# A REGULARIZED KERNEL-BASED APPROACH TO UNSUPERVISED AUDIO SEGMENTATION

*Zaïd Harchaoui*[1], *Félicien Vallet*[1,2], *Alexandre Lung-Yut-Fong*[1], *Olivier Cappé*[1]

[1]LTCI, TELECOM ParisTech & CNRS
46 rue Barrault
75634 Paris cedex 13, France

[2]Institut National de l'Audiovisuel
4 avenue de l'Europe
94366 Bry-sur-Marne cedex, France

**ABSTRACT**

We introduce a regularized kernel-based rule for unsupervised change detection based on a simpler version of the recently proposed kernel Fisher discriminant ratio. Compared to other kernel-based change detectors found in the literature, the proposed test statistic is easier to compute and has a known asymptotic distribution which can effectively be used to set the false alarm rate *a priori*. This technique is applied for segmenting tracks from TV shows, both for segmentation into semantically homogeneous sections (applause, movie, music, etc.) and for speaker diarization within the speech sections. On these tasks, the proposed approach outperforms other kernel-based tests and is competitive with a standard HMM-based supervised alternative.

***Index Terms***— Change detection, kernel methods, audio segmentation.

## 1. INTRODUCTION

Monitoring changes in a continuously observed signal is one of the fundamental problems of signal processing. In the framework of audio processing, for instance, the segmentation task consists in partitioning an audio stream into acoustically homogeneous segments. Segmentation is an important preliminary task for archiving audio or audiovisual content in databases while allowing for content-based retrieval of the data [1]. In particular, a large body of work in the field has focused on the identification of speech segments in radio broadcast [2, 3]. A related application is speaker change detection, also called speaker diarization, which is generally handled through parametric modelling based on Gaussian Mixture Models (GMMs) [4].

In this contribution, we clearly distinguish segmentation from classification and focus on methods that can solve the former problem in an unsupervised way, that is, with minimal knowledge of the characteristics of the underlying segments. Unsupervised segmentation is attractive because it may handle a variety of different situations without requiring prior training on each of these. In the experiments described in Section 5 below for instance, we will consider segmenting soundtracks of TV shows that have been annotated in terms of the following classes: applause, music, movie, speech. In such a situation, trying to simultaneously solve the unsupervised segmentation and classification tasks would be hard. The unsupervised approach is also potentially useful for applications involving multimodal data, such as segmentation of video streams based on joint audio and visual contents.

To provide the required flexibility to our unsupervised segmentation method, we propose a kernel-based test statistic for change detection. Kernel-based approaches for change detection have been introduced for audio processing by [5, 6] (see also [7]). Another related approach, coined the Maximum Mean Discrepancy (MMD), was also proposed by [8] for two-sample testing problems on microarray data. In contrast however to both MMD and KCD, we consider a test statistic based on the regularized kernel Fisher discriminant ratio, introduced in [9], yet with a novel and simpler regularization strategy, which is properly normalized and hence has a predictable false alarm rate. Note that, as in the other previously mentioned methods, we consider here a sliding window approach where the potential occurrence of a *single* change-point within the window is tested. We do not adopt here the multiple change-point estimation framework investigated in [10] both because of its computational complexity (which is quadratic in the frame size) and of the difficulty of assessing the statistical significance of the detected change-points within this approach.

In Section 2, we introduce relevant aspects of reproducing kernel Hilbert spaces (RKHS) theory so as to describe the proposed test statistic. Then, in Section 3, we consider the computation of the test statistic using the so-called kernel-trick as well as that of the associated decision threshold for a specified false-alarm rate. Section 4 is devoted to a discussion of the main features of the proposed approach together with a comparison with the KCD and MMD algorithms. Finally, in Section 5, we provide experimental results both on segmentation of whole audio tracks from TV shows and on speaker diarization within the interview segments.

