Posterior Based Keyword Spotting with A Priori Thresholds

Hamed Ketabdar, Jithendra Vepa, Samy Bengio and Hervé Bourlard

IDIAP Research Institute, Martigny, Switzerland
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
haketa,vepa,bengio,bourlard@idiap.ch

Abstract

In this paper, we propose a new posterior based scoring approach for keyword and non keyword (garbage) elements. The estimation of these scores is based on HMM state posterior probability definition, taking into account long contextual information and the prior knowledge (e.g. keyword model topology). The state posteriors are then integrated into keyword and garbage posteriors for every frame. These posteriors are used to make a decision on detection of the keyword at each frame. The frame level decisions are then accumulated (in this case, by counting) to make a global decision on having the keyword in the utterance. In this way, the contribution of possible outliers are minimized, as opposed to the conventional Viterbi decoding approach which accumulates likelihoods. Experiments on keywords from the Conversational Telephone Speech (CTS) and Numbers'95 databases are reported. Results show that the new scoring approach leads to better trade off between true and false alarms compared to the Viterbi decoding approach, while also providing the possibility to precalculate keyword specific spotting thresholds related to the length of the keywords.

Index Terms: keyword spotting, keyword posterior, frame level decision, outliers, a priori thresholds.

1. Introduction

Word spotting is the detection of occurrences of selected words or phrases in speech. Hidden Markov Model (HMM) based approaches have been extensively used for this task [1, 2, 3, 4, 5]. The conventional way of spotting keywords using the HMM configuration is Viterbi decoding. Each path in the HMM contains a sequence of keyword and non keyword elements. Non keyword elements are modeled by the so called 'garbage' models. The decoder finds scores for all possible paths and the one with the highest score is selected as the output. This score is a global score accumulated over all likelihoods and transitions in the whole utterance, and not an specific keyword. Therefore, strong outliers can possibly contribute a lot in the final global score (thus, final decision made based on this score). Moreover, the score is not normalized with respect to the probability of the acoustic observation, thus it is relative to the particular acoustic observation [6]. It means that some factors like the length of the utterance, the length of keyword and garbage elements and the numerical range for the values of likelihoods, can affect this score. The values of these scores are penalized by changing keyword and garbage entrance penalties, which are effectively acting as spotting thresholds. The optimal choice of these thresholds are obtained by empirically adjusting the operating point (trade off between true and false alarms) to maximize the performance criteria on a development set.

Based on studies in [7, 8], in this paper we propose a new posterior based scoring approach for keyword and garbage elements. This posterior can be estimated through the same HMM configuration which is used in Viterbi decoding. The estimation of this posterior is based on HMM state posterior probability definition [9], taking into account prior knowledge (e.g. keyword model topology) and long contextual information. The state posterior probabilities are then integrated to keyword and garbage posteriors for each frame. This is a frame level score for a keyword or garbage element and not a global score for the whole utterance. Moreover, the estimation of these posteriors involves normalization with respect to the probability of acoustic observation, therefor it is irrelative to a particular acoustic observation space. These frame level posteriors are then used to make a frame level decision about the detection of the keyword. These frame level (binary) decisions are then accumulated (in this case by counting) to have a global decision about the detection of the keyword in the utterance. Therefor, the main difference between our approach and the Viterbi decoding approach is accumulating frame level decisions instead of frame level likelihoods. This leads to decreasing the contribution of the possible outliers, because even strong temporal outliers can only change few frame level decisions, while they can significantly change the accumulated likelihoods.

We show that the new posterior based scoring approach results in a better trade-off between true and false alarms (larger area under the ROC curve), compared to the Viterbi based approach. Moreover, it provides the possibility to precalculate keyword specific spotting thresholds based only on the keywords length, which can be known a priori, or computed from the minimum length and number of phonemes composing the keyword. In contrast, in the Viterbi based approach, there is no meaningful interpretation of thresholds (entrance penalties) in terms of a priori known keyword characteristics, and they should be adjusted empirically.

The paper is organized as follows: Section 2 explains the garbage and keyword modeling approach used in this work. Section 3 reviews the Viterbi based scoring approach and introduces the keyword and garbage posterior based scoring approach. Section 4 talks about keyword detection based on frame level keyword posteriors, and threshold precalculation. Section 5 explains the experiments comparing the two scoring approaches. Finally, Section 6 summarizes the paper.

