Derivative and Parametric Kernels for Speaker Verification

C. Longworth and M.J.F. Gales

Engineering Department, Cambridge University
Trumpington St, Cambridge, CB2 1PZ
{cl336,mjfg}@eng.cam.ac.uk

Abstract

The use of Support Vector Machines (SVMs) for speaker verification has become increasingly popular. To handle the dynamic nature of the speech utterances, many SVM-based systems use dynamic kernels. Many of these kernels can be placed into two classes, *parametric kernels*, where the feature-space consists of parameters from the utterance-dependent model, and *derivative kernels*, where the derivatives of the utterance log-likelihood with respect to parameters of a generative model are used. This paper contrasts the attributes of these two forms of kernel. Furthermore, the conditions under which the two forms of kernel are identical are described. Two forms of dynamic kernel are examined in detail, based on MLLR-adaptation and mean MAP-adapted models. The performance of these kernels is evaluated on the NIST SRE 2002 dataset. Combining the two forms of kernel together gave a 35% relative reduction in Equal Error Rate compared to the best individual kernel.

Index Terms: Speaker Verification, Support Vector Machines, dynamic kernels.

1. Introduction

Speaker verification is a binary classification task in which the objective is to determine whether or not a speech utterance was uttered by a specific claimed speaker. The standard approach to text-independent speaker-verification uses Gaussian Mixture Models (GMMs) as the acoustic model. Normally, a Universal Background Model (UBM) is trained to represent all speakers using a large amount of development data. A speaker-dependent model is then obtained by using Maximum A-Posteriori (MAP) adaptation to robustly adapt the UBM using a small amount of enrolment data associated with each speaker. The verification data is then classified, as to whether the claimed identity is correct, using Bayes' decision rule[1].

Recently, there has been considerable interest in the use of Support Vector Machines (SVMs) for speaker verification. SVMs are general purpose classifiers that have been found to perform well on a wide range of classification tasks. However, SVMs are normally only able to classify data of fixed dimensionality whereas speech utterances are typically parameterised as variable length sequences of observation vectors. This has led to the use of *dynamic kernels*, also known as sequence kernels. Dynamic kernels implicitly map variable length observation sequences into a fixed-dimensional vector and are often based on generative models.

This paper describes how many proposed dynamic kernels for speaker verification can be classified into one of two types, *parametric kernels* and *derivative kernels*. These two forms of kernel generally extract different speaker-dependent features and thus may be complementary to one another. However, under certain conditions the features expressed by the kernels are identical. The nature of the kernels is also dependent on the generative model used to represent the individual speakers. Two forms of dynamic kernel speaker models are examined in detail in this paper. The first is based on parameters of the GMM, related to the GMM/mean-supervector kernel [2] and generative kernels [3]. The second uses MLLR-adaptation similar to the MLLR kernel described in [4]. This paper is organised as follows. The next section introduces dynamic kernels and describes common examples of parametric and derivative kernels. It is then shown that under certain restricted conditions, the features generated by the two forms of kernel are identical. In Section 3, experimental results on the NIST 2002 SRE dataset are presented. Finally, conclusions are drawn.

Chris Longworth would like to thank the Schiff Foundation for funding.

2. Dynamic kernels

SVMs have been successfully applied to a wide range of machine learning problems. One reason for this popularity is that they can be kernelised. In SVM training and inference all references to data are in the form of inner-products between data examples. It is then possible to define a *kernel function* $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that implicitly calculates the inner-product between two vectors in some, possibly very high dimensional, *feature-space*. One issue when applying SVMs to speech processing tasks is that the SVM can only perform classification on data of fixed dimensionality. However speech utterances are typically variable length sequences. This has led to the development of dynamic kernels, also known as sequence kernels. These kernels are often based on generative models and have the form

