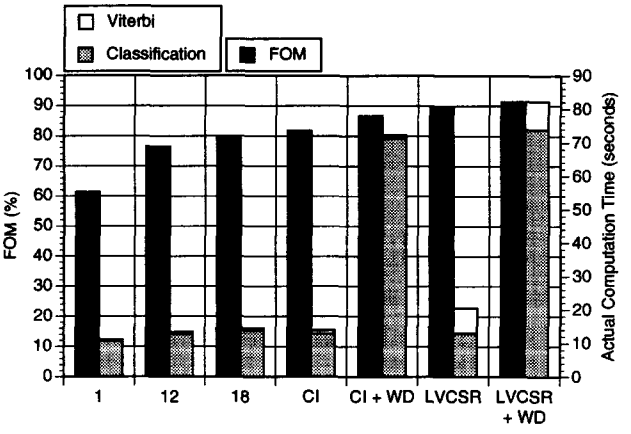

Figure 4: FOM and computation time measurements for all developed wordspotters.

6. REFERENCES

[1] R. Rose, "Definition of subword acoustic units for wordspotting," *Proc. EUROSPEECH'93*, pp. 1049–1052, 1993.

[2] P. Jeanrenaude, K. Ng, M. Siu, J.R. Rohlicek, and H. Gish, "Phonetic-based word spotter: various configurations and application to event spotting," *Proc. EUROSPEECH'93*, pp. 1057–1060, 1993.

[3] E. Lleida, J.B. Marino, J.Salavedra, A. Bonafonte, E. Monte, and A. Martinez, "Out-of-vocabulary word modelling and rejection for keyword spotting," *Proc. EUROSPEECH'93*, pp. 1265–1268, 1993.

[4] M. Weintraub, "Keyword-spotting using SRI's DECIPHER large-vocabulary speech recognition system," *Proc. ICASSP93*, pp. 463–466, 1993.

[5] A. S. Manos, *A Study on Out-of-Vocabulary Word Modelling for a Segment-Based Keyword Spotting System*, SM Thesis, Department of Electrical Engineering and Computer Science, MIT, May, 1996.

[6] V. Zue, J. Glass, D. Goodine, M. Phillips, and S. Seneff, " The SUMMIT speech recognition system: phonological modelling and lexical access," *Proc. ICASSP90*, pp. 49–52. IEEE, 1990.

[7] D. A. Dahl, M. Bates, M. Brown, W.Fisher, K. H. Smith, D. Pallet, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the atis task: The atis-3 corpus," *Proc. DARPA Human Language Technology Workshop*, pp. 43–48, Morgan Kaufmann, March 1994.