ENHANCED SVM TRAINING FOR ROBUST SPEECH ACTIVITY DETECTION

Andrey Temko, Dušan Macho[1], and Climent Nadeu

TALP Research Center, Universitat Politècnica de Catalunya
Barcelona, Spain

{temko, dusan, climent}@talp.upc.es

ABSTRACT

Speech Activity Detection (SAD) is a key objective in speech-related technologies. In this work, an enhanced version of the training stage of a SAD system based on a Support Vector Machine (SVM) classifier is presented, and its performance is tested with the RT05 and RT06 evaluation tasks. A fast algorithm of data reduction based on proximal SVM has been developed and, furthermore, the specific characteristics of the metric used in the NIST SAD evaluation have been taken into account during training. Tested with the RT06 data, the resulting SVM SAD system has shown better scores than the best GMM-based system developed by the authors and submitted to the past RT06 evaluation.

Index Terms— speech activity detection, support vector machines, speech processing

1. INTRODUCTION

In smart-room environments, the availability of a robust Speech Activity Detection (SAD) system is a basic requirement. In the previous work done at our lab [1], a SAD algorithm was developed and compared with other reported techniques using a subset of the SPEECON database. The SAD system was based on a decision tree classifier and Frequency-Filtered (FF) log spectral energies. That system was posteriorly improved [2] by adding two additional features (measures of energy dynamics at low and high frequencies), and by developing two alternative classifiers based, respectively, on Gaussian Mixture Models (GMM) [2] and Support Vector Machines (SVM) [3].

A set of several hundred of thousand of examples is a usual amount of data for classical audio and speech processing techniques that involve GMM. However, it is an enormous number of feature vectors to be used for a usual SVM training process and hardly makes such training feasible in practice. A number of methods of dataset reduction for SVM have been recently proposed. In [4], a Speech / Non-Speech classification with SVM has been done by changing from frame-based to segment-based decisions and computing mean and deviation of all feature vectors inside the chosen segment. The proposed method, however, results in a temporal resolution decrease of the SAD system and thus is better suited to audio indexing (for what it was actually designed) than to SAD. In [5], SVMs have been also applied to the SAD problem using a training set that consists of an arbitrarily chosen small portion of the whole database (12 utterances out of 4914). In [6], a method based on regression trees has been proposed to reduce the available dataset for audio classification, and a cross-training method has been exploited in [7]. Unfortunately, none the above mentioned methods is suitable for our SAD task, either because they show a small ratio of data reduction or they have been applied to relatively small datasets on which it was possible to train a classical SVM. Active learning literature [8] propose several alternatives to deal with moderately large databases, however they involve continuous retraining that with accurate sub-sampling strategy and large initial dataset becomes computationally very expensive.

In this work, the usual training algorithm of the SVM classifier has been enhanced in order to cope with that problem of dataset reduction, proposing a fast algorithm based on Proximal SVM (PSVM) [9]. Besides that, the SVM learning process has been adjusted in order to take into account the specific characteristics of the metric used in the NIST Rich Transcription (RT) evaluations. The resulting SVM SAD system has been tested with the RT06 data, and it has shown better scores than the GMM-based system which, submitted by the authors, ranked among the best systems in the RT06 evaluation.

2. DATABASES

Several databases have been used in this work. A subset of the Spanish SPEECON database, already used in [1][2], was used for classifier training. The single distant microphone evaluation database from the RT05 "conference room" meeting task was used for development in the first stage and for training in the second one. It contains 10 extracts from 10 English language meetings recorded at 5 different sites. Each extract is about 12 minutes long. The proportion of Speech / Non-Speech is highly unbalanced: approximately 90% of the whole signal is Speech.

[1] During the work reported in this paper D. Macho was with the TALP Research Center, Universitat Politècnica de Catalunya; now he is with Motorola Inc., Schaumburg, Illinois, USA.

