COMPARISON OF DIFFERENT STRATEGIES FOR A SVM-BASED AUDIO SEGMENTATION

Mathieu Ramona ‡, Gaël Richard †

‡ RTL (Ediradio)
22 rue Bayard, 75008 Paris
phone: +33 (0) 1 56 69 42 26
mathieu.ramona@rtl.fr

† Telecom ParisTech / LTCI-CNRS
37 rue Dareau, 75014 Paris
phone: +33 (0) 1 45 81 73 65
gael.richard@telecom-paristech.fr

ABSTRACT

We compare in this paper diverse hierarchical and multi-class approaches for the speech/music segmentation task, based on Support Vector Machines, combined with a median filter post-processing. We show the efficiency of kernel tuning through the novel Kernel Target Alignment criterion. Quantitative results provide an F-measure of 96.9%, that represents an error reduction of about 50% compared to the results gathered by the French ESTER evaluation campaign. We also show the relevance of the SVM with very low feature vector dimension on this task.

1. INTRODUCTION

Audio segmentation is now becoming a major component for numerous audio applications such as broadcast audio streams transcription [13, 14], music signals structuring/summarization, musical instrument recognition or audio coding. For instance, speech/music segmentation is particularly interesting in the context of broadcast audio streams transcription, where it appears indeed mandatory to first obtain a precise segmentation into homogeneous segments before applying to those segments a specific processing (e.g. speaker identification for speech segments, audio identification for music segments,...).

A number of approaches were already proposed in the past and some of them were extensively tested in a national collaborative evaluation campaign of broadcast transcription systems called ESTER [3]. Some of such systems are based on ergodic Hidden Markov Models with Gaussian Mixtures modeled observations [5]. Some straightforward applications of GMM are also found on very diverse sets of audio features [13] but nevertheless, Support Vector Machines (SVM) remain rarely used for this task despite their good discrimination power [7, 8]. Indeed, SVM originally impose a constraint of two-class discrimination but various solutions have been proposed to extend the application scope of SVM to multiclass cases.

In this paper, we propose an extension of a previous system [12] by first comparing different hierarchical or multiclass strategies and assessing their robustness to the feature vector dimension reduction. In particular, we evaluate the interest of considering multiple music sub-classes and its impact on the overall performance.

This work was partly realized as part of the Quaero Program, funded by OSEO, French State agency for innovation.

We therefore consider 4 different learning classes in this paper: speech only (called `speech` below), music only (`music`), speech with musical background (`mix`) and music with singing voice (`singing`). Second, we validate on a large database the Kernel Target Alignment criterion recently introduced [2] that allows to fine tune the SVM parameters at a very low complexity compared to the classical grid search. The different results obtained are compared, whenever possible, to the results of the best systems submitted to the national evaluation campaign ESTER [3] using the exact same protocol and databases. It is shown that the best taxonomy brings a systematic gain compared to the best ESTER system (up to a +2.7% absolute gain for a feature vector dimension of 50).

The article is structured as follows. The global architecture, as well as the essential details of our system are presented in section 2. The experimental protocol and databases used are then exposed in section 3. Quantitative results are presented and commented in section 4 and some discussion and perspectives will be given in section 5.

2. CLASSIFICATION SCHEME

2.1 General architecture

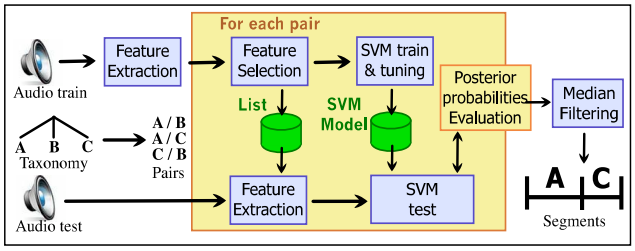

Figure 1: Architecture of the proposed system

The system presented here combines SVM discriminations by comparing multiclass approaches to various hierarchical approaches. We follow here the classical scheme in statistical learning based on the classification of acoustic feature vectors on a collection of overlapping frames covering the audio signal. Each frame is associated with one of the 3 classes `speech`/`music`/`mix`. The presence of singing voice on some of the music segments of the corpus might induce a confusion with the `mix` class, considering that singing is somehow sim-

ilar to speech. We have thus introduced a forth class (`singing`), used or not, depending on the taxonomy adopted, during the learning phase, and assimilated to the `music` class for the decision. The posterior probability sequence is further smoothed by a median filtering, preceding the class decision and the gathering of labels within homogeneous temporal segments.

