CONFIDENCE MEASURES FOR KEYWORD SPOTTING USING SUPORT VECTOR MACHINES

Y. Benayed, D. Fohr and J. P. Haton

LORIA-CNRS/ INRIA Lorraine
BP239, F54506, Vandœuvre France
(ybenayed, fohr, jph)@loria.fr

G. Chollet

ENST, CNRS-LTCI
46 rue Barrault, F75634 Paris France
chollet@tsi.enst.fr

ABSTRACT

Support Vector machines (SVM) is a new and very promising classification technique developed from the theory of Structural Risk Minimisation [1]. In this paper, we propose an alternative out-of-vocabulary word detection method relying on confidence measures and support vector machines. Confidence measures are computed from phone level information provided by a Hidden Markov Model (HMM) based speech recognizer. We use three kinds of average techniques as arithmetic, geometric and harmonic averages to compute a confidence measure for each word. The acceptance/rejection decision of a word is based on the confidence feature vector which is processed by a SVM classifier. The performance of the proposed SVM classifier is compared with methods based on the averaging of confidence measures.

1. INTRODUCTION

During recent years, it has become increasingly essential to equip speech recognition systems with the ability to accommodate spontaneous speech input. Although providing this capability facilitates a friendly user-interface, it also poses a number of new problems, such as the inclusion of out-of-vocabulary words, false starts, disfluency, and acoustical mismatch.

Significant progress has been made in keyword spotting for unconstrained speech using HMM. Keyword spotting systems introduce a filler model for enhancing keyword detection and absorbing out-of-vocabulary events. To reduce false alarm, a large number of studies have incorporated *keyword verification* following detection and segmentation of speech into keyword hypothesis via a conventional Viterbi search. These studies employ some types of confidence measure to verify whether or not a given keyword exists within a segment of speech [2, 3].

In this work, three kinds of average techniques were investigated as arithmetic,geometric and harmonic averages. Firstly, we use phone confidences without weights. Secondly, we weight all phone confidence measures by normalizing them with phone duration. These averages are used to extract a phone-based confidence measure for rejection. It is important to note that this type of keyword verification introduces two new kinds of errors : *false rejection* and *false acceptance*. False rejection refers to the rejection of a valid keyword, and false acceptance to the acceptance of an incorrect keyword.

In this paper, a support vector machine based method is proposed for keyword spotting. The feature vectors for SVM classifier are constructed with the acoustic confidence measures. The SVM minimizes the structural risk, i.e., the probability of misclassifying patterns for fixed but unknown probability distribution of the data. This is in contrast to traditional pattern recognition techniques of minimizing the empirical risk, i.e., of optimising the performance on the training data. This minimum structural risk principle is equivalent to minimizing an upper bound on the generalisation error [4, 5]. The proposed approach is evaluated and compared against the word level confidence measure methods.

The remainder of this paper is organized as follows : section 2 describes the recognition system, and the confidence measures are given in section 3. In section 4, the basic principles of the SVM are briefly described. Speech database and experimental results are presented in section 5. Finally, section 6 concludes this presentation.

2. RECOGNITION SYSTEM

The recognizer used in this work is a speaker independent HMM system. The modelled unit is a phone, and each phone is represent by 3-state, strictly left-to-right, continuous density HMM. The topology of the hidden Markov models is defined by the number of states and the allowed transitions. A word is represented by the concatenation of phone models. The number of probability density function (pdf) per state is determined during the training phase.

The parameterization is based on MFCC (Mel-Frequency Cepstral Coefficients) parameters. The user can modify this parameterization : size of the analyzing window, shift, number of triangular filters, lower and upper frequency cutoff of the filter bank, and number of the cepstral coefficients. Finally the delta (the first derivative) and acceleration coefficients (the second derivative) can be added. In the following experiments, the acoustic feature vectors are built as follows: 32ms frames with a frame shift of 10ms, each frame is passed through a set of 24 triangular band-pass filter resulting in a vector of 35 features, namely 11 static mel-cepstral coefficients (C_0 is removed), 12 delta and 12 delta delta coefficients. Phonemes and silence are modeled by continuous density mixtures. Models are left-to-right with no skip state transitions. 31 context independent phoneme models and a silence model are used.

