LIMITED TRAINING DATA ROBUST SPEECH RECOGNITION USING KERNEL-BASED ACOUSTIC MODELS

Martin Schafföner, Sven E. Krüger, Edin Andelic, Marcel Katz, Andreas Wendemuth

Dept. of Electrical Engineering and Information Technology
Otto-von-Guericke-University
P. O. Box 4120, 39016 Magdeburg, Germany
martin.schaffoener@e-technik.uni-magdeburg.de

ABSTRACT

Contemporary automatic speech recognition uses Hidden-Markov-Models (HMMs) to model the temporal structure of speech where one HMM is used for each phonetic unit. The states of the HMMs are associated with state-conditional probability density functions (PDFs) which are typically realized using mixtures of Gaussian PDFs (GMMs). Training of GMMs is error-prone especially if training data size is limited. This paper evaluates two new methods of modeling state-conditional PDFs using probabilistically interpreted Support Vector Machines and Kernel Fisher Discriminants. Extensive experiments on the RM1 [1] corpus yield substantially improved recognition rates compared to traditional GMMs. Due to their generalization ability, our new methods reduce the word error rate by up to 13% using the complete training set and up to 33% when the training set size is reduced.

1. INTRODUCTION

Kernel Fisher Discriminants (KFDs) and Support Vector Machines (SVMs) represent two recent approaches to pattern classification. They have attracted much interest because they are capable of generalizing well, which often results in much better performance compared to conventional techniques such as, e. g., artificial neural networks.

KFDs and SVMs have been successfully used in a wide range of applications, e. g. handwriting recognition or protein classification. SVMs have also been applied to speech-related problems such as phonetic classification [2] or post-scoring of speech recognition hypotheses [3]. However, both approaches do not integrate KFDs or SVMs into the process of continuous speech recognition. Attempts have been made in [4] where a tied-posterior framework is applied, and in [5] for phoneme recognition.

The task of automatic speech recognition is to deduce the most likely text (sequence of words) $\hat{\boldsymbol{w}}$ from a given sequence $\boldsymbol{X}$ of M observations $\boldsymbol{x}$ [6]:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\ P(\boldsymbol{w}|\boldsymbol{X}) = \underset{\boldsymbol{w}}{\operatorname{argmax}}\ \frac{P(\boldsymbol{X}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{X})} \tag{1}$$

where $P(\boldsymbol{X}|\boldsymbol{w})$ is called the acoustic model and $P(\boldsymbol{w})$ is known as the language model. The acoustic model is usually a combination of a lexicon breaking words into (sub-)phonetic units and HMMs modeling each of these units. HMMs are finite-state automata especially capable of handling temporal dynamics. Each state s is associated with an emission-probability $P(\boldsymbol{x}|s)$ which for continuous variables $\boldsymbol{x}$ is replaced with its PDF. These PDFs are usually realized using weighted sums of elementary Gaussian PDFs (Gaussian Mixture Models – GMMs). However, determining the parameters of such mixture-models is error-prone [7], especially if estimation material is limited.

To overcome this problem, our approach models the emission probabilities of the HMMs using probabilistically interpreted KFDs or SVMs [8, 9]. Both KFD and SVM project data (samples) $\boldsymbol{x} \in \mathbb{R}^n$ onto a one-dimensional direction $\boldsymbol{w}$ which optimally separates two classes (labels) w. r. t. KFD's or SVM's specific criteria. The direction $\boldsymbol{w}$ may be non-linearly related to the input space $\mathbb{R}^n$ (cf. section 2) and can be constructed by the SVM using only very few samples. Using this simple one-dimensional representation of the problem, models for class-conditional probabilities (a. k. a. emission probabilities) may be estimated more easily and more robustly than is often possible using the original setting in $\mathbb{R}^n$.

The paper is organized as follows: A brief review of KFD and SVM is given. Different methods for their probabilistic interpretation are discussed. Baseline experiments and results are discussed before the two new methods are evaluated and compared. A conclusion finishes the paper.