2.2 Pairs discrimination

2.2.1 Classification taxonomies

Since the classification method employed here is fundamentally discriminative, different strategies (taxonomies), presented in figure 2, are proposed in this study to extend the scope of SVM to more than 2 classes. Binary tree nodes consist in successive discrimination (a/b being the class union of a and b). Non-binary tree nodes are based on the combination of the discriminations of all pairs of classes involved (the combination scheme is further presented in section 2.3).

We propose here 2 main schemes A and B with variants according to the use of the `singing` class : *nosing* when the class is not used for training, *mixsing* when it is first grouped with the `mix` class, and *musing* when it is grouped with the `music` class.

Thus, for exemple in A^{nosing} the `singing` class is not included in the learning process, whereas in $A^{mixsing}$ it is first classified with the `mix` class and then seperated to be identified as `music` in the end. The A scheme consists basically in a multiclass level based on pair combinations, whereas the B scheme consists in a hierarchical tree seperating pure `music` from `speech` (with an eventual music background), and then separating pure `speech` from `mix`. Scheme C is a less intuitive variant of the hierarchical approach and scheme D adds the `singing` class to the multiclass classification of scheme A. Please note that the taxonomy selection presented here is obviously not exhaustive.

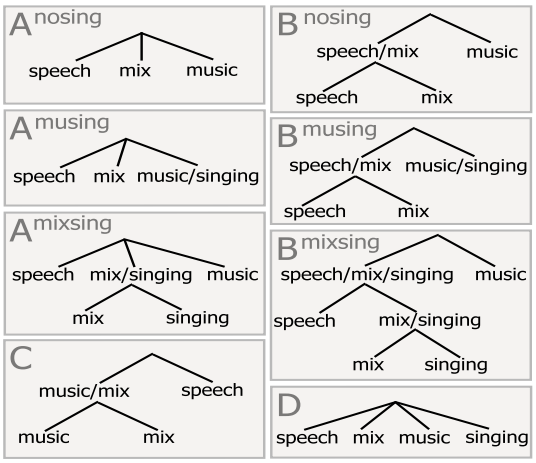

Figure 2: Taxonomies selected for our classification in 4 classes. Two labels for a single node or leaf indicate that the two classes were grouped in the learning phase (e.g. music/singing)

Obviously, a large number of pairs are common to several trees. These taxonomies gather a total of 14 distincts pairs (that can imply unions of classes) ; for each of them a SVM classifier is tuned and trained.

2.2.2 Acoustic features

Our system uses a large collection of about 600 features of various types (temporal, spectral, cepstral and perceptual[1]). Most of these features are computed on short-term frames of 32ms (16ms overlap) and some of them on large-term frames of 1s (0.5s overlap). The short-term features are replaced by their mean and variance over each long-term frame and then associated with the long-term features in a common feature vector. Each feature is mean-centered and normalized by its standard deviation evaluated on the whole training set.

Then for each class pair the most relevant features are selected by using the IRMFSP algorithm, introduced in [9]. Each classification taxonomy has been evalated with a common feature vector dimension for all pairs, varying between 2 and 50.

2.2.3 Discrimination with Support Vector Machines

The classification scheme used in this study is based on Support Vector Machines, which apply a non-linear transform (through a kernel function κ, which is the inner product on two transformed vectors) on d dimension vectors into a higher dimension space where the two classes are linearly separated under the margin maximization constraint. The kernels used here are Radial Basis Gaussian kernels :

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{d\sigma^2}\right)$$

Empirical studies have not shown any significant advantage with other classical kernels (polynomial, sigmoidal...).

The σ parameter must be tuned carefully to optimize the kernel by reaching the best compromise between efficiency and generalization. We have used here a recent method, called Kernel Target Alignment (section 2.2.4), to automatically determine the optimal value for any kernel parameter without any previous SVM learning and thus no grid-search or need for a validation set.

