A new evaluation criteria for keyword spotting techniques and a new algorithm

Marius C. Silaghi

Department of Computer Science
Florida Institute of Technology
msilaghi@cs.fit.edu

Rachna Vargiya

Department of Computer Science
Florida Institute of Technology
rvargiya@cs.fit.edu

Abstract

Keyword spotting is an efficient approach for search of relevant recordings in databases of recorded unconstrained speech. Many algorithms have been proposed in the past for this problem and several techniques claim to be very efficient and accurate. Researchers have so far attempted to correctly compare their results by using standardized Receiver Operating Characteristic (ROC) curves, and performing experiments on publicly available databases with known keywords.

However, when it comes to compare the expected behavior of a technique for new keywords and utterances, the generalization of published comparisons is not very clear, and the choice of the benchmark-keywords has considerable effects on the comparison. In this paper we propose a new measure of the accuracy of a keyword spotter, removing the benchmark-keywords selection bias and offering a qualitative estimation of how well the technique is expected to perform on new keywords.

We apply our evaluation scheme to compare previously known algorithms as well as a new technique that we propose now. The new technique is based on a confidence measure that evaluates a keyword match to the worst of its phoneme scores (where the score of a phoneme is taken as the ratio between the log probability of that phoneme and the length of the phoneme). It is remarkable that the newly proposed technique can detect all occurencies of 100 keywords with less than .5 false alarms/keyword/hour.

1. Introduction

Keyword spotting (KWS) is the recognition of predefined keywords in unconstrained utterances. It is used in applications that do not require the entire sequence of words to be recognized (e.g. search in speech databases, classification of speech messages). It is observed that the choice of keywords influences the performance of keyword spotting algorithms. Certain words are recognized more easily than other words, making the selection of keywords for fair benchmarking a difficult task. Comparison of two or more techniques on a set of keywords may be inaccurate if the keywords favor one algorithm. Other than proposing a new posterior based confidence measure and pruning techniques, we investigate an alternative evaluation method which attempts to remove the influence of keywords selected for recognition. The contributions of this paper are:

- A new KWS technique based on a novel confidence measure, called Real Fitting is proposed. The measures results in high detection rates.
- Limitations of standard evaluation of keyword spotting techniques are highlighted. A novel evaluation, Equal Opportunity Evaluation, is proposed.

Definition 1 (HMM) *A Hidden Markov Model (HMM) M is defined by a set of states $Q = \{q_1, ..., q_J\}$, with a unique starting and ending state, a space of possible outputs V, a set of transition probabilities $a_{ij} = P(q_i|q_j)$, $\forall i, j \in [1..J]$, and a set of probability density functions defining the likelihood of each possible output for each state, $b_i(v) = P(v|q_i), v \in V$.*

Given a sequence of acoustic vectors $X = \{x_1, \ldots, x_N\}$. The **KWS problem** consists of deciding whether the keyword represented by a HMM M generates any segment X_b^e of X. Often, on detection it is also desired to find the segmentation (i.e., the beginning b and end e) and the sequence $Q = \{q^b, ..., q^e\}$ of states that represents the most probable path in M generating X_b^e.

Typically a keyword spotter decides a match if an algorithm-dependent score of the match is above (or below) some threshold. Assuming a scoring function $S_{M,X}(Q, b, e)$ for the match (path) Q between M and X_b^e, a keyword spotter returns:

$$\begin{cases} \underset{Q,b,e}{\text{argmax}}(S_{M,X}(Q, b, e)) & \text{if} \max\limits_{Q,b,e}(S_{M,X}(Q, b, e)) \leq T \\ fail & \text{if} \max\limits_{Q,b,e}(S_{M,X}(Q, b, e)) > T \end{cases}$$

2. Background

Keyword spotting has a rich history in speech processing. Techniques can use Dynamic Time Warping (DTW) [3], as well as HMMs [12, 11, 4, 7, 8]. Certain algorithms aim at computational efficiency [12, 4], others aim at theoretical significance [3, 4, 7], while others aim at experimentaly-proved accuracy [11, 2, 8].

3. KWS Algorithms

The main issue in the design of KWS algorithm is the choice of a scoring function $S_{M,X}(Q,b,e)$. The scoring function has a definitive impact not only on the accuracy of the recognition, but also on the computational complexity. Often one will choose scoring functions that are approximations of some 'optimal' measures, but that are easier to compute than those measures [4].