2. KERNEL-BASED ACOUSTIC MODELS

Both Fisher's Discriminant (FD) and SVM are linear classifiers which can be extended to nonlinear classification by way of the so-called *kernel trick*. Instead of applying the classifiers directly to the input space $\mathbb{R}^n$ they can be applied to a *feature-space* $\mathcal{F}$ of higher, possibly infinite dimensionality $d > n$. The feature-space $\mathcal{F}$ is nonlinearly related to the input space through a mapping $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathcal{F}^d$. A *kernel function* $\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}')$ satisfying Mercer's conditions then computes a dot-product in $\mathcal{F}$ [10]:

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{\Phi}(\boldsymbol{x}) \cdot \boldsymbol{\Phi}(\boldsymbol{x}')) \tag{2}$$

The most commonly used kernel function is the Gaussian radial basis function (RBF)

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{||\boldsymbol{x}-\boldsymbol{x}'||^2}{2\sigma^2}} \tag{3}$$

Since both FD and SVM can be expressed in terms of Euclidean dot-products $(\boldsymbol{x} \cdot \boldsymbol{x}')$ only, the extension to nonlinear classification can be achieved by replacing the Euclidean dot-product by the dot-product in $\mathcal{F}$, which is $\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}')$ from (2).

Martin Schafföner acknowledges funding by the Friedrich Naumann Foundation.

2.1. Kernel Fisher Discriminant

The Fisher Discriminant is a well known heuristic approach for two-class discrimination problems [11]. Consider a training set $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ belonging to an input space $\mathcal{X}$ and consisting of M samples which are split into two classes. Let the classes be labeled with -1 and 1 defining a corresponding label vector $\boldsymbol{y} = \{-1, 1\}^M$. The number of samples labeled with -1 and 1 is M_1 and M_2, respectively. The corresponding class means are $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, respectively. Successful discrimination of the samples can be achieved by finding a direction $\boldsymbol{w}$ where at the same time class means are maximally separated and class variances are minimal. Thus one has to maximize the coefficient

$$R(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{4}$$

with $\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$ and $\boldsymbol{S}_W = \sum_{i,y_i=-1}(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^T + \sum_{i,y_i=1}(\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^T$ denoting the unnormalized between-class and within-class covariance matrices (often referred to as scatter matrices), respectively. By differentiating (4) one can see that $\boldsymbol{w}$ is the leading eigenvector of the generalized eigenvalue problem $\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$.

Instead of working on the original input space we apply the above mentioned nonlinear mapping $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathcal{F}$ to the data $\boldsymbol{X}$, arriving at Kernel FD [12]. Thus (4) becomes

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}} \tag{5}$$

where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^M$, $\boldsymbol{M} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$, $\boldsymbol{m}_i = \boldsymbol{K}_i \boldsymbol{u}_i, \quad i = 1, 2$. $\boldsymbol{K}_i$ denotes the kernel matrix of class i and has elements $[k^i_{jk} = \mathrm{k}(\boldsymbol{x}_j, \boldsymbol{x}^i_k)]^{j=M, k=M_i}_{j,k=1}$. The kernel matrix of the complete data set is $\boldsymbol{K} = [k_{ij} = \mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j)]^M_{i,j=1}$. The column vector $\boldsymbol{u}_i$ contains M_i elements with a common value of M_i^{-1}. $\boldsymbol{N}$ is given by $\boldsymbol{N} = \sum_{i=1,2} \sum_{m, y_m = y_i} (\boldsymbol{\Phi}(x_m) - \boldsymbol{m}_i)(\boldsymbol{\Phi}(x_m) - \boldsymbol{m}_i)^T$.