The output value of the decision function is unbounded, though, and thus does not represent a probabilistic value. The sigmoid bijection proposed by Platt in [10] is used here to get a probabilistic output from SVM.

2.2.4 Kernel Target Alignment

In [2], Cristianini *et al.* describe a new criterion to estimate the efficiency of a kernel, that is only based on its Gram matrix $\boldsymbol{K}$ computed from a set of n examples $\boldsymbol{x} = [x_i]$. Defining the ideal target matrix as $\boldsymbol{K^*} = \boldsymbol{y}\boldsymbol{y^t}$, where $\boldsymbol{y} = [y_i] \in \{1; -1\}^n$ represents the class labels corresponding to the examples x_i, Cristianini *et al.* propose to quantify the relevance of $\boldsymbol{K}$ as follows, thus defining

[1] For an extensive list, please consult [12] ; the estimated pitch f_0 and aperiodicity measure estimated with *YIN* [1] have been added as well.

the Kernel Target Alignment (KTA) :

$$A(\boldsymbol{K}, \boldsymbol{K^*}) = \frac{\langle \boldsymbol{K}, \boldsymbol{K^*} \rangle_F}{||\boldsymbol{K}||_F ||\boldsymbol{K^*}||_F}$$

where $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \sum_i \sum_j a_{ij} b_{ij}$ is the Frobenius inner product of two matrices, and $||\boldsymbol{A}||_F$ the corresponding norm. This criterion constitutes an estimation of the similarity between the Gram matrix of a Kernel and the target matrix, and thus allows the tuning of the kernels parameters (such as σ for the Gaussian kernel), without any training of the SVM.

Pothin and Richard have later observed in [11] that the KTA of a kernel $\boldsymbol{K_\theta}$ can easily be differenciated with respect to the kernel parameter θ. Indeed, if we define the matrix $\partial_\theta \boldsymbol{K_\theta} = [\partial_\theta \kappa_\theta(x_i, x_j)]$, then

$$\partial_\theta \langle \boldsymbol{K_\theta}, \boldsymbol{K^*} \rangle_F = \langle \partial_\theta \boldsymbol{K_\theta}, \boldsymbol{K^*} \rangle_F$$

$$\partial_\theta ||\boldsymbol{K_\theta}||_F = \langle \partial_\theta \boldsymbol{K_\theta}, \boldsymbol{K_\theta} \rangle_F \, / \, ||\boldsymbol{K_\theta}||_F$$

The derivative of the alignment with respect to θ is then expressed as follows

$$\partial_\theta A(\boldsymbol{K_\theta}, \boldsymbol{K^*}) = \frac{\langle \partial_\theta \boldsymbol{K_\theta}, \boldsymbol{K^*} \rangle_F}{||\boldsymbol{K_\theta}||_F ||\boldsymbol{K^*}||_F} - \frac{\langle \boldsymbol{K_\theta}, \boldsymbol{K^*} \rangle_F \langle \boldsymbol{K_\theta}, \partial_\theta \boldsymbol{K_\theta} \rangle_F}{||\boldsymbol{K_\theta}||_F^3 ||\boldsymbol{K^*}||_F}$$

We have thus tuned the σ parameter of our kernels through a gradient ascent on the KTA. The latter has a Gaussian-like shape with respect to $\log(\sigma)$. We have thus fixed an initial value of 0.1 for σ since the steep is more pronounced from this side. The ascent stops when the gain of the KTA is inferior to a threshold $\epsilon = 10^{-3}$. The gradient step has been determined empirically and results in an average of 5 iterations to reach convergence.

Using this method, the σ parameter is tuned for each SVM prior to any training. This results in a very significant reduction of the overall training time. We will see in section 4.1 that the performances reached with this technique are comparable to the grid-search approach and even better for some of the taxonomies.

The *SVMlight*[2] package has been used in this experiment to train the Support Vector Machines.

2.3 Posterior probabilities

Two structure types are present in the classification taxonomies trees presented in figure 2 :

Multi-class decision : schemes A and D. In this case each class pair is discriminated by the corresponding SVM. The algorithm proposed by Hastie and Tibshirani in [4] is used to estimate the posterior probabilities of the classes, from the results of all pairs.

