Speech Recognition using SVMs

Nathan Smith
Cambridge University
Engineering Dept
Cambridge, CB2 1PZ, U.K.
nds1002@eng.cam.ac.uk

Mark Gales
Cambridge University
Engineering Dept
Cambridge, CB2 1PZ, U.K.
mjfg@eng.cam.ac.uk

Abstract

An important issue in applying SVMs to speech recognition is the ability to classify variable length sequences. This paper presents extensions to a standard scheme for handling this variable length data, the Fisher score. A more discriminatory mapping is introduced based on the likelihood-ratio. The score-space defined by this mapping avoids some of the problems with the Fisher score. It also allows the discriminative power of the generative models to be directly incorporated into the score-space. The mapping schemes, and appropriate normalisation schemes, are evaluated on a speaker-independent isolated letter task where the new mapping outperforms both the Fisher score and standard HMMs.

1 Introduction

Speech recognition is a complex, dynamic classification task. State-of-the-art systems use Hidden Markov Models (HMMs), either trained discriminatively or to maximise likelihood, to achieve good levels of performance. One of the reasons for the popularity of HMMs is that they readily handle the variable length data sequences which result from variations in word sequence, speaker rate and accent. Support Vector Machines (SVMs) [1] are a powerful discriminative technique that have been shown to work well on a variety of tasks. However they are typically only applied to static binary classification tasks. This paper examines the application of SVMs to speech recognition. There are two major problems to address. First, how to handle the variable length sequences. Second, how to handle multi-class decisions. This paper only concentrates on dealing with variable length data sequences. It develops our earlier research in [2].

There have been a variety of methods proposed to map variable length sequences to vectors of fixed dimension. These include vector averaging and selecting a 'representative' number of observations from each utterance [3]. However, these methods may discard useful information. This paper adopts an approach similar to that of [4] which makes use of all the available data. The scheme uses generative probability models of the data to define a mapping into a fixed dimension space, the Fisher score-space. When incorporated within an SVM kernel, the kernel is known as the Fisher kernel. It may be shown that the Fisher kernel has justifiable data-dependency and regularisation [5]. This paper examines the suitability of the Fisher kernel for classification in speech recognition and proposes an alternative, more discriminative kernel. In addition some normalisation issues associated with using this kernel for speech recognition are addressed.

Initially a general framework for defining alternative score-spaces is required. First, define an observation sequence as $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots \boldsymbol{o}_t, \ldots \boldsymbol{o}_T)$ where $\boldsymbol{o}_i \in \mathbb{R}^D$, and a set of generative probability models of the observation sequences as $\mathcal{P} = \{p_k(\boldsymbol{O}|\boldsymbol{\theta}_k)\}$, where $\boldsymbol{\theta}_k$ is the vector of parameters for the kth member of the set. The observation sequence $\boldsymbol{O}$ can be mapped into a vector of fixed dimension [4],

$$\varphi^f_{\hat{F}}(\boldsymbol{O}) \quad = \quad \varphi_{\hat{F}} f\big(\{p_k(\boldsymbol{O}|\boldsymbol{\theta}_k)\}\big) \tag{1}$$

$f\big(\{p_k(\boldsymbol{O}|\boldsymbol{\theta}_k)\}\big)$ is the *score-argument* and is a function of the members of the set of generative models $\mathcal{P}$. $\varphi_{\hat{F}}$ is the *score-mapping* and is defined using a *score-operator* $\hat{F}$. $\varphi^f_{\hat{F}}(\boldsymbol{O})$ is the *score* and occupies the fixed dimension *score-space*. Our investigation of score-spaces falls into three divisions. What are the best generative models, score-arguments and score-operators to use?