Equivalently to the FD, the coefficients $\boldsymbol{\alpha}$ of the introduced Kernel Fisher Discriminant (KFD) are given by the leading eigenvector of $\boldsymbol{N}^{-1}\boldsymbol{M}$. However, $\boldsymbol{N}$ may become ill-conditioned or even singular. So one has to impose some kind of regularization on $\boldsymbol{N}$, e. g. $\boldsymbol{N}_C = \boldsymbol{N} + C\boldsymbol{I}, C \in \mathbb{R}$ where $\boldsymbol{I}$ denotes the identity matrix. The KFD classifier is given by the projections $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{K}\boldsymbol{\alpha} + b\mathbf{1}$ onto the direction $\boldsymbol{w} = \sum_{i=1}^M \alpha_i \boldsymbol{\Phi}(\boldsymbol{x}_i)$. The vector $\mathbf{1}$ contains M elements with a common value of 1 and

$$b = -\boldsymbol{\alpha} \frac{M_1 \boldsymbol{m}_1 + M_2 \boldsymbol{m}_2}{M} \tag{6}$$

denotes the bias.

The KFD is equivalent to a regression to the labels contained in $\boldsymbol{y}$ [12]. Thus instead of solving a generalized eigenvalue problem imposed by (5) one can obtain a KFD by solving the convex quadratic optimization problem

$$\min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \quad \|\boldsymbol{\xi}\|^2 + CP(\boldsymbol{\alpha}) \tag{7}$$

with

$$\boldsymbol{\xi} = \boldsymbol{y} - (\boldsymbol{K}\boldsymbol{\alpha} + \mathbf{1}b) \tag{8}$$

and subject to the conditions

$$\mathbf{1}_1^T \boldsymbol{\xi} = 0, \quad \mathbf{1}_2^T \boldsymbol{\xi} = 0 \tag{9}$$

with $\mathbf{1}_1 = \max(-\boldsymbol{y}, 0)$ and $\mathbf{1}_2 = \max(\boldsymbol{y}, 0)$. C is a regularization constant which is still mandatory due to the equivalence to KFD. P is a regularization functional, typically $P(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|^2$. Thus the smaller C the tighter the solution $\boldsymbol{\alpha}$ is allowed to fit the training data.

To decrease computational complexity, a sparse solution $\boldsymbol{\alpha}$ is desired but not generally given. However, a sparse approximation to the complete solution can be achieved iteratively, e. g. using the greedy algorithm described in [12].

2.2. Support Vector Machines

We only give a brief overview; further information can be found in e. g. [13, 14].

Given a training set $\{\boldsymbol{x}_i, y_i\}$ of M samples $\boldsymbol{x}_i \in \mathbb{R}^n$ and corresponding labels $y_i \in \{-1, 1\}$, a kernel function $\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}')$ and a regularization parameter C (see below), the SVM finds a structurally optimal separating hyperplane in $\mathcal{F}$ by finding a vector $\hat{\boldsymbol{\alpha}}$:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^M \alpha_i - \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \alpha_i \alpha_j y_i y_j \mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{10}$$

subject to

$$\sum_{i=1}^M \alpha_i y_i = 0 \tag{11}$$

$$0 \le \alpha_i \le C \quad \forall i \tag{12}$$

The parameter $C > 0$ allows to specify how tightly the SVM is supposed to match the training data, with a larger C resulting in a tighter fit. Due to the restrictions on $\boldsymbol{\alpha}$ given in eq. (11), the solution of an SVM is sparse for most problems. The threshold b can be computed afterwards:

$$b = \frac{1}{|\{i : 0 < \hat{\alpha}_i < C\}|} \sum_{i \in \{i : 0 < \hat{\alpha}_i < C\}} (y_i - \sum_{j=1})^M y_j \hat{\alpha}_j \mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{13}$$

2.3. Probabilistic Output

Essentially, both KFD and SVM produce one-dimensional projections along a direction of greatest discrimination of the form

$$f(\boldsymbol{x}) = \sum_{i=1}^M \alpha_i \mathrm{k}(\boldsymbol{x}, \boldsymbol{x}_i) + b \tag{14}$$

In the case of the KFD the direction is determined directly, in the case of the SVM the direction is perpendicular to the separating hyperplane.