Table 1. SPEECON, RT05 and RT06 databases summary

DATABASE	SPEECON	RT05	RT06
Language	Spanish	English	English
Type	Single utterances	Conference	Conference & Lecture
Microphone	2-3 m in front of a speaker	On the table	On the table
Signal	16kHz, 16b	16kHz, 16b	16kHz, 16b

For testing we have used the RT06 dataset that consists of two kinds of data, conference meetings ("confmtg") and lecture meetings ("lectmtg"). The "confmtg" dataset is similar to the previously described RT05 data. The "lectmtg" data were collected from lectures and interactive seminars across the smart-rooms of different CHIL (Computers in the Human Interaction Loop) project partners.

SPEECON and the RT data are similar in the sense that they are recorded in closed environments using far-field microphones, thus the recordings have a relatively low SNR due to reverberation and environmental noise. However, there are some differences that should be mentioned: different Speech and Non-Speech proportion and also the fact that the main attention of a speaker in SPEECON was the recording itself, while in the RT databases the recording was secondary. As a consequence, the RT databases are more spontaneous, speakers speak not necessarily heading the microphone, and the data contain overlapped speech. Other features of the databases used in the work are presented in Table 1.

3. FEATURES

The same feature set from [2] was used. The first part of it extracts information about the spectral shape of the acoustic signal in a frame. It is based on Linear Discriminant Analysis (LDA) of FF parameters [1]. The size of the FF representation (16FF+16ΔFF+16ΔΔFF+ΔE=49) is reduced to a single scalar measure by applying LDA. The second part of the feature set focuses more on the dynamics of the signal along the time observing low- and high-frequency spectral components [2].

The contextual information is involved in several ways. First, before applying the LDA transform, the current delta and delta-delta features involve an interval of 50 and 70 ms, respectively, in their calculation. Next, for the representation of the current frame, eight LDA measures are selected from a time window spanning the interval of 310 ms around the current frame. Finally, low and high frequency dynamics involve a smoothed derivative calculation that uses 130 ms interval.

The first and the second part of the feature set form a vector of 10 components. Additionally, for RT06 evaluation task, a cross-frequency energy dynamic feature, which is obtained as a combination of low and high frequency dynamics and was also introduced in [2], is added to the final feature vector.

4. METRICS

As a primary metric we use the one defined for the SAD task in the NIST RT evaluation. It is defined as *the ratio of the duration of* incorrect *decisions to the duration of all speech segments in reference*. We denote this metric as NIST metric in our results.

Notice that the NIST metric depends strongly on the prior distribution of Speech and Non-Speech in the test database. For example, a system that achieves a 5% error rate at Speech portions and a 5% error rate at Non-Speech portions, would result in very different NIST error rates for test databases with different proportion of Speech and Non-Speech segments; in the case of 90-to-10% ratio of Speech-to-Non-Speech the NIST error rate is 5.6%, while in the case of 50-to-50% ratio it is 10%. Due to this fact we also report three metrics that are used for the CHIL project SAD evaluations: Mismatch Rate (MR), Speech Detection Error Rate (SDER), and Non-Speech Detection Error Rate (NDER) defined as:

- **MR** = Duration of Incorrect Decisions / Duration of All Utterances
- **SDER** = Duration of Incorrect Decisions at Speech Segments / Duration of Speech Segments
- **NDER** = Duration of Incorrect Decisions at Non-Speech Segments / Duration of Non-Speech Segments

5. SVM-BASED SPEECH ACTIVITY DETECTOR

A set of several hundreds of thousand of feature vectors hardly makes SVM training process feasible in practice. Alternative methods should be effectively applied to reduce the amount of data. In [2] a hard data reduction was imposed by randomly selecting 20 thousand examples where the two classes of interest are equally represented. In this section we propose two modifications of the SVM training process that aim to improve SAD performance of the SVM classifier from [2]. We use the same preprocessing steps. The training data are firstly normalized anisotropicly to be in the range from –1 to 1, and the obtained normalizing template was then applied also to the testing dataset. In all experiments the Gaussian kernel is used. To train the system the SVMlight software [10] was used.