3.1. Double Normalization (DN)

A double averaging of the probability of a match with first the number of frames per phone and then the number of phones in the keyword was known to yield good accuracy [1, 10, 7]. Assuming 1 state per phone, this is:

$$S_{M,X}(Q,b,e) \overset{\text{def}}{=} \frac{-1}{J}\sum_{j=1}^{J}\left(\frac{\sum_{n=b_j}^{e_j}\log P(q_j^n|x_n)}{e_j-b_j+1}\right) \tag{1}$$

where J represents the number of phones in M and q_j^n the hypothesized phone q_j for input frame x_n. The quality of a match increases with the decrease of this score.

The DN2 algorithm tested in this article differs from DN in [7] in the criteria for pruning the paths. If the HMM of each phoneme is considered a level in a word HMM, DN2 compares two paths for pruning as long as the paths end in states emitting the same phoneme in a frame, irrespective of the level. DN, on the other hand requires that the paths also end in the same level for them to be compared. Experiments detailed in [9] show DN2 to be slightly better than DN.

3.2. Extended Real Fitting (XRF)

We also define a new confidence measure that represents differently the exigencies of the recognition. Since the phonemes and the absent states can be modeled by the used HMMs offering sufficient flexibility, we wanted to verify the impact of requesting a good score of each matched phoneme. We measure the confidence level of a match as being equal with the maximum over all phonemes of the minus of the logarithm of the cumulated posterior of the phone, normalized with its length:

$$S_{M,X}(Q,b,e) \overset{\text{def}}{=} \max_{q_j\in\mathcal{Q}} \frac{\sum_{n=b_j}^{e_j}\log P(q_j^n|x_n)}{e_j-b_j+1} \tag{2}$$

In comparison to DN2, the XRF algorithm[1] evaluated in this article has the property that pruning does not retain just one path per phoneme in the trellis. Instead, it retains a set[2] of paths per phoneme satisfying some constraints independent of the other paths. The constraints are as follows:1) The length of every state in a candidate path is shorter than a maximum allowed value and 2) Average posterior of every phoneme is also within a threshold. The values of these thresholds is predetermined. The best paths satisfying these constraints per frame are retained.

4. Extended Evaluation Methodology

The standard evaluation of keyword spotting techniques is done by plotting ROC curves on the false positives (FP) and true positives/false negatives (FN) generated by the techniques on a set of randomly selected keywords. As it is highlighted by experiments described later, the performance of a technique is strongly influenced by the choice of keywords. Thus, to evaluate techniques on a set of randomly selected keywords is not fair. To overcome this issue, we now propose an extension of the standard evaluation methodology, called *Equal Opportunity Evaluation*.

Equal opportunity evaluation (EOE) selects the best keywords for each technique and then compares the performances of each technique on its respective list of best keywords thereby removing bias. Best keywords for an algorithm are picked by selecting keywords with lowest area under the ROC curve formed by FPs and FNs at different thresholds (see Figures 1 and 2). The area under the curves in Figures 1 and 2 evaluate to $(2*3)+(4*2)+(8*1)=26$ and $(4*3)+(6*2)+(10*1)=34$ respectively, indicating the ROC of the former word in Figure 1 is better, as required. 100 words with lowest area under the ROC are selected. We compared the techniques by plotting ROCs of these 100 words.

5. Experiments

The two algorithms, XRF and DN2, were evaluated on 242 sentences of the BREF database [5]. Based on our novel evaluation technique, we compare the two confidence measures - Double Normalization and Real Fitting for their best keywords. We picked the best prunings in confidence measures DN and RF, referred as DN2 and XRF, and compared their results. The ROCs are shown in Figures 3 and 4. As shown, although XRF starts at higher detection for lower FPs, DN2 goes on to achieve 100 percent detection before XRF. Depending on the threshold, one technique is better than the other. Also, if more number of paths are permitted per frame in XRF, its detection rate could surpass the detection rate of DN2. We also compared the performance of our techniques with Sliding Window (SW) Technique. DN2 and XRF exceeded the performance of SW as suggested by the results in Figure 5. We also compared the time taken by the three techniques and XRF took the least time followed by DN2. On

[1] Patent [6].