$$K(\boldsymbol{O}_i, \boldsymbol{O}_j; \boldsymbol{\lambda}) = < \boldsymbol{\phi}(\boldsymbol{O}_i; \boldsymbol{\lambda}), \boldsymbol{\phi}(\boldsymbol{O}_j; \boldsymbol{\lambda}) > \quad (1)$$

where $\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda})$ is a function that maps a speech utterance into a fixed dimensional feature space and $\boldsymbol{\lambda}$ is the set of parameters associated with the generative model. The kernel also defines the distance metric between two feature vectors. One such metric that is maximally non-committal is

$$K(\boldsymbol{O}_i, \boldsymbol{O}_j; \boldsymbol{\lambda}) = \boldsymbol{\phi}(\boldsymbol{O}_i; \boldsymbol{\lambda})^{\mathrm{T}} \boldsymbol{Q}^{-1} \boldsymbol{\phi}(\boldsymbol{O}_j; \boldsymbol{\lambda}) \quad (2)$$

where $\boldsymbol{Q}$ is the Fisher information matrix defined as

$$\boldsymbol{Q} = \mathcal{E}\left\{(\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}) - \boldsymbol{\mu}_\phi)(\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda}) - \boldsymbol{\mu}_\phi)^{\mathrm{T}}\right\} \quad (3)$$

$$\boldsymbol{\mu}_\Phi = \mathcal{E}\{\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda})\} \quad (4)$$

where $\mathcal{E}\{\}$ is the expectation with respect to $\boldsymbol{O}$. A number of different dynamic kernels of this form have been proposed for speaker verification, for example the MLLR-kernel [4], GMM/mean-supervector kernel [2] and Fisher kernel [5]. These kernels can be characterised into two broad classes depending upon the form of $\boldsymbol{\phi}(\boldsymbol{O}; \boldsymbol{\lambda})$. These will be referred to as *parametric kernels* and *derivative kernels*.

2.1. Parametric kernels

Parametric kernels are a form of dynamic kernel where the feature-space is the parameters $\boldsymbol{\lambda}$ associated with the generative model trained to represent the verification utterance $\boldsymbol{O}^{(v)}$. Thus

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{O}^{(v)}) = \left[\hat{\boldsymbol{\lambda}}\right], \quad \hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}}\{\log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda})\} \tag{5}$$

One property of this form of kernel is that the derivative at the ML estimate is zero, i.e.

$$\nabla_{\boldsymbol{\lambda}} \log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda})\Big|_{\hat{\boldsymbol{\lambda}}} = \mathbf{0} \tag{6}$$

The precise nature of the parametric kernel is determined by the generative model used to represent the speaker.

One parametric kernel that has been successfully used for speaker verification is the GMM/mean-supervector kernel [2]. In this kernel, the feature-space is the concatenated means of an utterance-dependent GMM. As there are typically not enough observations per component to robustly estimate the parameters, MAP adaptation, using the UBM as a prior, is used instead. Here

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}}\{\log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda})\} \tag{7}$$

where $p(\boldsymbol{\lambda})$ is based on the UBM parameters. For a GMM the ML, or MAP estimate, has no closed-form solution. Multiple iterations of model adaptation using EM are therefore used. This iterative approach means that equation 6 does not necessarily hold for limited numbers of iterations. For component m the MAP-adapted mean at iteration k is given by

$$\boldsymbol{\mu}_m^{(k)} = \frac{\sum_{t=1}^{T} \gamma_m^{(k-1)}(t)\boldsymbol{o}_t^{(v)} + \tau\tilde{\boldsymbol{\mu}}_m}{\sum_{t=1}^{T} \gamma_m^{(k-1)}(t) + \tau} \tag{8}$$

where $\tilde{\boldsymbol{\mu}}_m$ are the UBM means associated with component m (which are also used as the initial parameters $\boldsymbol{\mu}_m^{(0)}$), $\gamma_m^{(k)}(t) = P(m|\boldsymbol{o}_t^{(v)};\boldsymbol{\lambda}^{(k)})$, the posterior probability of component m at time t given observation $\boldsymbol{o}_t^{(v)}$ and $\boldsymbol{\lambda}^{(k)}$, and τ is the standard MAP adaptation constant that controls the influence of the prior on the final model. If k-iterations of mean-only MAP adaptation are performed the feature-space is