2. Modelling garbage and keywords

We have used acoustic sub-word speech units (phonemes) as garbage models [1, 12], thus the garbage is represented as a sequence of separate phonemes. Keywords are also modeled by concatenating phoneme models which are composing the keyword. Therefore, the whole HMM configuration is a parallel network of keyword models (composed of phone models) and separate phone models (garbage models).

3. Keyword and garbage scoring

3.1. Viterbi based scoring

The conventional approach to detect keywords is Viterbi decoding through the HMM configuration [1, 2, 4, 12]. Each path in the decoder is a sequence of keyword and garbage elements. The decoder finds scores for all possible paths and the one with the highest score is selected as the output. This score is related to the joint probability of the path and the feature vectors (evidences). This scoring approach has the the following drawbacks concerning the keyword spotting task:

- The score is a global score estimated by accumulating all likelihoods for the whole utterance, and not specifically for a keyword or garbage element. Therefore, the temporal outliers can possibly affect the final global score significantly, and result in having a wrong spotting case.

- The score is not normalized with respect to the probability of the acoustic observation and thus relative to the particular acoustic observation space [6]. For example, it can be related to the length of the utterance, the length and number of keywords and garbage elements, the numerical range for values of evidences, etc.

- The values of these scores are penalized by changing keyword and garbage entrance penalties, which are effectively spotting thresholds in this approach. There is no meaningful interpretation for the entrance penalty values and they should be adjusted empirically to optimize the performance criteria. It implies that for each keyword there should be a sufficiently large development or training set. It would be ideal if we could find a reasonable threshold based on keyword characteristics like length which can be known a priori or easily estimated or measured, instead of adjusting on a development set.

3.2. Posterior based scoring

Based on the previous work in [7, 8], we propose a new frame level posterior probability score for keyword and garbage elements. This posterior probability can be estimated through the same HMM configuration which is used for the Viterbi decoding. The estimation of these posteriors are based on HMM state posterior probability definition, integrating long contextual information and also prior knowledge (such as keyword structure and model topology). The HMM state posterior probability $p(q_t^i|x_{1:T}, M)$ is the probability of being in specific HMM state q^i at specific time t having seen the whole observation sequence $x_{1:T}$ and the model M encoding prior knowledge (e.g. keyword structure and model topology) [9]. It can be written in terms of HMM forward and backward recursions as follows:

$$p(q_t^i|x_{1:T}, M) = \frac{\alpha(i,t)\beta(i,t)}{\sum_j \alpha(j,T)} \tag{1}$$

$$\begin{aligned} \alpha(i,t) &= p(x_{1:t}, q_t^i) \\ &= p(x_t|q_t^i) \sum_j p(q_t^i|q_{t-1}^j)\alpha(j,t-1) \end{aligned} \tag{2}$$

$$\begin{aligned} \beta(i,t) &= p(x_{t+1:T}|q_t^i) \\ &= \sum_j p(x_{t+1}|q_{t+1}^j)p(q_{t+1}^j|q_t^i)\beta(j,t+1) \end{aligned} \tag{3}$$

where, x_t is a feature vector at time t, $x_{1:T} = \{x_1, \ldots, x_T\}$ is an acoustic observation sequence, q_t is HMM state at time t, which value can range from 1 to N_q (total number of possible HMM states), and q_t^i shows the event "$q_t = i$". In the following, we will drop the M, keeping in mind that all recursions are processed through some prior (Markov) model M. Similar recursions can be written for posterior based systems (such as hybrid HMM/ANN system) where the HMM state emission probabilities are estimated by Neural Networks [13].

The state level posterior probabilities are then integrated to frame level keyword and garbage posteriors:

$$\begin{aligned} p(w_t^i|x_{1:T}) &= \sum_{j=1}^{N_q} p(w_t^i, q_t^j|x_{1:T}) \\ &= \sum_{j=1}^{N_q} p(w_t^i|q_t^j, x_{1:T})p(q_t^j|x_{1:T}) \end{aligned} \tag{4}$$

where w_t is a keyword at time t and w_t^i represents the event "$w_t = i$". $p(w_t^i|q_t^j, x_{1:T})$ represents the probability of being in a given keyword i at time t knowing to be in the state j at time t. Assuming that there is no parameter sharing between keywords and garbage elements (which is the case in this work), it is deterministic and equal to 1 or 0. Hence, a keyword frame level posterior is estimated by adding up all the posteriors for the states associated with the keyword in the whole model. The same argument is valid for the garbage elements posterior estimation.