2 The score-space

As HMMs have proved successful in speech recognition, they are a natural choice as the generative models for the proposed kernels. In particular HMMs with state output distributions formed by Gaussian Mixture Models. There is also the choice of the score-argument. Previous schemes have used the log of a single generative model, $p(\boldsymbol{O}|\boldsymbol{\theta})$, representing either both classes [4] or one of the classes [6]. This score-space is termed the *likelihood* score-space, $\varphi^{\text{lik}}_{\hat{F}}(\boldsymbol{O})$. It was used to define the original Fisher score-space. The score-space proposed in this work uses the log of the ratio of two class-conditional models $p_1(\boldsymbol{O}|\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{O}|\boldsymbol{\theta}_2)$ where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T]^T$. The corresponding score-space is called the *likelihood-ratio* score-space, $\varphi^{\text{lr}}_{\hat{F}}(\boldsymbol{O})$. Thus,

$$\varphi^{\text{lik}}_{\hat{F}}(\boldsymbol{O}) \quad = \quad \varphi_{\hat{F}} \ln p(\boldsymbol{O}|\boldsymbol{\theta}) \tag{2}$$

$$\varphi^{\text{lr}}_{\hat{F}}(\boldsymbol{O}) \quad = \quad \varphi_{\hat{F}} \ln \left[\frac{p_1(\boldsymbol{O}|\boldsymbol{\theta}_1)}{p_2(\boldsymbol{O}|\boldsymbol{\theta}_2)}\right] \tag{3}$$

The likelihood-ratio score space avoids some of the limitations of the likelihood score-space, and may be viewed as a generalisation of the standard generative model classifier. These issues will be discussed later.

Having proposed these generative models and score-arguments, the score-operators must be selected. The original score-operator in [4] was the 1st-order derivative operator applied to HMMs with discrete output distributions. Consider a continuous density HMM with N emitting states, $j \in \{1 \ldots N\}$. Each state, j, has an output distribution formed by a mixture of K Gaussian components, $\mathcal{N}(\boldsymbol{\mu}_{jk}, \Sigma_{jk})$ where $k \in \{1 \ldots K\}$. The component parameters are the weight w_{jk}, mean $\boldsymbol{\mu}_{jk}$ and covariance Σ_{jk}. The 1st-order derivatives of the log probability of the sequence $\boldsymbol{O}$ with respect to each parameter are given below[1], where the derivatives have been defined as column vectors,

$$\nabla_{\boldsymbol{\mu}_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta}) \quad = \quad \sum_{t=1}^{T} \gamma_{jk}(t) S^T_{[t,jk]} \tag{4}$$

$$\nabla_{\text{vec}(\Sigma_{jk})} \ln p(\boldsymbol{O}|\boldsymbol{\theta}) \quad = \quad \sum_{t=1}^{T} \gamma_{jk}(t) \frac{1}{2}\left[-[\text{vec}(\Sigma^{-1}_{jk})]^T + \{S_{[t,jk]} \otimes S_{[t,jk]}\}\right]^T \tag{5}$$

[1] For the derivation of this score-space defined using this operator and an HMM generative model, see [2].

$$\nabla_{w_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \gamma_{jk}(t) \left[\frac{1}{w_{jk}} - \frac{\gamma_{j1}(t)}{w_{j1}\gamma_{jk}(t)} \right] \quad (6)$$

$$\text{where} \qquad S_{[t,jk]} = (\boldsymbol{o}_t - \boldsymbol{\mu}_{jk})^T \Sigma_{jk}^{-1} \quad (7)$$

$\gamma_{jk}(t)$ is the posterior probability of component k of state j at time t. Assuming the HMM is left-to-right with no skips and assuming that a state only appears once in the HMM (i.e. there is no state-tying), then the 1st-order derivative for the self-transition probability for state i, a_{ii}, is,

$$\nabla_{a_{ii}} \ln p(\boldsymbol{O}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \left[\frac{\gamma_i(t)}{a_{ii}} - \frac{1}{T a_{ii}(1 - a_{ii})} \right] \quad (8)$$

The 1st-order derivatives for each Gaussian parameter and self-transition probability in the HMM can be spliced together into a 'super-vector' which is the *score*[2].