## 2. REGULARIZED KERNEL FISHER DISCRIMINANT

Temporal segmentation is the problem of partitioning a sequence of vector-valued observations $(X_t)_{t=1,\ldots,n}$ into, say, $K$ segments, delimited by $K+1$ boundaries, as

$$[X_1, \ldots, X_n] = \bigcup_{k=0,\ldots,K-1} [X_{\tau_k+1}, \ldots, X_{\tau_{k+1}}] \,, \qquad (1)$$

where each segment $[X_{\tau_k+1}, X_{\tau_{k+1}}]$, with boundaries $[\tau_k + 1, \ldots, \tau_{k+1}]$, is considered homogeneous in distribution (and $\tau_0 = 0, \tau_K = n$).

This research was supported by the European Comission under contract *FP6-027026-K-SPACE* and by the Agence Nationale de la Recherche project *ANR-06-BLAN-0078-03 KERNSIG*.

### 2.1. Test Statistic

Following the principle at the core of several recent development in machine learning [11], the observations are mapped in an abstract space, namely a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a reproducing kernel $k(\cdot,\cdot)$ and a feature map $\Phi(X) = k(X,\cdot)$.

We now describe how the kernel Fisher discriminant ratio, put forth in [9] for testing the homogeneity of two samples, may be further simplified to yield an efficient kernel-based change detection algorithm. This is described in the operator-theoretic framework, developed for the statistical analysis of kernel-based learning algorithms in [12, 9].

Consider two samples of independent observations

$$X_1^{(1)}, \ldots, X_{n_1}^{(1)} \sim \mathbb{P}^{(1)} \qquad \text{and} \qquad X_1^{(2)}, \ldots, X_{n_2}^{(2)} \sim \mathbb{P}^{(2)} \; .$$

corresponding respectively to the first part and the second part of the sliding window. The goal is to design a test statistic to decide between

$$\mathbf{H}_0: \quad \mathbb{P}^{(1)} = \mathbb{P}^{(1)} \qquad \text{and} \qquad \mathbf{H}_A: \quad \mathbb{P}^{(1)} \neq \mathbb{P}^{(2)} \; .$$

Define the corresponding empirical mean elements and covariance operators associated to $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ as follows

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{\ell=1}^{n_1} k(X_\ell, \cdot) \; ,$$
$$\hat{\Sigma}_1 = \frac{1}{n_1} \sum_{\ell=1}^{n_1} \{k(X_\ell,\cdot) - \hat{\mu}_1\} \otimes \{k(X_\ell,\cdot) - \hat{\mu}_1\} \; ,$$

where $u \otimes v$ for $u, v \in \mathcal{H}$ is defined for all $f \in \mathcal{H}$ as $(u \otimes v) f = \langle v, f \rangle_{\mathcal{H}} \, u$ The quantities $\hat{\mu}_2$ and $\hat{\Sigma}_2$ are similarly defined for the second sample $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ with obvious changes.

Their population counterparts, *i.e.* the population mean element and the population covariance operator, are defined for any probability measure $\mathbb{P}$ as $\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}} = \mathbb{E}[f(X)]$ for all $f \in \mathcal{H}$, and $\langle f, \Sigma_{\mathbb{P}} g \rangle_{\mathcal{H}} = \mathrm{Cov}_{\mathbb{P}}[f(X), g(X)]$ for $f, g \in \mathcal{H}$. The (maximum) *regularized kernel Fisher discriminant ratio*, which we abbreviate as KFDR is defined as

$$\mathsf{KFDR}(X_1^{(1)}, \ldots, X_{n_1}^{(1)}; X_1^{(2)}, \ldots, X_{n_2}^{(2)}) = \frac{n_1 n_2}{n_1 + n_2} \left\langle \hat{\mu}_2 - \hat{\mu}_1, \mathcal{I}(\hat{\Sigma}_W; \gamma)(\hat{\mu}_2 - \hat{\mu}_1) \right\rangle_{\mathcal{H}} \qquad (2)$$

where

$$\hat{\Sigma}_W = \frac{n_1}{n_1 + n_2} \hat{\Sigma}_1 + \frac{n_2}{n_1 + n_2} \hat{\Sigma}_2$$

and $\mathcal{I}(\hat{\Sigma}_W; \gamma)$ denotes a regularized version of the inverse within-class covariance operator, with regularization parameter $\gamma$.