Hierarchical decision : schemes B and C. The nodes are processed sequentially, any father node being processed before its children. Each node discriminates one of the classes (source class) within two new classes. The posterior probabilities of the other classes are unchanged. Those of the two new classes are the result of the discrimination, weighted by the posterior probability of the source class. Thus the sum of the classes' probabilities remains equal to one.

[2] `http://svmlight.joachims.org/`

2.4 Smoothing by median filtering

In order to smoothen the posterior probabilities computed on the frame sequence, a median filtering is applied for each class. The window size has been empirically tuned to 9 frames, which approximately corresponds to a 5s window.

Then, for each frame, the class maximizing the posterior probability is elected. Homogeneous adjacent frames are gathered within temporal segments.

3. EXPERIMENT

3.1 ESTER Corpus

This experiment is based on the corpus of the evaluation campaign ESTER, complying the SES task of sound event segmentation. We have updated the whole learning set annotation by differentiating the music only and singing music segments. The ressources are divided between a 77h learning set and a 12h30 development set that respects the distribution of the different radios in the audio files. The test set annotations have been kept intact, in order to keep our results relevant with those of the evaluation campaign.

3.2 Scoring

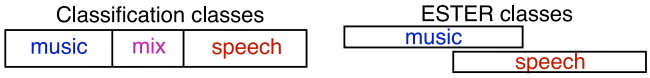

Figure 3: Conversion from non-overlapping classes to the ESTER classes

The scoring is done following the protocol of the ESTER campaign. The 3 non-overlapping classes considered for classification (`speech`, `mix` and `music`) are converted into 2, possibly overlapping, classes (`speech` and `music`), as shown in figure 3. On each of these classes, and on their union, the *Recall (R)* and *Precision (P)* are measured. These are defined as the ratio of the cumulated duration where the class is correctly detected on the cumulated duration where the class is, respectively, really present (Recall) and detected (Precision). The *F-measure (F)* is the harmonic mean between these two measures (i.e. $F = \frac{2RP}{R+P}$). The false alarm (*fa*) and false reject (*fr*) rates are also considered here

We use the `trackeval` tool, provided as part of the ESTER corpus, to compute these criterions.

4. RESULTS

4.1 Evaluation of the kernel tuning

Choosing an arbitrary dimension of $d = 20$, we have tuned each SVM with both σ values determined through a classical grid-search and the KTA maximization.

The grid-search is done within a set of 12 values locarithmically spread between 0.2 and 1.5 (indeed, thanks to the d normalization factor in the kernel denominator, the optimal σ is always in this interval). An SVM is trained on the learning set for each σ value ; the value maximizing the performances of the SVM on the validation set is picked.

The KTA maximisation is performed over 5000 frames from the learning set for each class, when available.

Figure 4 shows the global F-measure computed for each of the 8 taxonomies with both tuning approaches. The pink bars are for the *grid-search* and the light blue ones for *KTA*. It is clear here that the grid-search surpasses the KTA for determining the optimal σ value. Nevertheless, once the median post-processing is applied (red and dark blue bars), KTA provides a better result is most cases. This is explained by the fact that KTA has provided a better tuning for some of the discriminators. The median post-processing corrects accidental errors (on one or two adjacent frames) very efficiently and is thus able to cancel the slight disadvantage of KTA.

The KTA maximization, proposed here for this application, thus provides comparable performances without any validation set and with a unique SVM training. The computation of the KTA being based on the Gram matrix of the Kernel, is it a little less than quadratic in complexity (since the Gram matrix is symmetric) with regard to the number of samples, which is about the same than a SVM training phase with *SVMlight* [6].

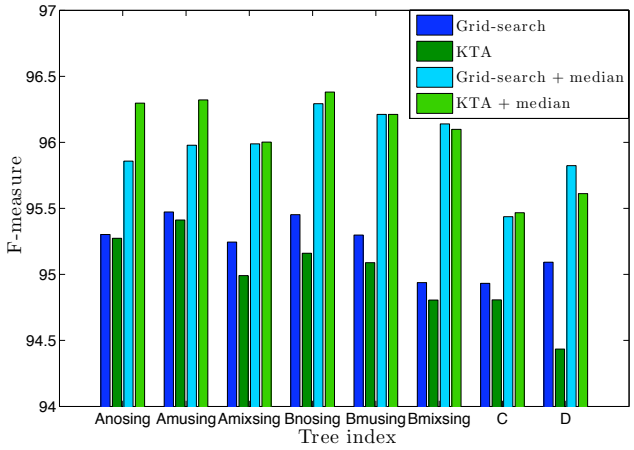

Figure 4: Comparison of the different taxonomies performances with grid-search and KTA kernel tuning.