[2] Our experiments retained 3 paths per trellis

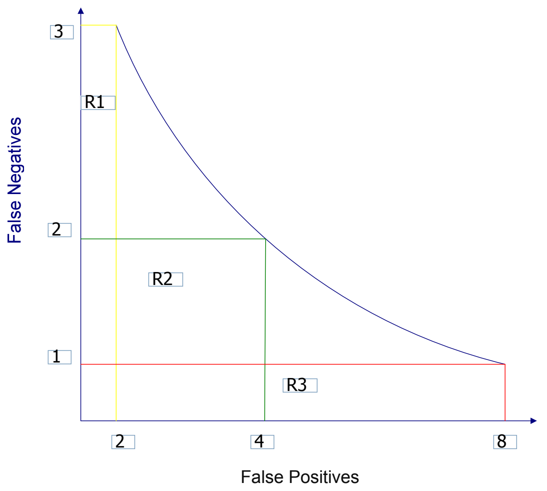

Figure 1: ROC curve of first word, area being computed by alternate method

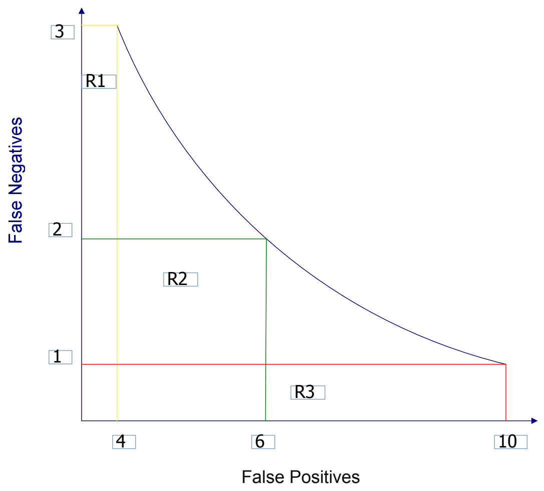

Figure 2: ROC curve of second word, area being computed by alternate method

an utterance of 740 frames, XRF (with 3 paths per state per frame) takes 5.22 seconds/keyword while DN2 takes 5.82 seconds/keyword. For the same utterance, SW takes 20 seconds/keyword! Clearly, not only do DN2 and XRF have better detection rates, they have a much lower time complexity as compared to Sliding Window technique.

5.1. Characteristics in Best Words

We also conducted some experiments to analyze the performance of individual phonemes. An interesting fact observed is that inspite of the fact that the words constituting the best 100 words for each algorithm were different, the ratio of recognized phonemes is very similar across algorithms. Some of the examples of well recognized phonemes in the table are /E/, /R/, /i/. Not only are some words more easily spotted by algorithms, some phonemes are also easier recognized by the used classifier (this may help combining classifiers).

Another such analysis was done on the length of the most easily recognized words. It is observed that all algorithms perform well on keywords with lengths 6 and 7 phones. Thus it can be said that keywords of length about 6-7 form good keywords. XRF is better at detecting shorter keywords than DN2 while DN2 is better at detecting longer keywords. A system comprising of both kinds of algorithms would thus perform well on a wider range of lengths of keywords. Also users could be advised to prefer keywords of the right length when searching in databases.

6. Conclusion

Keyword spotting (KWS) techniques deal with recognition of known vocabulary words in unconstrained utterances. This paper focuses on are evaluation criteria and confidence measures for KWS.

Two confidence measures are compared in this research: Real Fitting and Double Normalization. Real Fitting represents the score of a path by its worst phone match. XRF achieves 100 percent detection rate at 0.5 false positives/keyword/hour. For certain thresholds, it also outperforms algorithms based on Double Normalization. Results also show that our implementation based on Real Fitting was faster than the one based on Double Normalization. DN2 achieves 100 percent detection at 0.3 false alarms/keyword/hour in 5.82 seconds/keyword. XRF takes 5.22 seconds/keyword, outperforming the rest.

The results of a keyword spotting algorithm can be strongly influenced by the choice of keywords on which the experiments are run. This assumption is supported by the ROCs drawn to compare DN2 and XRF on two sets of keywords. Another contribution of this research is an evaluation technique that allows different algorithms to be compared without this influence. Equal opportunity evaluation selects the best keywords for each technique and compares the performance of the techniques on their respective best keywords. Although the best keywords for each technique differ, the phonemes constituting the best keywords across techniques are consistent, which leads as to conclude that some phonemes are easier to recognize by the used classifier. Most keywords which are recognized well by algorithms vary in length between 6-7