The problem now is that both KFD and SVM are non-probabilistic discriminators. However, for inclusion in an HMM based framework, *generative models* yielding class-conditional probabilities are needed.

2.3.1. The KFD case

State-conditional probabilities can be obtained if one assumes $p(\boldsymbol{x}|s_i) = p(f(\boldsymbol{x})|s_i)$. The one-dimensional projections $f(\boldsymbol{x})$ computed by the KFD exhibit strong Gaussianity [12] so that a single normal density function

$$p(f(\boldsymbol{x})|s_i) = \mathcal{N}(f(\boldsymbol{x})|\mu_i, \sigma_i) \tag{15}$$

can be fitted to each class' projection.

If a one-vs-rest setting is used for the binary classifiers, no further actions need to be taken. If, however, a one-vs-one binary setting is used, different estimates $p_{ij}(f(\boldsymbol{x})|s_i)$ of $p(f(\boldsymbol{x})|s_i)$ are computed. Since it is not clear how to combine these different estimates,

we resort to pairwise coupling using posterior probabilities (cf. section 2.3.3).

The posterior probability of state s_i given vector $\boldsymbol{x}$ can then be computed in each binary problem using Bayes' rule, marginalization and l'Hospital's rule:

$$P_{ij}(s_i|\boldsymbol{x}) = \frac{p(f(\boldsymbol{x})|s_i)P_{ij}(s_i)}{\sum_{c\in\{ij\}} p(f(\boldsymbol{x})|s_c)P_{ij}(s_c)} \tag{16}$$

2.3.2. The SVM case

The projections obtained by evaluating SVMs are not normally distributed. Instead, the projection densities between the margins, which is the area of greatest interest, are empirically found to be approximately exponential: $p(f(\boldsymbol{x})|s_i) = e^{a_i(f(\boldsymbol{x})-b_i)}$ [15]. Therefore, the following model can be applied to estimate posterior probabilities:

$$P_{ij}(s_i|\boldsymbol{x}) = \frac{1}{1+e^{A_{ij}f(\boldsymbol{x})+B_{ij}}} \tag{17}$$

The same equation is obtained if Gaussian projection-densities with equal covariance are assumed. The parameters A_{ij} and B_{ij} can be computed by minimizing a cross-error entropy function using a model-trust algorithm [15].

2.3.3. Pairwise Coupling

If the binary problems are considered to be one-vs-rest settings, the posterior probabilities from (16) or (17) can be used directly for posterior probabilities in the multi-class setting. If, however, the binary problems are one-vs-one settings (which is often done for computational tractability and prediction performance reason), the multi-class posterior probability must be computed from the individual binary posterior probability estimates. An efficient and accurate way of combining the individual estimates can be found in [16], which boils down to this formula:

$$P(s_i|\boldsymbol{x}) = \frac{1}{\sum_{j\neq i}\frac{1}{P_{ij}(s_i|\boldsymbol{x})} - S + 2} \tag{18}$$

State-conditional probabilities are then obtained using Bayes' rule:

$$P(\boldsymbol{x}|s_i) = \frac{P(s_i|\boldsymbol{x})P(\boldsymbol{x})}{P(s_i)} \propto \frac{P(s_i|\boldsymbol{x})}{P(s_i)} \tag{19}$$

The marginal probability $P(\boldsymbol{x})$ is not easy to determine. However, it can be neglected in (19) because it does not contribute to the maximization w. r. t. $\boldsymbol{w}$ in eq. (1).

3. EXPERIMENTS

3.1. General Setup and Baseline Results

Experiments were carried out on the DARPA Resource Management (RM1) task [1]. Training was carried out on the 72-speaker training set. The Oct89 set was used for development testing, model selection and parameter tuning. All systems were then evaluated on the Feb89 set using the previously determined optimal models.