4.2 Evaluation of the different taxonomies

Figure 5 shows the evolution of the F-measure, for each of the 8 taxonomies, with the dimension d of the feature vector.

As expected, the main schemes A and B frankly exceed the performances of the C and D variants, at almost any dimension. The C scheme is indeed counter-intuitive and the D scheme does not have enough training samples for the class `singing` to be relevant. These two taxonomies still are only about 1% below the best result, which shows that the chosen taxonomy is not crucial to the system performances. Nevertheless, a 1% increase over $F = 96\%$ means an error rate decrease of about 25%, thus justifying the choice of an optimal scheme if possible.

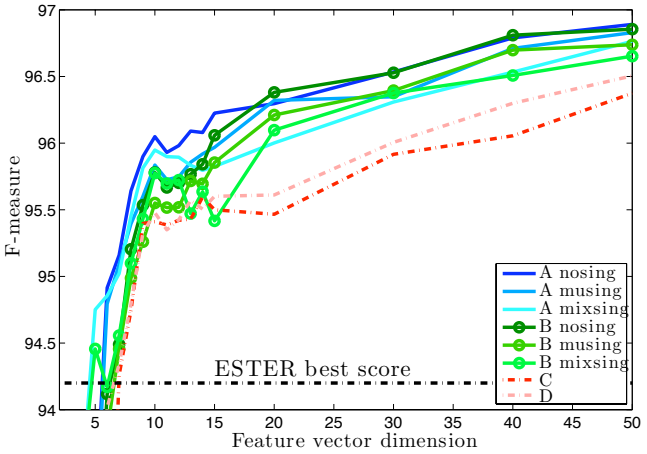

Figure 5: Results on the ESTER corpus test set

There is no significant gap between the A and B schemes except at low dimension ($d < 20$) where the multi-class approach (A) provides better results than the hierarchical one. Moreover, the influence of the `singing` class use in very similar in both cases. At high dimensions ($d > 15$) the system performs best when no `singing` samples are used for training and the union of the `singing` and `mix` samples is less efficient than the more intuitive `singing/music` union. However, at very low dimensions ($d < 7$) the first union performs best with both schemes, probably because the diversity brought by the `singing` samples partly compensates the lack of characteristics expressed with a low dimensional feature vector.

The scores increase with the feature vector dimension, asymptotically approaching a 97% F-measure. All the taxonomies overcome the best score reached during the ESTER campaign [3] (94.2%, see Table 1 for more details) above $d = 7$. This shows the efficiency of the proposed system. Interestingly, some of them ($A^{mixsing}$ and $B^{mixsing}$) even keep a good behavior for a very low dimension ($d = 5$), with F-measures around 94.5%.

Finally, for a very reasonable complexity ($d = 10$), our best and worse system (A^{nosing} and C) provides respective absolute gains of 2% and 1.3% over the best ESTER score. All the systems proposed by the ESTER participants where based on classical MFCC vectors with first and second derivatives, thus using between 33 and 40 features. We show here better performances with lower feature vector dimensions.

Further analysis of this system performances are given in the following section.

4.3 Detailed performances assessment

Table 1 gathers the performances of the 3 best ESTER contestants, compared to our best system (A^{nosing}) at different feature vector dimensions. Even at very low dimension $d = 2$, our system surpasses the 2^{nd} contestant, which tends to show the relevance of the first features selected by the IRMFSP algorithm, and the efficiency of the SVM, applied to this problem. The most significant gain, though, lies in the music class recognition, for

Systems	general			speech			music		
	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
$d = 50$	96.9	2.0	4.5	99.4	13.0	0.5	78.8	1.5	29.6
$d = 10$	96.0	3.0	5.4	99.2	20.2	0.7	73.6	2.1	34.9
$d = 2$	93.3	11.9	4.1	98.9	16.2	1.5	64.8	11.6	20.3
ESTER 1^{st}	94.2	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
ESTER 2^{nd}	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
ESTER 3^{rd}	92.7	11.7	5.7	99.2	36.6	0.7	54.8	10.9	38.7

Table 1: Performances of the A^{nosing} taxonomy at various dimensions compared to the best ESTER results

which our system provides increases of F-measure ranging from 12 to 26%, due mostly to a false rejection rate much lower than other systems, that have generally been designed to maximize speech regions recognition (since all the other tasks of the campaign deal with speech processing).