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{O}^{(v)};\boldsymbol{\lambda}^{(k)}) = \left[\boldsymbol{\mu}_1^{(k)\mathsf{T}}, \ldots, \boldsymbol{\mu}_M^{(k)\mathsf{T}}\right]^{\mathsf{T}} \tag{9}$$

In [2], a distance metric is defined such that the kernel function is an upper bound on the KL divergence between the two utterance-dependent models. This normalises each component mean by the associated mixture weight and the inverse of the covariance matrix. In this work the distance metric given in equation 2 is used, which is consistent with the metrics used for the other kernels.

Another parametric kernel proposed for speaker-verification is the MLLR kernel [4]. MLLR is an adaptation technique in which a linear transform of the canonical model means, here the UBM means, is used to represent a speaker. Like MAP, an iterative EM-based training scheme can be used to gradually increase the likelihood of $\boldsymbol{O}^{(v)}$. At iteration k the adapted mean, $\hat{\boldsymbol{\mu}}_m^{(k)}$, associated with component m is given by

$$\hat{\boldsymbol{\mu}}_m^{(k)} = \boldsymbol{A}^{(k)}\boldsymbol{\mu}_m + \boldsymbol{b}^{(k)} = \boldsymbol{W}^{(k)}\boldsymbol{\xi}_m \tag{10}$$

where $\boldsymbol{\xi}_m$ is the extended mean vector $\boldsymbol{\xi}_m = [\tilde{\boldsymbol{\mu}}_m^{\mathsf{T}} \;\; 1]^{\mathsf{T}}$, and the ith row, $\boldsymbol{w}_i^{(k)}$, of the MLLR transform is

$$\boldsymbol{w}_i^{(k)\mathsf{T}} = \boldsymbol{G}_{\boldsymbol{i}}^{(\boldsymbol{k})-\mathbf{1}}\boldsymbol{k}_i^{(k)} \tag{11}$$

where $\boldsymbol{G}_i^{(k)}$ and $\boldsymbol{k}_i^{(k)}$ are sufficient statistics defined as

$$\boldsymbol{G}_i^{(k)} = \sum_{m=1}^{M} \frac{1}{\sigma_{mi}^2}\boldsymbol{\xi}_m\boldsymbol{\xi}_m^{\mathsf{T}} \sum_{t=1}^{T} \gamma_m^{(k-1)}(t) \tag{12}$$

$$\boldsymbol{k}_i^{(k)} = \sum_{m=1}^{M}\sum_{t=1}^{T} \gamma_m^{(k-1)}(t)\frac{1}{\sigma_{mi}^2} o_{ti}^{(v)}\boldsymbol{\xi}_m \tag{13}$$

The MLLR parametric feature-space is then defined as $\phi_{\boldsymbol{\lambda}}(\boldsymbol{O}^{(v)};\boldsymbol{W}^{(k)}) = [\mathbf{vec}(\boldsymbol{W}^{(k)})]$, where **vec**() converts the matrix to a vector.

2.2. Derivative kernels

Derivative kernels provide an interesting contrast to parametric kernels. Rather than using model parameters as the feature-space, the partial derivatives of the utterance log-likelihood with respect to individual model parameters are used instead. For a set of model parameters, $\boldsymbol{\lambda}$, the derivative feature-space generated from a verification utterance $\boldsymbol{O}^{(v)}$ has the form

$$\phi_{\nabla}(\boldsymbol{O}^{(v)};\hat{\boldsymbol{\lambda}}) = \frac{1}{T}\left[\nabla_{\boldsymbol{\lambda}} \log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda})\Big|_{\hat{\boldsymbol{\lambda}}}\right] \tag{14}$$