The training set consists of 2880 sentences. Limited training data availability was simulated by randomly picking $1/2$, $1/4$, $1/8$, $1/16$ and $1/32$ of the complete training set. Each random sampling was repeated 10 times to get reliable results, resulting in 50 subsets of the training data. Each subset was used to train HMMs with GMMs as emission probabilities with up to 16 mixture components for monophone models and up to 8 mixture components for cross-word triphone models. For both modeling schemes the optimal number of mixture components was determined on the Oct89 set and then used for evaluation on the Feb89 set. Baseline results can be found in table 1.

HMM	Subset	min	avg	max	avg. # mixt.
Mono	Full		**93.8**		16
	1/2	91.45	**92.36**	92.78	15.1
	1/4	89.26	**90.46**	91.25	14.1
	1/8	84.26	**86.15**	87.62	7.6
	1/16	78.45	**80.34**	82.66	4.5
	1/32	70.40	**71.91**	74.00	2.9
XwrdTri	Full		**96.8**		8
	1/2	94.61	**95.19**	95.70	6.7
	1/4	91.92	**92.94**	93.71	6.3
	1/8	85.86	**88.23**	89.57	6.3
	1/16	78.64	**81.78**	83.83	4.2
	1/32	68.22	**70.77**	72.51	2.7

Table 1. Baseline results using the complete and randomly-sampled training set for monophone and triphone HMMs using GMMs as emission probabilities.

3.2. KFD and SVM

Both KFD and SVM based monophone models were evaluated for the full, the $1/8$ and the $1/32$ training data sets. In all cases a one-vs-one binary setting was chosen. For each of the 21 subsets a state-time alignment was produced using the GMM models optimal on the Oct89 set. Parameters for normalizing data to mean zero and variance one were estimated on the respective training subsets and applied to both training and test data. SVMs were trained using Torch [17]; for KFDs training, a high-performance BLAS-based training software was used. Kernel and regularization training parameters were optimized on the Oct89 set using only one subset for each subset size. Probabilistic interpretors from eq. (15) for KFDs and eq. (16) for SVMs were fitted using the same respective data sets as for the classifier training. Transition probabilities were re-used from the respective GMM trainings in order to keep parameter estimation effort low and to make results more comparable.

A modified version of HTK [18] utilizing a runtime plug-in library for external computation of emission probabilities was employed for recognition assessment. The library, which is common for KFD and SVM, uses a state-of-the-art XML-based storage scheme for its parameters which aids in editing and validating the data.

Full details of the recognition results using kernel-based acoustic models can be found in table 2, while a graphical comparison of

Model	Subset	min	avg	max
MonoKFD	Full		**94.21**	
	1/8	88.79	**89.41**	90.33
	1/32	76.65	**78.56**	80.48
MonoSVM	Full		**94.58**	
	1/8	89.38	**89.94**	90.98
	1/32	78.72	**81.13**	82.74

Table 2. Recognition results for monophone HMMs using kernel-based emission probabilities

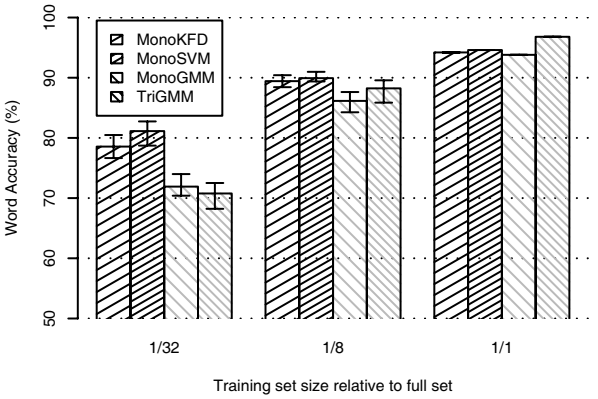

Fig. 1. Comparison of four different methods of acoustic modeling in relation to the amount of available training material

different methods can be found in figure 1.