5.1. Dataset reduction by PSVM

Proximal Support Vector Machine (PSVM) has been recently introduced in [9] as a result of the substitution of the inequality constraint of a classical SVM $y_i(wx_i+b)\geq 1$ by the equality constraint $y_i(wx_i+b)=1$, where y_i stands for a label of a vector x_i, w is the norm of the separating hyperplane H_0, and b is the scalar bias of the hyperplane H_0.

This simple modification significantly changes the nature of the optimization problem. Unlike conventional SVM, PSVM solves a single square system of linear equations and thus it is very fast to train. As a consequence, it turns out that it is possible to obtain an explicit exact solution to the optimization problem [9].

Figure 1 shows a geometrical interpretation of the change. H_{-1} and H_1 planes do not bound the negatively- and the positively-labeled data anymore, but can be viewed as "proximal" planes around which the points of each class are clustered and between which the separating hyperplane H_0 lies.

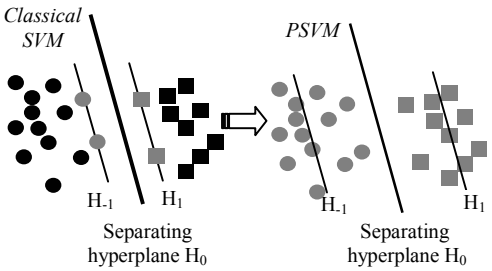

Figure 1. Geometrical interpretation of PSVM

In the nonlinear case of PSVM (we use a Gaussian kernel) the concept of Support Vectors (SVs) (Figure 1, in gray) disappears as the separating hyperplane depends on all data. In that way, all training data must be preserved for the testing stage.

Our proposed algorithm of dataset reduction consists of the following steps:

Step 1. Divide all the data into chunks of 1000 samples per chunk.

Step 2. Train a PSVM on each chunk performing 5-fold cross-validation (CV) to obtain the optimal kernel parameter and the C parameter that controls the training error.

Step 3. Apply an appropriate threshold to select a pre-defined number of chunks with the highest CV accuracy

Step4. Train a classical SVM on the amount of data selected in *Step 3*.

The proposed approach is in fact similar to Vector Quantization (VQ) used for dataset reduction for SVM in [11]. With *Step 2* some kind of clustering is performed, and *Step 3* chooses the data that corresponds to the most separable clusters. However, unlike VQ, SVs, which are obtained with the proposed algorithm in *Step 4*, are taken from the initial data. Besides, additional homogeneity is achieved because the PSVM data clustering is performed in the transformed feature spaces with the transformation functions that correspond to the Gaussian kernel and the same kernel type is applied to the chosen data in *Step 4*. Additionally, as it will be shown in the experimental part, the proposed algorithm gives flexibility to select an efficient dataset for different levels of difficulty of the tested databases.

5.2. NIST metric SVM adjustment

The second modification makes use of the knowledge of the specific NIST metric during the training phase. As it has been mentioned in Section 4, NIST metrics depends on the prior distribution of Speech and Non-Speech in the test database. For this reason, if we want to improve the NIST scores we should penalize the errors from the Speech class more than those from the Non-Speech class. That is possible for a discriminative classifier as SVM in the training stage by introducing different costs for the two classes. In that way, the separating hyperplane H_0 will no longer lie exactly in the middle of the H_{-1} and H_1 hyperplanes (Figure 1). In our case the SVMlight coefficient *j* was fixed to 10.

For a GMM classifier, however, it is possible to favor one of the classes only in the testing stage as it was done in [2]. In that work the final decision was made from the condition $\alpha p_1(x)-(1-\alpha)p_2(x) > 0$, where α is a balancing factor, $p_1(x)$ and $p_2(x)$ are the likelihoods calculated with Non-Speech and Speech GMMs, respectively. When positive, a Non-Speech label is assigned. α was fixed to 0.4. Although it was not done in this work, it is worth to mention that favoring a class in the testing stage could be done for SVM in a similar way through the bias *b* of the separating hyperplane.