Comparing with the Viterbi decoding approach, the new scoring approach provides the following advantages:

- It provides a frame level keyword or garbage specific score, instead of a global score for the whole utterance. As (1-3) show, it is not possible to get a high posterior for a keyword without having a high emission probability (evidence) for it, while the score in the decoder based approach is global and can be affected by many factors.
- This score is normalized with respect to the probability of acoustic observation (1), and thus irrelative to the particular observation sequence.
- Having frame level normalized scores allows the possibility of relating the spotting thresholds to the length of the keywords (explained in more details in the next section).

Next section explains how these frame level posteriors are used to decide about detection of a keyword in the utterance.

4. Keyword detection and threshold precalculation

Having the frame level keyword or garbage posteriors $p(w_t^i)$, the next step is to decide about existence of the keyword in the utterance. The frame level posteriors are used to make a frame level decision about the detection of the keyword (by comparing frame level keyword and garbage posteriors). The frame level (binary) decisions are then accumulated (in this case by counting continuous frame level keyword detections). The outcome is showing the detected length of the keyword in the utterance. The main difference between our approach and the Viterbi decoding approach is accumulating frame level decisions instead of frame level likelihoods. Strong temporal outliers can contribute significantly in the Viterbi based scores leading to a wrong spotting case, while they can only affect few frame level decisions in our case.

As mentioned, the above process provides a score showing the detected length of the keyword in the utterance. Therefore, the spotting threshold to compare with this length based score, can be precalculated based on the length of the keywords. The length of the keywords can be known a priori or computed using the number and minimum duration of phonemes composing the keyword. These thresholds can be further adjusted having in mind that they are related to the length, in order to achieve different desired operating points. In a practical keyword spotting system, specially if the keyword set is not fixed, or we are interested to spot names or words which are not appearing very frequently in the database, or in applications like learning to read tutors, we cannot have a huge development set for each new keyword and new condition to properly adjust the spotting thresholds. In these cases, precalculating keyword specific thresholds based on some priorly known characteristics of the keywords (e.g. length) can be useful.

5. Experiments and results

For the experiments, we model garbage and keyword elements with monophone units as explained in Section 2. We mainly compare the Viterbi scoring approach with the new posterior based scoring approach for spotting keywords.

We used Conversational Telephone Speech (CTS) [10] and Numbers'95 [11] databases for the experiments. There are 1000 and 31 words, and 46 and 27 phones in these databases, respectively. The acoustic feature vectors are PLP cepstral coefficients and their first and second order derivatives. The HMM emission probabilities are phone posteriors estimated by a Multi Layer Perceptron (MLP). We used 15 hours of data to train the MLP in the CTS case and 3 hours in the case of Numbers'95 database. The test set contains 2 hours of data for CTS database and 2 hours for Numbers'95 database.

We have used 7 keywords from the CTS database and 5 keywords from Numbers'95 database. These keywords are 'you', 'yeah', 'like', 'think', 'something', 'because', 'people', 'one', 'five', 'four', 'fifteen', and 'zero'. Their selection is based on having a large variability in terms of frequency, number of phonemes and length.

In the first set of experiments, the performance of our posterior based scoring system is compared with the Viterbi decoder based system in terms of trade-off between true and false alarms. The HMM configuration is the same for the two methods. We use Receiver Operating Characteristic (ROC) curves in order to measure and compare the performance of the two systems. Figure 1 shows ROC curves for different keywords obtained by the two methods. In most of the cases, the area under the curve is higher for the posterior based approach, showing that it can achieve better trade-off between true and false alarms. In the Viterbi based approach, the score which is used to decide about detecting a keyword is a global score obtained for the whole utterance, and accumulated over all evidences for garbage and keywords, transition probabilities, etc. Therefore, even when there is no keyword in the utterance, a 'fake' existence of a keyword can be possibly made by a strong temporal outlier (having very large or very small likelihood) which can change global scores for the paths. In contrast, in the posterior based approach, a temporal outlier, no matter how strong it is, can only affect possibly few frame level decisions, thus less probable to lead in a wrong spotting case.