From the definitions above, the score is the weighted sum of scores for the individual observations. If the scores for the same utterance spoken at different speaking rates were calculated, they would project into different regions of score-space simply because of their different number of observations. It is therefore important to normalise the score-space by the number of observations in each sequence. The method used in this work, *sequence length normalisation*, redefines the component posterior probabilities using state posterior probabilities normalised by the total state occupancy over the utterance. The new component posterior probabilities $\hat{\gamma}_{jk}(t)$ may be used instead of $\gamma_{jk}(t)$ in Equations 4 to 6 and 8,

$$\hat{\gamma}_{jk}(t) = \frac{\gamma_j(t)}{\sum_{\tau=1}^{T} \gamma_j(\tau)} \left(\frac{w_{jk}\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{jk}, \Sigma_{jk})}{\sum_{i=1}^{K} w_{ji}\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{ji}, \Sigma_{ji})} \right) \quad (9)$$

where $\gamma_j(t)$ is the posterior probability of state j at time t. The generative models should also be trained using $\hat{\gamma}_{jk}(t)$ (see [7] for details).

The nature of the score-space affects the discriminative power of the classifier, in this case an SVM, built in the score-space. The likelihood score-space is susceptible to *wrap-around* [7], caused when two different positions in acoustic-space map to a single point in score-space. For example, consider two classes modelled by two widely-spaced Gaussians. If an observation is generated at the peak of the first Gaussian, then the score relative to the mean of that Gaussian is zero by virtue of a zero gradient. However, the score relative to the mean of the distant second Gaussian is also zero by virtue of a zero component posterior probability. A similar problem occurs with an observation generated at the peak of the second Gaussian. This ambiguity in mapping two possible locations in acoustic-space to the zero of score-space represents a wrapping of the acoustic space onto the score-space. This also occurs in the score-space of the variances. Thus wrap-around increases the confusion between classes. A likelihood-ratio score-space constructed using these two Gaussians will not suffer wrap-around because the component posterior probabilities for each of the individual generative models will be forced to be unity.

So far, only 1st-order derivative score-operators have been considered. It is possible to extend the score operator to use, for example, zeroth, 1st and 2nd-order derivatives. Of course there is an interaction between the score-operator and the score-space. For example, the zeroth-order derivative for the likelihood score-space

[2] Due to the sum to unity constraints, one of the weight parameters in each Gaussian mixture is discarded from the definition of the super-vector, as are the forward transitions in the left-to-right HMM.

is expected to be less useful than its counter-part in the likelihood-ratio score-space because of its greater sensitivity to variations in acoustic conditions. It is useful to note that an SVM, with zero bias, constructed in the single-dimensional score-space formed by the likelihood-ratio would, with equal class priors, give a minimum Bayes error classifier. 2nd-order derivative operators may also be used to derive score-spaces such as $\nabla^T_{\boldsymbol{\mu}_{jk}}\left[\nabla_{\boldsymbol{\mu}_{jk}}\left(\ln p(\boldsymbol{O}|\boldsymbol{\theta})\right)\right]$, where for simplicity the component posterior probability $\gamma_{jk}(t)$ has been assumed independent of $\boldsymbol{\mu}_{jk}$,

$$\nabla^T_{\boldsymbol{\mu}_{jk}}\left[\nabla_{\boldsymbol{\mu}_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta})\right] = -\sum_{t=1}^{T} \gamma_{jk}(t)\Sigma_{jk}^{-1} \tag{10}$$

The zeroth, 1st and 2nd-order derivatives of the score-argument project into different partitions of the score-space. If the score-space is normalised or whitened, then the relative weightings of the partitions are unimportant. Alternatively, a Taylor-series expansion could be used to define the weightings.