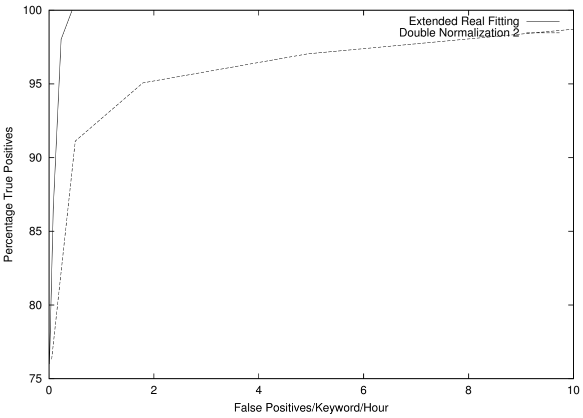

Figure 3: ROC curve comparing XRF and DN2 for the first set of 100 keywords

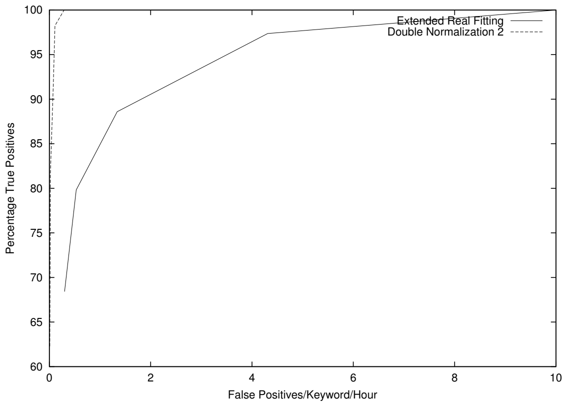

Figure 4: ROC curve comparing XRF and DN2 for the second set of 100 keywords

phones long. Another interesting fact to note is that RF techniques are better recognizing shorter keywords than DN while the latter recognize longer keywords more easily than the former. Hence a system that combines the two is expected to have a broad range of detection.

7. Acknowledgements

We thank Dr. Hervé Bourlard for his comments about an early version of this report and for providing us with the trained ANN used in these experiments.

8. References

[1] G. Bernardis and H. Bourlard. Improving posterior based confidence measures in hybrid hmm/ann speech recognition systems. In *ICSLP*, 1998.

[2] H. Bourlard, B. D'hoore, and J.-M. Boite. Optimizing recognition and rejection performance in wordspotting systems. In *ICASSP*, 1994.

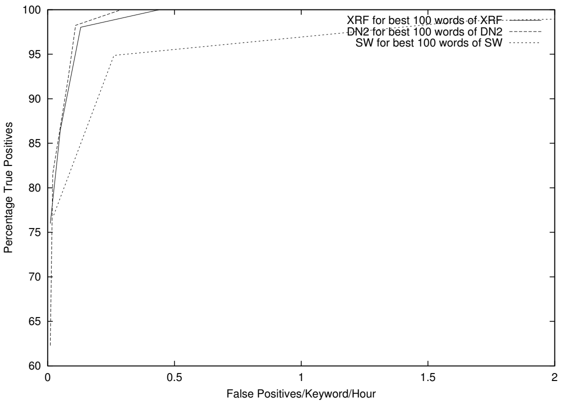

Figure 5: EOE comparing XRF, DN2, and SW's ROC curves for their best 100 keywords

[3] J. S. Bridle. An efficient elastic template method for detecting given words in running speech. In *Brit. Acoust. Soc. Meeting*, pages 1–4, 1973.

[4] J. Junkawitsch, L. Neubauer, H. Hoge, and G. Ruske. A new keyword spotting algorithm with pre-calculated optimal thresholds. In *ICSLP*, volume 4, pages 2067–2070, Philadelphia, PA, 1996.

[5] L. F. Lamel, J.-L. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french. In *EuroSpeech*, 1991.

[6] M.-C. Silaghi. Speech recogn. and sign. anal. by search of subsequences with maximal confidence measure. PCT patent WO 00/51107, Feb. 1999.

[7] M.-C. Silaghi and H. Bourlard. Iterative posterior-based keyword spotting without filler models. In *ICASSP*, Istanbul, Turkey, 2000.

[8] K. Thambiratnam and S. Sridharan. Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary kws. In *ICASSP*, 2005.

[9] R. Vargiya. Keyword spotting using confidence measures based on normalization of posterior probability. Technical Report CS-2005-9, FIT, 2005.

[10] G. Williams and S. Renals. Confidence measures for hybrid hmm/ann speech recognition. In *Proceedings of Eurospeech*, pages 1955–1958, 1997.

[11] J. G. Wilpon, C. H. Lee, and L. R. Rabiner. Application of hidden markov models for recognition ofa limited set of words in unconstrained speech. In *Proc. of ICASSP'89*, pages 254–257, 1989.

[12] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Trans. on ASSP*, 38(11):1870–1878, 1990.