where $\hat{\boldsymbol{\lambda}}$ is the model parameter value at which the derivative is evaluated. Equation 14 includes an optional term to normalise by the number of frames T in $\boldsymbol{O}^{(v)}$. This is important if the utterances in the dataset vary greatly in duration. A derivative kernel may also include higher-order derivative terms in the feature-space, however generally only first-order derivatives are used. It is necessary to define the point around which the derivative kernel feature-space will be evaluated. This may be based on the UBM parameters, which is similar to using the Fisher kernel [5]. Another possibility is to use the speaker-specific parameters. As a GMM is typically used, iterative approaches are used to obtain the speaker-specific parameters. To clearer specify the iteration at which the derivative is evaluated, $\log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda}^{(k)})$, will be used for the feature-space evaluated at the k^{th} iteration. This approach resembles the log-likelihood ratio kernel [3].

The precise nature of derivative kernels is again determined by the generative model used to represent a speaker. Derivatives with respect to the means of the GMM can be used [6]. Here

$$\nabla_{\boldsymbol{\mu}_m} \log p(\boldsymbol{O}^{(v)};\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}^{(k)}} = \sum_{t=1}^{T} \gamma_m^{(k)}(t)\boldsymbol{\Sigma}_m^{-1}(\mathbf{o}_t^{(v)} - \boldsymbol{\mu}_m^{(k)}) \tag{15}$$

Alternatively derivatives with respect to an MLLR transform may also be used, effectively this is the derivative form of the parametric MLLR-kernel. The derivatives of $\log p(\boldsymbol{O}^{(v)};\boldsymbol{W})$ with respect to the ith row of transform $\boldsymbol{W}$ evaluated at the point $\boldsymbol{W}^{(k-1)}$ can be expressed as

$$\nabla_{\boldsymbol{w}_i} \log p(\boldsymbol{O}^{(v)};\boldsymbol{W})\Big|_{\boldsymbol{W}^{(k-1)}} = \boldsymbol{G}_i^{(k)}\boldsymbol{w}_i^{(k-1)\mathsf{T}} - \boldsymbol{k}_i^{(k)} \tag{16}$$

where $\boldsymbol{G}_i^{(k)}$ and $\boldsymbol{k}_i^{(k)}$ are the sufficient statistics defined in equations 12 and 13. Note these statistics are identical to those used for the MLLR parametric kernel. In this work a global transform is used, the extension to multiple regression classes is trivial. The MLLR derivative feature-space is then defined by simply mapping the derivatives corresponding to each element of the MLLR transform into a single feature vector and optionally normalising by the utterance duration. This *MLLR-derivative kernel* is then given by

$$\phi_{\nabla}(\boldsymbol{O}^{(v)};\boldsymbol{W}^{(k)}) = \frac{1}{T}\left[\mathbf{vec}\left(\nabla_{\boldsymbol{W}} \log p(\boldsymbol{O}^{(v)};\boldsymbol{W})\Big|_{\boldsymbol{W}^{(k)}}\right)\right]$$

It is interesting to briefly contrast derivative kernels with parametric kernels. From equation 6, the derivative of the parametric kernel features at the ML-estimate of the model parameters will be zero for the verification data $\boldsymbol{O}^{\text{v}}$. In general this will not be the case for the derivative kernel. Instead the features of the derivative kernel will be zero for the enrolment data at the ML-estimate, i.e.

$$\hat{\boldsymbol{\lambda}}_e = \arg\max_{\boldsymbol{\lambda}}\{\log p(\boldsymbol{O}^{\text{e}};\boldsymbol{\lambda})\}, \quad \boldsymbol{\phi}_{\nabla}(\boldsymbol{O}^{\text{e}};\hat{\boldsymbol{\lambda}}_e) = \mathbf{0} \qquad (17)$$

In addition the derivative kernels commonly use a length normalisation term. This is not necessary for parametric kernels, where there is an implicit normalisation for the lengths. A consequence of this is that when a component, or generally a transform class, is not observed ML-based parametric kernels are undefined, whereas derivative kernels tend to zero.