Once the score-space has been defined, an SVM classifier can be built in the score-space. If standard linear, polynomial or Gaussian Radial Basis Function (GRBF) kernels are used, then the score-space is assumed to have a Euclidean tensor metric. The score-space should therefore first be whitened (i.e. decorrelated and scaled) before the kernels are applied. Failure to perform such *score-space normalisation* results in the Plain kernel [5], which is expected to perform poorly when the dimensions of score-space have different dynamic ranges [2]. Simple scaling has been found to be a reasonable approximation to this normalisation and avoids the need for inverting large matrices [2]. The Fisher kernel in [4] uses the Fisher Information matrix to normalise the score-space. This is only an acceptable normalisation for a likelihood score-space with 'matched' training and test sets (i.e. when the mean score is zero for all dimensions of the score space). The appropriate SVM kernel to use between two utterances $\boldsymbol{O}_i$ and $\boldsymbol{O}_j$ in the normalised score-space is therefore the *Normalised kernel*, $k_{\mathrm{N}}(\boldsymbol{O}_i, \boldsymbol{O}_j)$,

$$k_{\mathrm{N}}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \varphi^f_{\hat{F}}(\boldsymbol{O}_i)\Sigma_{\mathrm{sc}}^{-1}\varphi^f_{\hat{F}}(\boldsymbol{O}_j) \tag{11}$$

where Σ_{sc} is the covariance matrix in score-space.

3 Experimental Results

The performance of the proposed schemes was evaluated on a speaker-independent classification task of isolated letter utterances from the ISOLET database [8]. The data was coded into feature vectors at a 10msec frame rate with a 25.6msec window-size. The data was parameterised into 39-dimensional feature vectors comprising 12 MFCCs and a log energy term with corresponding delta and acceleration parameters. 240 utterances per letter from `isolet{1,2,3,4}` were used for training and 60 utterances per letter from `isolet5` for testing. There was no overlap between the training and test speakers. Two sets of experiments were conducted, the harder E-set subset, {`B C D E G P T V Z`}, and the full set of 26 letters. The baseline HMM system was trained using Maximum Likelihood (ML) estimation. Each letter was modelled by a 10-emitting state left-to-right continuous density HMM with no skips. Silence was modelled by a single emitting-state HMM with no skips. Each state output distribution had the same number of Gaussian components which was varied in the experiments. The models were tested using a Viterbi recogniser constrained to a silence-letter-silence network.

The baseline HMMs were used as generative models for SVM kernels. A modified version of SVM^{light} Version 3.02 [9] was used to train SVMs. A set of 1v1 SVM classifiers were trained on each possible class pairing. Sequence length normalisation was used during training and testing. Linear kernels were used in the normalised score-space, since they gave better performance than GRBFs of variable width and polynomial kernels of degree 2 (including homogeneous, inhomogeneous, and inhomogeneous with zero-meaned score-space). The linear kernel did not require parameter-tuning and was found to be insensitive to variations in the SVM trade-off parameter C. C was fixed at 100, and biased hyperplanes were permitted. A variety of score-spaces were examined. The abbreviations `m`, `v`, `w` and `t` refer to the score-spaces $\nabla_{\boldsymbol{\mu}_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta})$, $\nabla_{\mathrm{vec}(\Sigma_{jk})} \ln p(\boldsymbol{O}|\boldsymbol{\theta})$, $\nabla_{w_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta})$ and $\nabla_{a_{ii}} \ln p(\boldsymbol{O}|\boldsymbol{\theta})$ respectively. `r` refers to the log likelihood-ratio, $\ln[p_1(\boldsymbol{O}|\boldsymbol{\theta}_1)/p_2(\boldsymbol{O}|\boldsymbol{\theta}_2)]$ and `l` to the log likelihood $\ln p(\boldsymbol{O}|\boldsymbol{\theta})$. As this is a multi-class problem, the binary SVM classification results (and, as a baseline, the binary HMM results) must be combined to obtain a single classification for each utterance. A simple majority voting scheme among the full set of 1v1 binary classifiers was used (for tied letters, the relevant 1v1 classifiers were inspected and then, if necessary, random selection performed).