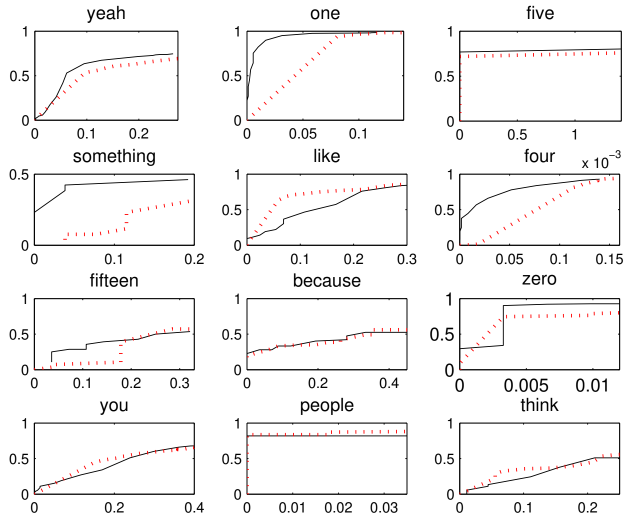

Figure 1: *ROC curves for different keywords. The dotted curves are showing Viterbi based approach results and full curves are showing posterior based approach. The y axis is the percentage of true alarms and the x axis is the percentage of false alarms.*

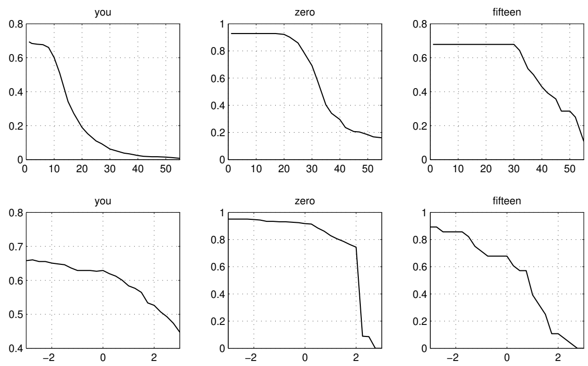

Figure 2: *Relation between the spotting rates and the thresholds for the two methods. The first row is showing posterior based approach and the second row shows decoder based approach. The y axis shows the spotting rates and the x axis shows the thresholds.*

In the second group of experiments, we study the relation between the spotting rates and the thresholds for the two approaches, and the possibility for precalculating keyword specific thresholds in the posterior based system. Figure 2 shows this relation, obtained for keywords with different lengths. The threshold for the posterior based system is the period for having continuous frame level keyword detection (in frames), while the threshold for the decoder based approach is the entrance penalty values. As can be seen, the threshold for the posterior based system is a meaningful value related to the length of the keyword (long words need higher threshold while shorter words need less) while it is not easy to find a meaningful interpretation of thresholds for the other system. Table 1 shows the performance of the posterior based system obtained with precalculated thresholds for different words[1 2]. The last column in the table shows the maximum achievable spotting rate with the posterior based approach (to have an idea how well the precalculated threshold works). We set the thresholds to the minimum length of the keywords. The minimum length of the keywords are assumed to be equal to the sum of the minimum lengths of its phonemes (3 frames per phoneme in this case). The precalculated thresholds can be adjusted further based on the desired trade-offs, taking into account that they are related to the length of keywords. In contrast, since the score in the decoder based approach can be related to different factors (as mentioned in Section 3.1), the spotting threshold is also a complex function of different factors. Therefore, the threshold precalculation cannot be applied in this case and it is necessary to have a development set for any new keyword to adjust the thresholds.

Table 1: True and false alarm rates for different keywords with the spotting thresholds set to the minimum keywords length. Length values are in frames.