It is important to notice that keyword HMMs are obtained as a concatenation of phone HMMs, so no special training data are needed to model keywords. In the recognition phase, parameters are adjusted in order to have no deletion keywords (as consequence we obtain a large number of insertion keywords).

3. CONFIDENCE MEASURE

The confidence measure is useful for rejecting utterances that are out of domain, or that contain out-of-vocabulary words or speech disfluences. Phone confidence is computed for each frame of speech as the posterior phone probability given the acoustic observation.

There are many different ways to compute confidence measure of each word by combining phone-level confidence measures [6]. In this work, we use three different averaging methods : arithmetic mean , geometric mean and harmonic mean. In a first attempt, we use phone confidence. Then we have confidence measure : arithmetic (CMa), harmonic (CMh) and geometric (CMg) respectively :

$$CMa(w) = \frac{1}{N}[\sum_{i=1}^{N} CM_i], \quad CMh(w) = \frac{N}{\sum_{i=1}^{N} \frac{1}{CM_i}}$$

$$CMg(w) = exp(\frac{1}{N}[\sum_{i=1}^{N} log(CM_i)])$$

Where :

N is the total number of phone sequence of the word.

$$CM_i = P(PH_i|O_t) = \frac{P(O_t|PH_i)P(PH_i)}{\sum_j P(O_t|PH_j)P(PH_j)}$$

The posterior probability $P(O_t|PH_i)$, is computed using a Viterbi algorithm.

$w = \{PH_1, ...PH_N\}$: sequence of phones for a spoken utterance.
$O = \{O_1, ...O_T\}$: acoustic observation sequence.
$O_i = \{O_{b[i]}, ..., O_{e[i]}\}$: sequence of frames, where b[i] and e[i] represent respectively the beginning and the end of frames of the phone number i.

$$T_i = e[i] - b[i] + 1$$

In order to incorporate durational information in the confidence measure, we propose to weight all phones equally by normalizing them with phone duration. The word level confidence measures based on these means from duration normalized phone level confidence measures are :

$$CMa_n(w) = \frac{1}{N}[\sum_{i=1}^{N} \frac{CM_i}{T_i}], \quad CMh(w) = \frac{N}{\sum_{i=1}^{N} \frac{T_i}{CM_i}}$$

$$CMg_n(w) = exp(\frac{1}{N}[\sum_{i=1}^{N} log(\frac{CM_i}{T_i})])$$

The confidence scores computed as above are used to take the final decision of accepting or rejecting an hypothesis. Each average is then postprocessed with a sigmoid mapping function, which is a well known thresholding means.

$$CM_f(w) = f(CM(w)) = \frac{1}{1 + exp(-\alpha \;\; CM(w))}$$

For each confidence measure, a specific threshold γ is set up. If the confidence score is below this threshold, the keyword is rejected :

$$w = \begin{cases} Accept & \text{if } CM_f(w) > \gamma \\ Reject & \text{otherwise} \end{cases}$$

4. LEARNING USING SUPPORT VECTOR MACHINES

4.1. Linear support vector machines

Consider the problem of separating the set of m training vectors belonging to two different classes, $\{(x_1, \;\; y_1), . . .,$ $(x_m, y_m)\}$ where $x_i \in R^n$ is a feature vector and $y_i \in \{-1, 1\}$ a class label, with a hyperplane of equation $w.x + b = 0$. Of all the boundaries determined by w and b, the one that maximizes the margin would generalise better, as compared to other possible separating hyperplanes.

A separating hyperplane in a canonical form must satisfy the following conditions :

$$y_i(w.x_i + b) \geq 1 \quad \forall i \in \{1, ..., m\}$$

The optimal separating hyperplane is given by maximizing the margin M given by the equation : $M = \frac{2}{||w||}$. Hence, the hyperplane that optimally separates the data is the one that minimizes : $\Phi(w) = \frac{w^2}{2}$. For the solution of the optimisation problem, the reader can refer to [5].