The confusion matrix is given in Table 2 for the case $d = 50$. To better identify the sources of errors, each global class (e.g. `music`, `mix` and `speech`) was further subdivided in two subclasses for which data was manually annotated. Scores in bold are the global class results. Column *prop* indicates the proportion of the class in the test set, and the last column shows its contribution to the global error rate of 6.5% (i.e. weighted by its proportion). The confusion between speech and pure music is expectedly very low. Indeed the main cause of error is, by far, the confusion of mix for speech (4.6% in the global error). Singing is better identified than pure music and we notice a confusion of noisy speech for mix (8.8%) that is much higher than that of pure speech (0.8%), but is tempered by its weak proportion in its contribution to the global error. The system reaches a correct classification rate of 93%.

Class	**prop**	**mix**	**music**	**speech**	**err**
mix	**12.5%**	**63.0**	**3.5**	**33.5**	**4.6**
music	**3.0%**	**16.6**	**77.3**	**6.1**	**0.7**
music	2.6%	19.7	74.8	6.1	0.6
singing	0.4%	1.4	92.2	6.4	0.1
speech	**84.5%**	**1.2**	**0.2**	**98.6**	**1.2**
speech	80.3%	0.8	0.1	99.1	0.7
sp+noise	4.2%	8.8	1.9	89.3	0.4

Table 2: Confusion matrix for the best configuration

5. CONCLUSION

We have shown here the relevance of SVM for the problem of speech/music segmentation, even at very low dimension. Indeed, the results for 10 selected features are higher than the best score of the ESTER campaign, with all the taxonomies proposed. Higher dimensions almost reach a 97% F-measure, i.e. a reduction of nearly 50% of the segmentation error.

A comparison of the different approaches shows a slight advantage of the approaches with no `singing` training material, but no major difference is observed between the hierarchical and multi-class schemes, except at low dimension. Most of the error is shown to come from a confusion between speech with music and pure speech. Future work will be focused on that specific problem, as well as on noisy speech misclassification.

REFERENCES

[1] A. Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA*, pages 1917–1930, April 2002.

[2] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On kernel target alignment. *JMLR*, 2002.

[3] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ESTER Phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. Interspeech '05*, 2005.

[4] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *NIPS*. The MIT Press, 1998.

[5] D. Istrate, N. Scheffer, and C. Fredouille. Broadcast news speaker tracking for ester 2005 campaign. In *Proc. Interspeech '05*, 2005.

[6] T. Joachims. *Advances in Kernel Methods: Support Vector Machines*, pages 169–184. MIT Press, Cambridge, MA, 1999.

[7] L. Lu, S. Li, and H.-J. Zhang. Content-based audio segmentation using support vector machines. In *Proc. ICME '01*, pages 749–752, August 22-25 2001.

[8] N. Mesgarani, M. Slaney, and S. Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Trans. ASLP*, pages 920–930, May 2006.

[9] G. Peeters and X. Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument database. In *Proc. DAFX '03*, September 8-11 2003.

[10] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 1999.

[11] J.-B. Pothin and C. Richard. A greedy algorithm for optimizing the kernel alignment and the performance of kernel machines. In *Proc. EUSIPCO '06*, September 4-8 2006.

[12] G. Richard, M. Ramona, and S. Essid. Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. In *Proc. ICASSP '07*, pages 461–464, April 15-20 2007.

[13] J. Saunders. Real-time discrimination of broadcast speech music. In *Proc. ICASSP '96*, pages 993–996, May 7-10 1996.

[14] G. Williams and D. Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech '99*, pages 687–690, September 5-9 1999.