2.3. Conditions for complementary feature sets

Both parametric and derivative kernels have been used successfully for speaker-verification. The respective feature-spaces can express different types of speaker-discriminant information and thus may be complementary. It is useful to establish under what conditions the two forms of kernel are the same, as this yields information as to how to make the features complementary to one another. The parametric kernel feature-space at the kth iteration of training can be expressed in the form of a gradient ascent update.

$$\boldsymbol{\phi}_{\boldsymbol{\lambda}}(\boldsymbol{O}^{\text{v}};\boldsymbol{\lambda}^{(k+1)}) = \left[\boldsymbol{\lambda}^{(k)} + \tilde{\eta}\nabla_{\boldsymbol{\lambda}}\log p(\boldsymbol{O}^{\text{v}};\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}^{(k)}}\right] \qquad (18)$$

where $\tilde{\eta}$ is the learning rate. This may additionally be expressed as a function of a derivative feature-space evaluated at $\boldsymbol{\lambda}^{(k)}$

$$\boldsymbol{\phi}_{\boldsymbol{\lambda}}(\boldsymbol{O}^{\text{v}};\boldsymbol{\lambda}^{(k+1)}) = \left[\boldsymbol{\lambda}^{(k)} + \eta\boldsymbol{\phi}_{\nabla}(\boldsymbol{O}^{\text{v}};\boldsymbol{\lambda}^{(k)})\right] \qquad (19)$$

where $\eta = T\tilde{\eta}$ if duration normalisation is used in equation 14, otherwise $\eta = \tilde{\eta}$. The two classes of dynamic kernel are thus related to each other. Compared to the derivative kernel feature-space, the parametric kernel features includes a term $\boldsymbol{\lambda}^{(k)}$ which introduces a translation of the feature space. If a kernel that is invariant to translation is used, such as a stationary kernel, this will have no effect. Note stationary kernels have the general form $K(\boldsymbol{O}_i,\boldsymbol{O}_j) = \mathcal{F}(\boldsymbol{\phi}(\boldsymbol{O}_i) - \boldsymbol{\phi}(\boldsymbol{O}_j))$ where $\mathcal{F}()$ is the function that defines the kernel.

Even if a stationary kernel is used, it is not sufficient to ensure that the two sets of features will be identical. Equation 19 contains a learning rate. Using an appropriate metric, the kernels will not depend on the learning rate if the learning rate is independent of the observation sequence since this dependency is removed by the metric (the metric used in equation 2 has this property but is not stationary). However this is not generally the case. To illustrate this consider the situation where the parametric kernel is obtained using an EM-based ML-estimation of the mean. At iteration $k+1$ the mean parametric feature-space for component m can be expressed as

$$\boldsymbol{\mu}_m^{(k+1)} = \boldsymbol{\mu}_m^{(k)} + \left(\frac{T\boldsymbol{\Sigma}_m}{\sum_{t=1}^{T}\gamma_m^{(k)}(t)}\right)\left[\frac{1}{T}\nabla_{\mu_m}\log p(\boldsymbol{O}^{\text{v}};\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}^{(k)}}\right] \qquad (20)$$

EM is thus equivalent to gradient ascent using the derivative features with a learning rate that depends upon the total component occupancy for that observation sequence as well as the length of the observation sequence (when length normalisation is being used).

If both parametric and derivative features are to be used, it is important that the features differ. This can be achieved using a non-stationary kernel, such as the kernel in equation 2, evaluating the derivative terms at a different point to the parametric features (effectively using a different number of iterations), or simply using the standard EM updates. Combinations of these may increase how complementary the features are.