4.2. The non-linear separable case

In this case, the set of training vectors of two classes are non-linearly separable. To solve this problem, Cortes and Vapnik [5] introduce non-negative variables, $\xi_i \geq 0$, which measure the miss-classification errors. The optimisation problem is now treated as a minimization of the classification error [7]. The separating hyperplane must satisfy the following inequalities :

$$(w.x_i) + b \geq +1 - \xi_i, \quad if \quad y_i = +1$$

$$(w.x_i) + b \leq -1 + \xi_i, \quad if \quad y_i = -1$$

The generalised optimal separating hyperplane is determined by the vector w, that minimizes the functional :

$$\phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{m} \xi_i$$

Where $\xi = (\xi_1, ..., \xi_m)$ and C are constants. The reader can refer to [4] for more details on the non-linear separable case.

4.3. Kernel support vector machines

If a linear boundary is inappropriate, the SVM replaces the inner product $x_i.x_j$ by a kernel function $K(x_i.x_j)$, and then constructs an optimal hyperplane in the mapped space. According to Mercer theorem [5], the kernel function implicitly maps the input vectors, via a Φ associated with the kernel, into a high dimensional feature space in which the mapped data is linearly separable. Kernel functions play a very important role in avoiding explicit production of the mappings and the curse of dimensionality.
There are several possible kernel functions :
- Linear : $K(x, y) = x.y$.
- Polynomial : $K(x, y) = (x.y + 1)^d$, where d is the degree of the polynomial.
- Radial Basis Function (RBF) : $K(x, y) = exp[-\frac{|x-y|^2}{2\sigma^2}]$, where σ is the width of the Gaussian function.

For a given kernel function, the classifier is given by :

$$class(x) = Sign[\sum_{SV} \alpha_i^0 y_i K(x_i.x) + b^0]$$

5. EXPERIMENTAL RESULTS

All experiments are carried out with the French database SPEECHDAT, recorded through the telephone network, at 8kHz. The database used for training contains 8800 sentences pronoun-ced by 800 speakers. For the test we use a set of 4780 sentences pronounced by 1000 speakers (different from the speakers of the training database). These test data contain 3,180 utterances of 20 keywords and 25,700 out-of-vocabulary words.

In the first set of experiments, we made a comparative study on the three different kinds of average methods : arithmetic, geometric, and harmonic means. First, we use phone confidence, next we use duration normalized phone confidence.

In the second set of experiments, the proposed SVM approach was compared against methods based on confidence measures. In effect, we have noticed that our three kinds of confidence measures have given complementary results, so we propose to combine them in only one classification method. SVM is known to be a good and a promising classifier, that's why we decide to use it with a linear and a RBF kernel. The input feature vector of the SVM classifier was then composed of all confidence measures used in the first set of experiments.

To evaluate the performance of our spotting system, we use two evaluation rates :

- The False Acceptance Rate, also called False Alarm Rate (FAR), defined as :

$$FAR = \frac{Total \;\; False \;\; Acceptance}{Total \;\; False \;\; Attempts}$$

- The False Rejection Rate (FRR), defined as :

$$FRR = \frac{Total \;\; False \;\; Rejection}{Total \;\; True \;\; Attempts}$$

Plotting FRR versus FAR gives a Receiver Operating Characteristics (ROC) graph.

The resulting ROC curves, using three kinds of means, arithmetic, geometric and harmonic by varying the value of the threshold γ are presented in figure 1. Best performance is obtained by arithmetic mean.
The Equal Error Rate (EER), given by FAR=FRR, is about 32.7% with a confidence interval of (+/- 0.6%) obtained by the arithmetic mean.