Keyword	Min length (threshold)	True and false alarms (%)	Max true alarms (%)
one	9	98.0 - 9.5	98.3
four	9	92.7 - 13.7	93.0
five	9	82.7 - 0.16	84.0
zero	12	94.0 - 1.5	94.5
fifteen	21	67.3 - 33.1	67.5
you	6	65.5 - 40.0	68.5
yeah	6	72.0 - 25.0	74.0
like	9	84.3 - 30.0	84.8
think	12	51.1 - 25.6	53.3
people	15	81.8 - 0.0	81.8
because	15	47.3 - 28.1	52.6
something	18	61.5 - 96.1	65.4

6. Conclusions

In this paper, we proposed estimating a new frame level posterior based score for keyword and garbage elements. We showed how this posterior can be estimated based on HMM state posterior probability definition, taking into account long contextual information and prior knowledge (e.g. keyword model topology). The frame level keyword and garbage posteriors are then used to make a frame level decision about detecting the keyword. These frame level decisions are accumulated to a global decision for having the keyword in the utterance, by counting the number of frame level keyword detections. Comparing with the Viterbi decoding approach which makes a global decision by accumulating frame level likelihoods, here we make a global decision based on frame level decisions. In this way, an outlier can just affect few frame level decisions while in the conventional Viterbi based approach, it can affect the whole global score. We showed that the new posterior based scoring approach results in a better trade-off between true and false alarms. In addition, we also studied the relation between spotting rates and the thresholds for the posterior based and Viterbi based approaches, and showed that the posterior based approach provides the possibility to precalculate keyword specific spotting thresholds based on the length of the keywords.

[1] In order to have a rough idea about the difficulty of these tasks (CTS and Numbers'95) it is useful to mention that the state-of-the art speech recognition performance for CTS and Numbers'95 databases are about 50 and 95 percent recognition rate, respectively.

[2] True and false alarm percentages for each keyword are obtained by dividing the number of true and false alarms by the total occurrences of the that keyword in the test set.

7. Acknowledgments

This project was funded by the European AMI project, and the IM2 Swiss National Center of Competence in Research. The authors would also like to thank Hynek Hermansky for helpful discussions.

8. References

[1] Rose, R., and Paul, D., "A Hidden Markov Model Based Keyword Recognition System", 1990 IEEE ICASSP, pp. 129- 132.

[2] Wilpon, J.G., Rabiner, L.R., Lee, C.H., and Goldman, E.R., "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", 1990 IEEE Trans. ASSP, Vo138. No. 11, pp. 1870-1878.

[3] Wilpon, LG., Miller, L.G., and Modi, P., "Improvements and Applications for Key Word Recognition Using Hidden Markov Modeling Techniques", 1991 ICASSP, pp. 309-312.

[4] Rohlicek, R., Russell, W., Roukos, S., Gish, H., "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting", 1989 IEEE ICASSP, pp. 627-630.

[5] Wilcox, L. D., and Bush, M.A., "Training and Search Algorithms for an Interactive Word Spotting System", 1992 IEEE ICASSP, pp. H-97-II-100.

[6] Williams, G. and Renals, S., "Confidence measures for hybrid HMM/ANN speech recognition" Proceedings of Eurospeech'97, pp. 1955-1958, 1997.

[7] Bourlard, H., Bengio, S., Magimai Doss, M., Zhu, Q., Mesot, B., and Morgan, N., "Towards using hierarchical posteriors for flexible automatic speech recognition systems", *DARPA RT-04 Workshop*, November 2004, also IDIAP-RR 04-58.

[8] Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Developing and enhancing posterior based speech recognition systems", *Interspeech'05*, Lisbon, Portugal, 2005.

[9] Rabiner, L. R., "A tutorial on hidden Markov models and selective applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-286, 1989.

[10] J. Godfrey, E. Holliman, and J. McDaniel. "SWITCHBOARD: Telephone speech corpus for research and development" In Proc. ICASSP-92, pages 517–520.

[11] Cole, R., Fanty, M., Noel, M. and Lander T. "Telephone Speech Corpus Development at CSLU", In Proc. of ISCLP (Yokohama, Japan, 1994), pp. 1815-1818.

[12] H. Bourlard, B. D'hoore, and J.M. Boite, "Optimizing recognition and rejection performance in word spotting systems", In Proc. of ICASSP, volume 1, pages 373–376, 1994.

[13] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Publishers, 1994.