### 2.2. Spectral Truncation Regularization

Note that in contrast to the usual Fisher test on vector-valued data, the kernel version of the test requires that $\hat{\Sigma}_W$ be regularized before inversion due to the infinite-dimensional nature of the RKHS $\mathcal{H}$. In [9], the authors propose to take

$$\mathcal{I}(\hat{\Sigma}_W; \gamma) = (\hat{\Sigma}_W + \gamma \mathrm{I})^{-1} \; , \qquad (3)$$

which however gives rise to a complicated limiting distribution under the null hypothesis (an infinite weighted sum of $\chi^2$ random variables). Denote $(\lambda_p, e_p)_{p=1,\ldots,\infty}$ respectively the sequence of eigenvalues and eigenfunctions of the operator $\hat{\Sigma}_W$. We propose to use instead a spectral truncation type of regularization:

$$\mathcal{I}(\hat{\Sigma}_W; 1/d) = \sum_{p=1}^{d} \lambda_p^{-1} (e_p \otimes e_p) \; , \qquad (4)$$

which corresponds to the inverse of a finite-dimensional approximation on $\mathrm{Span}\{e_1, \ldots, e_d\}$ of $\hat{\Sigma}_W$ (see also [13]). As discussed below in Section 3, the practical computation of the test statistics boils down to performing a principal component analysis of a correctly centered version of the Gram matrix.

### 2.3. Large-Sample Distribution Under The Null Hypothesis

In order to set the decision threshold for a prescribed false-alarm rate $\alpha$, we may use the limiting null distribution, that is the distribution as $n_1, n_2 \to \infty$, of the test statistic under the null hypothesis $\mathbf{H}_0$. Using similar arguments as in [9], we may prove the following result.

**Proposition 1** *Assume $\boldsymbol{H}_0$ holds. Under suitable assumptions (cf. [9, Theorem 1])), we have as $n_1 + n_2 \to \infty$*

$$\mathit{KFDR}_d(X_1^{(1)}, \ldots, X_{n_1}^{(1)}; X_1^{(2)}, \ldots, X_{n_2}^{(2)}) \xrightarrow{\mathcal{D}} \chi_d^2 \; .$$

Here $\chi_d^2$ denotes a $\chi^2$ random variables with $d$ degrees of freedom. Therefore, thanks to our spectral truncation regularization scheme, for any *degrees of freedom* $d > 1$ of our test statistic, and for any prescribed false-alarm rate $0.0 < \alpha < 0.5$, we may readily compute the associated decision threshold $c(1-\alpha; d)$, such that

$$\mathbb{P}_{\mathbf{H}_0}(\mathsf{KFDR}_d > c(1-\alpha; d)) = \alpha \; .$$

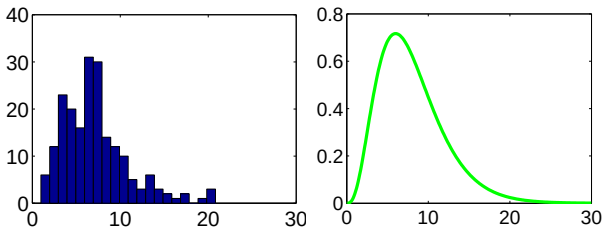


**Fig. 1**. Comparison of the finite-sample distribution of the test statistic against its large-sample distribution, based on 200 homogeneous windows of length 128 of the data considered in Section 5.1.

As Figure 1 shows, the large-sample distribution provides a reasonably accurate picture of the finite-sample behavior of our test statistic under the null hypothesis.

In the next section, we provide details on the efficient computation of the test statistic for all $d > 1$, using a bi-centered KPCA, and the computation of the corresponding decision thresholds.

## 3. COMPUTATION OF THE TEST STATISTIC

We first describe how to compute $\mathsf{KFDR}_d$ using the kernel trick, and a suitable variant of KPCA [14].