3. Experimental results

The parametric and derivative dynamic kernels were evaluated on the 2002 NIST SRE one-speaker detection task[7]. Each utterance was parameterised using a frame rate of 10ms and a window size of 30ms. 31 features were extracted per frame, these consisted of 15 static, 15 delta Mel-PLP coefficients and the delta energy. Cepstral Mean Subtraction was performed on each utterance followed by Cepstral Feature Warping[8] using a three second window to introduce additional robustness to channel noise. This setup is identical to that used in [9]. System performance was primarily evaluated using the Equal Error Rate metric. NIST SRE evaluations are evaluated by means of a Detection Cost Function (DCF) This is the weighted sum of the False alarm and Miss probabilities at a defined threshold. The normalised cost used in this paper takes the form [7]

$$\text{DCF} = P_{\text{Miss}} + 9.9P_{\text{False Alarm}} \qquad (21)$$

To aid comparison with other work the minDCF score, obtained a-posteriori by adjusting the decision threshold, is quoted in addition to the EER.

Initially, gender-dependent UBMs were trained by EM using all SRE 2002 enrolment utterances of the appropriate gender. Each UBM consisted of a 128-component GMM with each component having diagonal covariance matrices. For the MAP-adapted mean kernels, speaker-dependent GMMs were constructed by MAP adapting the means of the appropriate gender-dependent UBM. Two iterations of static prior MAP were used with τ set at 25. These speaker specific models were used for the parametric kernels and the point at which the derivative kernels were evaluated[1]. Dynamic kernels based upon MLLR features were also evaluated. For these kernels, a single iteration of MLLR adaptation was used to adapt the means of the appropriate UBM. Preliminary experiments showed that additional iterations of adaptation provided only negligible gains. During adaptation, a full transform was trained, tied over all components. For both the MAP-adapted mean based kernels and the MLLR-based kernels, the non-stationary kernel given in equation 2 was used. A diagonal approximation was used for the Fisher information matrix, estimated based on the covariance matrix features extracted from the enrolment utterances. For each speaker, an imposter training set was created using all the non-speaker enrolment utterances. When combining multiple feature space-types, the features were simply concatenated together to form a single feature-space. Unlike [10], weights were not assigned to individual feature sets. In addition to the use of SVM for verification, baseline results were obtained using standard verification with either the MAP-adapted mean speaker models or the MLLR-adapted speaker models and the UBM. In this setup, non-stationary kernels were used, parametric and derivative terms were evaluated at different points and standard EM updates were used. This was intended to increase the complementariness of the parametric and derivative feature sets.

[1] From equation 19 the parametric and derivative kernel features were evaluated at different iterations.

Feature Set	MAP		MLLR	
	EER(%)	minDCF	EER(%)	minDCF
GMM	12.17	0.5014	17.51	0.5855
ϕ_{λ}	9.89	0.4220	20.12	0.7755
ϕ_{∇}	8.62	0.3723	15.36	0.5825
$\phi_{\lambda} + \phi_{\nabla}$	**5.60**	**0.2416**	14.99	0.5817

Table 1: Performance for parametric (ϕ_{λ}), derivative (ϕ_{∇}) and composite ($\phi_{\lambda} + \phi_{\nabla}$) MAP and MLLR feature sets.

Evaluation performance with the baseline systems and the various dynamic kernels are presented in Table 1. Both parametric and derivative MAP kernels gave significant gains compared to the baseline GMM system. This is consistent with previous work such as [2] and [6]. For both MAP and MLLR adaptation, the derivative kernel outperformed the parametric kernel. Note for the parametric MAP-adapted mean kernel, a different metric was used compared to the results in [2]. When the MAP parametric and derivative feature sets were combined into a single kernel, a further 35% relative reduction in EER was obtained compared to the MAP derivative kernel indicating that the feature sets are complementary. Although overall performance was significantly worse for MLLR-based kernels combining parametric and derivative features still produced gains.