Table 1: Error-rates for score-spaces and generative models (ISOLET E-set)

Num comp. per class per state	HMM		SVM score-space		
	min. Bayes error	majority voting	lik-ratio	lik (1-class)	lik (2-class)
1	11.3	11.3	6.9	7.6	6.1
2	8.7	8.7	5.0	6.3	9.3
4	6.7	6.7	5.7	8.0	23.2
6	7.2	7.2	6.1	7.8	30.6

Table 1 compares the baseline HMM and SVM systems as the complexity of the generative models was varied. The baseline HMMs were comparable to state-of-the-art results on the ISOLET E-set [10]. The majority voting scheme gave the same performance as the minimum Bayes error scheme, indicating that majority voting was an acceptable multi-class scheme for the E-set experiments. For the SVMs, the likelihood-ratio score-space was defined for each binary problem using the competing class-conditional generative models projecting into the `mr` score-space. The likelihood (1-class) score-space was defined using only the generative model for the first of the two classes, and projected into the `ml` score-space. The likelihood (2-class) score-space was defined using a generative model for both classes, and also projected into the `ml` score-space (the original Fisher score, which is a projection into the `m` score-space, was also tested but was found to yield slightly higher error rates). SVMs built using the likelihood-ratio score-space achieved lower error rates than HMM systems, as low as 5.0%. The likelihood (1-class) score-space performed slightly worse than the likelihood-ratio score-space because it only contained about half the information and did not contain the log likelihood-ratio. The optimum number of components was 2 per state in both cases, possibly reflecting the gender division within each class. The likelihood (2-class) score-space performed poorest because of the confusion between classes due to wrap-around. The exception was for the generative model with 1 component per class per state (in total the model had 2 components per state because it modelled both classes). These 2 components generally did not reflect the gender division in the 2-class data as was at first supposed, but instead the class division. A probable explanation is that each Gaussian component modelled a bi-modal distribution produced by gender differences. Most of the data being modelled did not sit at the peaks of the two Gaussians and so did not produce zero scores. There were therefore relatively few observations at the ambiguous zero in the likelihood score-space, and hence there was still sufficient discrimination between classes in the score-space [7]. Contrary to [6] the HMM and

SVM confusion matrices revealed similar error structures, probably because of the dependence of the SVM kernel on the HMM.

Without score-space normalisation, the error-rate for class-conditional generative models with 2 components per state dropped from the baseline 5.0% to 11.1%. Tests using sequence length normalisation on artificially lengthened test data were inconclusive. The task was too simple but showed that score-space normalisation achieved a similar effect to sequence length normalisation. The baseline `mr` score-space was also augmented with the 2nd-order derivatives $\nabla^T_{\boldsymbol{\mu}_{jk}}[\nabla_{\boldsymbol{\mu}_{jk}} \ln p(\boldsymbol{O}|\boldsymbol{\theta})]$. The resulting classifiers showed increases in error rate up to 26% relative for generative models ranging from 1 to 4 components per state. The disappointing performance was probably due to the simplicity of the task and the assumption that component posterior probabilities were independent of component means.

It is commonly known that some dimensions of feature-space are noisy and degrade classification performance. For this reason, experiments were performed which selected subsets of the score-space and then built SVM classifiers in those subsets. First, the subsets were selected by parameter type. Error rates and dimensionalities for the resulting classifiers, otherwise identical to the baseline SVM, are detailed in Table 2. The generative models were class-conditional HMMs with 2 components per state. The log likelihood-ratio was shown to be a powerful discriminating feature[3]. Increasing the number of dimensions in score-space increased the discriminative power of the score-space. There was more discrimination in the derivatives of the component means than component variances. As expected in a dynamic task, the derivatives of the transitions and weights were also useful.