Let $\mathbf{K} = [K(X_i, X_j)]_{ij=1,\ldots,n}$ be the Gram matrix associated to the pooled sample $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}; X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$. Define the projection matrix $\mathbf{P}_1 = \mathrm{I}_{n_1} - n_1^{-1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T$ where $\mathbf{1}_{n_1}$ is the $(n_1 \times 1)$ vector whose components are all equal to one and $\mathrm{I}_{n_1}$ is the $(n_1 \times n_1)$ identity matrix. Define similarly $\mathbf{P}_2 = \mathrm{I}_{n_2} - n_2^{-1}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T$. Introduce the "bi-centering" matrix $\mathbf{P}_{\text{bi}}$ defined as

$$\mathbf{P}_{\text{bi}} = \begin{pmatrix} \mathbf{P}_1 & 0 \\ 0 & \mathbf{P}_2 \end{pmatrix}. \tag{5}$$

Denote $\tilde{\mathbf{K}} = \mathbf{P}_{\text{bi}}\mathbf{K}\mathbf{P}_{\text{bi}}$ the bi-centered Gram matrix. Denote by $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)^T$ the matrix representations of the first $d$ eigenvalues and $d$ eigenvectors of the normalized bi-centered Gram matrix $\mathbf{N} = \frac{1}{n_1+n_2-1}\tilde{\mathbf{K}}$. $\mathbf{\Lambda}$ and $\mathbf{V}$ may be interpreted as the empirical estimates of the first $d$ eigenvalues and $d$ eigenvectors of the pooled covariance estimate. Now, define the vector $\mathbf{m}_n = (\mathbf{m}_{n,i})_{1\leq i\leq n}$ with $\mathbf{m}_{n,i} = -n_1^{-1}$ for $i = 1, \ldots, n_1$ and $\mathbf{m}_{n,i} = n_2^{-1}$ for $i = n_1 + 1, \ldots, n_1 + n_2$. Then,

$$\mathsf{KFDR}_d = \mathbf{m}\mathbf{K}\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{K}^T\mathbf{m}^T \ .$$

It is worthwhile to emphasize that, to compute $\mathsf{KFDR}_d$ for a large range of values $d = 1, \ldots, N$ of the regularization parameter $d$, it suffices to compute once for all the eigendecomposition of $\mathbf{N}$ up to $N$, and then compute all $\{\mathsf{KFDR}_d\}_{d=1,\ldots,N}$ recursively using rank-one updates. In order to obtain a decision rule with specified false-alarm rate $\alpha$ for large samples, it suffices to choose $c(1-\alpha; d)$ as the $(1-\alpha)$-quantile of the limiting $\chi^2$-type distribution under $\mathbf{H}_0$.

## 4. DISCUSSION

### 4.1. Window Size

The choice of the size of the window on which the test statistic is to be computed should take into account two important constraints. First, if $n$ is the sliding-window length, the computational complexity of KFDR is $O(n^3)$ in time (eigenvalue problem) and $O(n^2)$ in space. Hence, for $n > 5000$, the computational burden is substantial, and one resorts to computing averages of KFDR of subsamples of size $m \ll 5000$ instead of the raw KFDR on the whole sample of size $n$. In the audio segmentation application, we may reduce the computational burden elegantly by working on medium samples coarse-scale statistical summaries of our features instead of working on large samples with fine-scale features. We describe in detail these "statistical summaries" later in Section 5.

Second, reducing change-detection on the whole signal to a sequence of single change-point detection on windows is valid only if we are guaranteed that there is *at most one* change-point within each window. Again, physical considerations allow to alleviate this issue by setting the window-length to a sufficiently small length for being guaranteed that no more than a single change-point occurs within the window, and sufficiently large length for our decision rule to be statistically significant (typically $n > 50$).

### 4.2. Choosing The Regularization Parameter

The spectral truncation regularization parameter $d$ plays a role similar to that of $1/\gamma$ in the Tikhonov-type version (3) of KFDR. Indeed, in order to effectively detect differences in distribution between two samples, one has to set $d$ (resp. $\gamma$) to a sufficiently large value (resp. small value) to capture the first component $p$ in the eigen-expansion $\Sigma_W$ on which $\langle \mu_2 - \mu_1, e_p \rangle_{\mathcal{H}}$ is significantly large. However, as the degrees-of-freedom $d$ gets higher than this optimal value, then the power slowly decreases. We recommend the following simple strategy: take $N = \max\{p, \lambda_p > \epsilon\}$, where $\epsilon = 10^{-10}$, and then set $d = N/2$.