From the results it becomes clear that both KFD and SVM outperform GMMs as emission probability estimators, independent of the amount of training material. For subset training it is also particularly interesting that the worst results obtained using either KFD or SVM are still better than the best results obtained using GMMs. Also, when training data is limited, monophone HMMs with KFD- or SVM-based emission probabilities surpass triphone HMMs which are otherwise superior to monophone HMMs in almost all cases.

The good generalization performance of kernel-based emission probabilities becomes even more evident when the relative gain in Word Error Rate is evaluated. Table 3 summarizes the results.

Subset size	MonoKFD	MonoSVM
1/1	6.5%	12.9%
1/8	23.5%	27.4%
1/32	23.7%	32.8%

Table 3. Average relative gains in Word Error Rate using kernel methods compared to GMMs

4. CONCLUSION

This paper showed that probabilistically interpreted Kernel Fisher Discriminants and Support Vector Machines can be used for modeling emission probabilities in HMM-based speech decoders. For the same HMM scheme they outperform Gaussian Mixture Models in all cases. The good generalization performance of the KFD and SVM leads to greatly improved recognition performance especially if only limited training material is available.

Future work will concentrate on evaluating KFD and SVM as emission probabilities for triphone HMMs, for larger problems and under different mismatch conditions, e. g. noisy or otherwise deteriorated acoustic conditions.

5. REFERENCES

[1] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word Resource Management database for continuous speech recognition," in *Proc. ICASSP*. IEEE, 1988, vol. 1, pp. 651–654.

[2] P. R. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proc. ICASSP*. IEEE, 1999.

[3] A. Ganapathiraju, *Support Vector Machines For Speech Recognition*, Ph.D. thesis, Mississippi State University, 2001.

[4] J. Stadermann and G. Rigoll, "A hybrid svm/hmm acoustic modeling approach to automatic speech recognition," in *Proc. Interspeech*. ISCA, 2004, pp. 661–664.

[5] S. E. Golowich and D. X. Sun, "A support vector/hidden markov model approach to phoneme recognition," in *ASA Proceedings of the Statistical Computing Section*, 1998, pp. 125–130.

[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[7] J. Weston, A. Gammerman, M. O. Stitson, V. Vapnik, V. Vovk, and C. Watkins, "Support vector density estimation," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., chapter 18, pp. 293–305. MIT Press, Cambridge, MA, USA, 1999.

[8] M. Schafföner, E. Andelic, M. Katz, S. E. Krüger, and A. Wendemuth, "Kernel fisher discriminants as acoustic models in hmm-based speech recognition," in *10th International Conference on Speech and Computer*, G. Kokkinakis, Ed., 2005.

[9] S. E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth, "Speech recognition with support vector machines in a hybrid system," in *Proc. EuroSpeech*, 2005, pp. 993–996.

[10] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[11] R. A. Fisher, "The use of multiple measurements in taxonomic problems," in *Annals of Eugenics*, vol. 7, pp. 179–188. Cambridge University Press, 1936.

[12] S. Mika, *Kernel Fisher Discriminants*, Ph.D. thesis, Technische Universität Berlin, Berlin, dec 2002.

[13] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Information Science and Statistics. Springer, Berlin, 2nd edition, 2000.

[14] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999.

[15] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large-Margin Classifiers*, P. Bartlett, B. Schölkopf, D. Schuurmans, and A. Smola, Eds., pp. 61–74. MIT Press, Cambridge, MA, USA, oct 2000.

[16] D. Price, S. Knerr, L. Personnaz, and G. Dreyfus, "Pairwise neural network classifiers with probabilistic outputs," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. Touretzky, and T. Leen, Eds., pp. 1109–1116. MIT Press, Cambridge, MA, USA, 7 1995.

[17] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *The Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[18] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.