Table 2: Error rates for subsets of the likelihood-ratio score-space (ISOLET E-set)

score-space	error rate, %	score-space dimensionality
`r`	8.5	1
`v`	7.2	1560
`m`	5.2	1560
`mr`	5.0	1561
`mv`	5.0	3120
`wmv`	4.4	3140
`mvt`	4.4	3140
`wmvt`	4.1	3160
`wmvtr`	4.1	3161

Next, subsets of the baseline `mr` and full `wmvtr` score-spaces were selected according to dimensions with largest Fisher-ratios. The lowest error rates for the `mr` and `wmvtr` score-spaces were for 250 dimensions at 3.3% and for 300 to 600 dimensions at 3.5% respectively. Although the small test set of 540 utterances gave noisy results, adding the most discriminative dimensions lowered error-rate until less discriminative dimensions were added. The single most discriminative dimension was the log likelihood-ratio. The low-order MFCCs and the log energy term were the most important coefficients, but the delta and acceleration derivatives of the log energy term were also very useful. The 2nd and 4th most discriminatory dimensions were those dependent on the self-transition probability of state 2, and the most discriminatory weight terms were those belonging to states 1 and 2. This confirmed that most of the discrimination in the E-set lay in the initial letter states.

[3]The error rate at 8.5% differed from that for the HMM baseline at 8.7% because of the non-zero bias for the SVMs, and noise due to the small size of the task.

The HMM and SVM classifiers were run on the full 26 letter task. The best HMM classifier, with 4 components per state, gave 3.4% error rate. The optimum number of components possibly reflected the gender and class division in the data. For an SVM classifier using the `mr` score-space pruned to 200 dimensions by Fisher-ratios, a lower error rate of 2.9% was obtained. Though optimised on the E-set, the SVM classifier achieved better performance than well-trained HMM systems.

4 Conclusions

This paper presents work on the use of SVMs for classifying speech data. The paper has concentrated on the nature of the score-space when handling variable length speech sequences. The standard likelihood score-space of the Fisher kernel has been extended to the likelihood-ratio score-space, and issues of normalisation investigated. The new score-space avoids some of the limitations of the standard score-space, and allows the discriminative power of the generative models to be directly incorporated into the SVM classifier. The different score-spaces have been compared on a speaker-independent speech recognition task. The likelihood-ratio score-space out-performed both the likelihood score-space and standard HMMs.

Acknowledgements

N. Smith would like to thank EPSRC and his CASE sponsor, the Speech Group at IBM U.K. Laboratories, and Thorsten Joachims, University of Dortmund, for his software SVM^{light}.

References

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

[2] N. Smith, M. Gales, and M. Niranjan. Data-dependent kernels in SVM classification of speech patterns. Technical Report CUED/F-INFENG/TR.387, Cambridge University Eng.Dept., April 2001.

[3] I. Bazzi and D. Katabi. Using Support Vector Machines for Spoken Digit Recognition. In *Proceedings.* 6th International Conference on Spoken Language Processing, October 2000. Beijing, China.

[4] T. Jaakkola and D. Haussler. Exploiting Generative Models in Discriminative Classifiers. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems 11.* MIT Press, 1999.

[5] N. Oliver, B. Schölkopf, and A. Smola. *Advances in Large-Margin Classifiers*, chapter Natural Regularization from Generative Models. MIT Press, 2000.

[6] S. Fine, J. Navrátil, and R. Gopinath. A hybrid GMM/SVM approach to speaker identification. In *Proceedings*, volume 1, International Conference on Acoustics, Speech, and Signal Processing, May 2001. Utah, USA.

[7] N. Smith and M. Gales. Using SVMs to classify variable length speech patterns. Technical Report CUED/F-INFENG/TR.412, Cambridge University Eng.Dept., June 2001.

[8] M. Fanty and R. Cole. Spoken Letter Recognition. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Neural Information Processing Systems 3*, pages 220–226. Morgan Kaufmann Publishers, 1991.

[9] T. Joachims. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning.* MIT-Press, 1999.

[10] P.C. Loizou and A.S. Spanias. High-Performance Alphabet Recognition. *IEEE Transactions on Speech and Audio Processing*, 4(6):430–445, Nov. 1996.