| | Nb. of sections | Ave. duration (sec.) |
|---|---|---|
| applause | 84 | 14 |
| movie | 29 | 155 |
| music | 38 | 194 |
| speech | 188 | 70 |
| speaker turns | 962 | 6 |

**Table 1**. Characteristics of the audio data.

### 4.3. Comparison With Related Approaches

When compared to KFDR, the Maximum Mean Discrepancy (MMD) test statistic of [8] is based on

$$\mathsf{MMD} = \frac{n_1 n_2}{n_1 + n_2} \langle \hat{\mu}_2 - \hat{\mu}_1, \hat{\mu}_2 - \hat{\mu}_1 \rangle_{\mathcal{H}} \ ,$$

instead of (2) in KFDR. The absence of renormalization by $[\mathcal{I}(\hat{\Sigma}_W; \gamma)]^{-1}$, the inverse of the regularized within-class covariance operator $\hat{\Sigma}_W$, results in a substantial loss of power against differences in distribution supported by higher components of $\Sigma_W$, as described both theoretically and experimentally in [15]. This lack of normalization also makes the *a priori* selection of a threshold with desired false alarm rate difficult.

The Kernel Change Detection (KCD) algorithm of [5], consists in, on the one hand, computing a $\nu$ 1-class SVM on the first sample $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ yielding a dual vector $\alpha_1$, and on the other hand, a $\nu$ one-class SVM on the second sample $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ yielding a dual vector $\alpha_2$. Then, the "test statistic", although not advertised as a test statistic but rather as a similarity measure, is defined as

$$\mathsf{KCD} = \frac{\alpha_1^T \mathbf{K}_{12} \alpha_2^T}{\sqrt{\alpha_1^T \mathbf{K}_{11} \alpha_1^T}\sqrt{\alpha_2^T \mathbf{K}_{22} \alpha_2^T}} \ .$$

When $\nu = 1$ for both 1-class SVMs, then KCD simply writes as

$$\mathsf{KCD} = \frac{\langle \hat{\mu}_1, \hat{\mu}_2 \rangle_{\mathcal{H}}}{\sqrt{\langle \hat{\mu}_1, \hat{\mu}_1 \rangle_{\mathcal{H}}}\sqrt{\langle \hat{\mu}_2, \hat{\mu}_2 \rangle_{\mathcal{H}}}} \ .$$

Hence, KCD may be related to variant of MMD, where empirical mean elements $\hat{\mu}_1$ and $\hat{\mu}_2$ are replaced by sparse weighted means $\tilde{\mu}_1 = n_1^{-1}\sum_{\ell=1}^{n_1} \alpha_{1,\ell} k(X_i, \cdot)$ and $\tilde{\mu}_2 = n_2^{-1}\sum_{\ell=1}^{n_2} \alpha_{2,\ell} k(X_i, \cdot)$.

## 5. EXPERIMENTAL RESULTS

Our dataset consists of the two soundtracks of the popular French 1980s entertainment TV-shows ("Le Grand Échiquier") of roughly three hours each, whose main features are gathered in Table 1. Audio tracks were extracted from MPEG video files, converted to mono by averaging right and left channels, and downsampled to 16 kHz. Then, we extracted every 10 ms the first 12 Mel Frequency Cepstral Coefficients (MFCC), as well as the 0-th order cepstral coefficient, giving 13 features in total.

### 5.1. Semantic segmentation

As is obvious from Table 1, the "semantic" changes that are to be detected in this first setting are diverse with some of them (movie,

| | Semantic seg. | | Speaker seg. | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| KFDR | 0.72 | 0.63 | 0.89 | 0.90 |
| MMD | 0.71 | 0.58 | 0.76 | 0.73 |
| KCD | 0.65 | 0.63 | 0.78 | 0.74 |
| HMM | 0.73 | 0.65 | 0.93 | 0.96 |

**Table 2**. Best Precision and Recall for all benchmarked method, for both semantic segmentation and speaker segmentation tasks.