Figure 2 presents ROC curves corresponding to the performance obtained using normalized means for the three kinds : arithmetic, geometric and harmonic, by varying the value of the threshold γ. The best results are achieved by

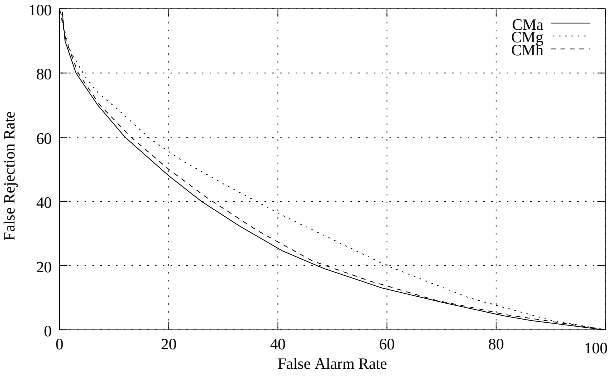

Fig. 1. ROC curves using three kinds of means : arithmetic, geometric and harmonic by varying the value of γ.

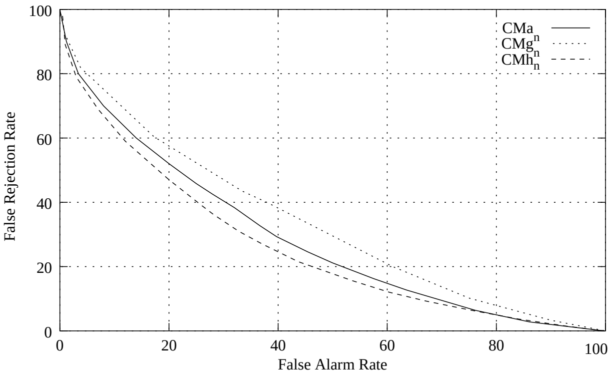

Fig. 2. ROC curves using three kinds of normalized means: arithmetic, geometric and harmonic by varying the value of γ.

the normalized harmonic mean. The EER is about 31.7% with a confidence interval of (+/- 0.6%).

Figure 3 shows that the results obtained by the RBF kernel and the linear SVM are better than those obtained by the best of means, which is the normalized harmonic mean. The EER concerning the RBF kernel is about 26.7% (+/- 0.5%) compared to 28.1% (+/- 0.5%) obtained by the linear kernel and 31.7% achieved by the normalized harmonic mean.

6. CONCLUSION

In this paper, we have described a keyword spotting system based on phone models. This system consists of two phases : recognition and verification. In the stage of recognition, multiple hypotheses with hypothesized word boundaries are obtained trough Viterbi decoder. In the stage of verification, unlikely hypotheses are rejected using a confidence measure. Firstly, three kinds of single and normalized average techniques were investigated as arithmetic, geometric and harmonic. Secondly, an alternative approach based on support vector machines was proposed for keyword spotting. The performances achieved were compared against word level confidence measure methods. The results show that the SVM approach provides the best result.

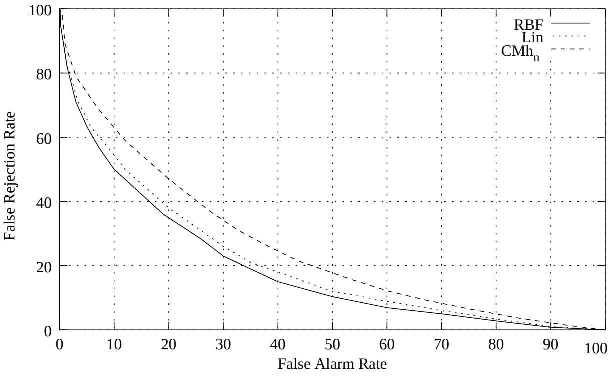

Fig. 3. ROC curves comparing the performances of linear, RBF SVM and normalized harmonic mean.

7. REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[2] Y. BenAyed, D. Fohr, J. P. Haton, and G. Chollet, “Keyword spotting using support vector machines,” in *5th International Conference on Text, Speech and Dialogue, Brno, Czech Republic*, 2002.

[3] Y. BenAyed, D. Fohr, J. P. Haton, and G. Chollet, “Recognition and rejection performance in wordspotting systems using support vector machines,” in *2nd WSEAS International Conference on Signal, Speech and Image, Skiathos Island, Greece*, 2002.

[4] C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, 1998.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, 1995.

[6] M. W. Koo, “An utterance verification system based on subword modeling for a vocabulary independent speech recognition system,” in *6th European Conference on Speech Communication and Technology, Budapest, Hungary*, 1999.

[7] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.