In addition to the results in Table 1 it is possible to combine mean MAP-based kernels and MLLR-based kernels[2]. A combined MAP and MLLR derivative kernel gave an EER of 8.02% representing an 7% relative reduction in EER compared to the MAP derivative kernel alone. Additionally combining this system with the MAP parametric kernel gave an improved performance of 6.36% EER. However this did not exceed the performance obtained by combining only the MAP parametric and derivative kernels. Overall, the best performing dynamic kernel was the composite MAP kernel.[3], which gave a 55% relative reduction in EER compared to the Baseline system.

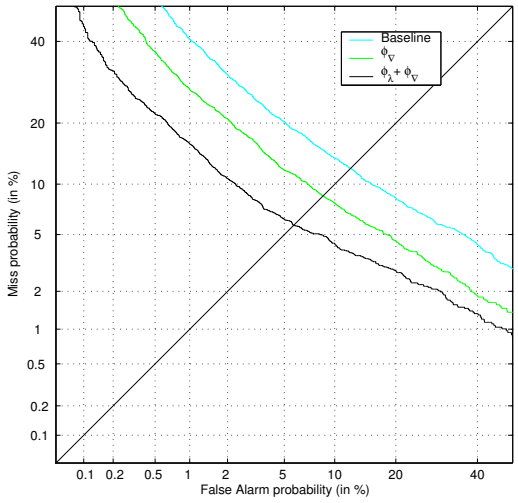

Figure 1: DET curve comparing baseline against derivative and composite MAP dynamic kernels.

[2]Combining likelihood features with parametric or derivative gave little change in performance due to maximally non-committal metric being used.

[3]The composite MAP kernel also outperformed the results reported in [9] which were obtained using much larger 1024-component GMMs

Figure 1 shows the DET curve for the baseline mean MAP-adapted system, the associated derivative kernel and combination of the derivative and parametric kernel. It is clear from the graph that on this task the combined kernel outperforms the derivative kernel at all operating points.

4. Conclusions

This paper has discussed two general forms of dynamic kernel, parametric and derivative kernels. The two sets of features produced have different properties and are generally complementary. However, under certain conditions, discussed in Section 2.3, the feature-spaces produced will be identical. Various dynamic kernels were evaluated using the NIST 2002 evaluation data. Both parametric and derivative MAP-based kernels individually provided gains compared to the baseline GMM system. Furthermore when the feature sets were combined a further 35% relative reduction in EER was observed compared to the best single feature set. MLLR-based dynamic kernels were also investigated. Although overall performance for these MLLR-based kernels was worse than for MAP-kernels, MLLR-based parametric and derivative kernels were again found to be complementary. The experimental results shown here are intended as an illustration of the merits of combining parametric and derivative feature sets rather than as an example of an evaluation-level system as the imposter and UBM training utterances were not obtained from an auxiliary dataset. Further performance gains should be achievable. For example, by implementing RASTA filtering, using larger Gaussian models or conducting T-normalisation.

5. References

[1] D. Reynolds, "Speaker identification and verification using gaussian mixture models," *Speech Communication*, vol. 17, pp. 91–105, 1995.

[2] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006.

[3] M. Gales and M. Layton, "Training augmented models using SVMs," *IEICE Special Issue on Statistical Models for Speech Recognition*, 2006.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A.Venkataramam, "MLLR transforms as features in speaker recognition," in *Interspeech*, 2005.

[5] T. Jaakkola and D. Hausser, "Exploiting generative models in disciminative classifiers," in *NIPS*, 1999.

[6] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Transactions Speech and Audio Processing*, 2004.

[7] A. Martin, "The NIST year 2002 speaker recognition evaluation plan," 2002, available from http://www.nist.gov/speech/tests/spk/2002/doc.

[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, 2001.

[9] C.Longworth and M.J.F.Gales, "Discriminative adaptation for speaker verification," in *Proc. ICSLP*, 2006.

[10] A. Hatch, A. Stolcke, and B. Peskin, "Combining feature sets with support vector machines: Application to speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.