music) corresponding to long segments. In order to detect only those high-level changes, we worked on subsampled statistical summaries of the above features instead of the raw features. This considerably lightened the computational burden while significantly enhancing the statistical power of our approach. Indeed, we computed pairs of *slope* and *intercept* obtained by a linear regression over (non-overlapping) windows of length 33. Then, we chose as a kernel a weighted linear combination of Gaussian rbf kernels on the 0-th order cepstral coefficient ($k_{c0}(\cdot,\cdot)$) and on the first 12 MFCCs ($k_{c1-c12}(\cdot,\cdot)$). Each specific kernel was actually a linear combination of a Gaussian RBF kernel on the slope and a kernel on the intercept. The kernel bandwidths were optimized globally over a fixed grid (hence, the same kernel bandwidth was used for all experiments), and the spectral truncation regularization parameter was set as described in Section 4. The test statistic was computed in sliding windows of length 128 overlapping by 20% and the performance was evaluated in terms of the number of windows with detection effectively containing at least one annotated audio change. We also provide a comparison with a supervised Hidden Markov Model (HMM) technique, where each state corresponds to a mixture of diagonal-covariance Gaussians with 16 components which is trained beforehand from the first excerpt of each type of segment. The detection threshold for each approach was chosen so as to maximize the precision/recall $F$-measure. Finally for KCD, we used $\nu = 0.5$ for the one-class SVM optimization. The experimental results are displayed in Table 5.2, in terms of precision and recall. Note that, while our high-level features are not accurate enough to detect very small segments, still, we outperform competing approaches in average in terms precision and recall.

### 5.2. Speaker segmentation

As before for semantic segmentation, we worked on "statistical summaries" of the above features instead of the raw features. Here, we computed pairs of *slope* and *intercept* obtained by a linear regression over (non-overlapping) windows of length 15. The same strategy was adopted for combining the different kernels and for setting the spectral truncation regularization parameter. The test statistic was computed in sliding windows of length 64 overlapping by 20%. We also provide comparison with Hidden Markov Models (HMM), trained similarly. The experimental results are displayed in Table 5.2. Note that HMM are much difficult to beat on this task. However, one has to keep in mind that the training procedure used here is rather unrealistic. We explicitly modelled beforehand all speakers involved in the speech sections. Therefore, our results are rather promising, since we obtained competitive performance results with a completely unsupervised approach.

## 6. CONCLUSION

We proposed an efficient regularized kernel-based test statistic for temporal segmentation of audio tracks. Thanks to a novel regularization strategy, our test statistic has a simple limiting distribution under the null hypothesis, which allows a straightforward calibration from a prescribed false-alarm rate $\alpha$. Our method outperforms previously proposed kernel-based approaches, and reach competitive performance when compared to "supervised" methods on the same tasks.

## 7. REFERENCES


[1] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9(4), pp. 441–457, May 2001.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE ICASSP*, 1996, pp. 993–996.

[3] E. Schreier and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," 1997, pp. 1331–1334, IEEE.

[4] P. Delacourt P. and C.J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32(16), pp. 111–126, September 2000.

[5] F. Désobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Trans. on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, August 2005.

[6] B. Fergani, M. Davy, and A. Houacine, "Unsupervised speaker indexing using one-class support vector machines," in *EUSIPCO*, 2006.

[7] P.-C. Lin, J.-C. Wang, J.-F. Wang, and H.-C. Sung, "Unsupervised speaker change detection using svm training misclassification rate," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1234–1244, 2007.

[8] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A.J. Smola, "A kernel method for the two-sample problem," in *Adv. NIPS*, 2006.

[9] Z. Harchaoui, F. Bach, and E. Moulines, "Testing for homogeneity with kernel fisher discriminant analysis," in *Adv. NIPS*, 2007.

[10] Z. Harchaoui and O. Cappé, "Retrospective change-point estimation with kernels," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2007.

[11] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Camb. UP, 2004.

[12] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Machine Learning*, vol. 66, 2007.

[13] G. Blanchard and L. Zwald, "Finite-dimensional projection for classification and statistical learning," *IEEE Transaction on Information Theory*, vol. 54, no. 9, pp. 4169–4182, 2008.

[14] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[15] Z. Harchaoui, F. Bach, and E. Moulines, *Testing for Homogeneity with Kernel Fisher Discriminant Analysis*, 2008, arXiv:0804.1026v1.