6. EXPERIMENTS

6.1. RT05 results

For the RT05 evaluation, the SPEECON database was used for training and development as it was done in [2].

For SVM training we select the same number of data: 20 chunks = 20 thousand samples. Table 2 shows results of the RT05 evaluation with the SVM system, modified according to Sections 5.1 and 5.2, along with the ones obtained with the best SVM and GMM systems in [2].

Table 2. Error rates obtained for RT05 with the modified SVM system

	NIST MR / SDER / NDER
GMM [2]	8.47 7.69 / 4.61 / 38.42
SVM [2]	11.45 10.41 / 7.99 / 34.56
SVM modified	8.03 7.30 / 2.51 / 55.07

From Table 2 we observe that, as it can be expected after the second modification, the NDER score has increased but the SDER score, which has the major influence on the NIST measure, has strongly decreased. In consequence, after both modifications, the NIST error for the modified SVM system decreases from 11.45% to 8.03%, showing comparable results to the best GMM system.

6.2. RT06 results

In [2] for the "confmtg" task and for the GMM classifier both SPEECON and RT05 databases were used for training. For the "lectmtg" task also a small amount of data collected in CHIL was added into training of the "lectmtg" system. For the SVM classifier, the dataset reduction algorithm was applied to the whole database available for training for RT06 task, namely, SPEECON, RT05, and the small amount of CHIL data. Only 10 thousand samples were selected for the final SVM training. Table 3 shows the results obtained with SVM for the RT06 task.

As it can be seen from Table 3, the SVM SAD system while performing well for the "confmtg" task becomes almost a

Table 3. SVM SAD results for two RT06 evaluation tasks

	"confmtg"	"lectmtg"
SVM	**4.88** (4.6 / 0.8 / 72)	13.86 (12.2 / 0.2 / 98)

Table 4. Error rates obtained for the RT06 evaluation for the "confmtg" and the "lectmtg" parts of the database. The results for matched conditions are given in bold.

	NIST MR / SDER / NDER			
	SVM		GMM	
Test / Train	confmtg	lectmtg	confmtg	lectmtg
confmtg	**4.88 (4.6 / 0.8 / 72)**	13.86 (12.2 / 0.2 / 98)	**5.45 (5.1 / 3.1 / 41.4)**	11.71 (10.3 / 0.1 / 83)
lectmtg	11.84 (11.2 / 11 / 14)	**6.16 (5.4 / 1.4 / 33)**	9.54 (9 / 8.2 / 22.4)	**7.1 (6.2 / 0.4 / 48)**

dummy system (the one that says everything is Speech) for the "lectmtg" task with a Non-Speech error rate of 98%. The "lectmtg" part actually is quite different from "confmtg" part and due to the spontaneous character of the former it is more difficult for SAD. As well as a small amount of CHIL data, which can be considered noisier than the RT05 data, was added to the training dataset of the GMM system, we decided to change *Step 3* of the algorithm of dataset reduction and choose for the "lectmtg" training the lowest CV accuracy instead of choosing the highest CV accuracy as it was done for the "confmtg" task.

Table 4 shows the error rate of the GMM and SVM systems for the "confmtg" and "lectmtg" parts of the database. The values in bold in the GMM part were submitted for the NIST evaluations where our GMM SAD system ranked among the best systems.

The diagonal elements of the SVM part show lower error rates than the diagonal elements of the GMM part. That indicates that the proposed algorithm managed to select the appropriate 10000 samples out of the whole training database available that consists of more than 1.5 million examples.

From Table 4 we observe that for the "lectmtg" case the change of *Step 3* of the proposed algorithm has an intermediate influence. Chunks with the lowest CV accuracy, which contain less separable data, are more important for the final classical SVM training in *Step 4* for the given subtask.

Notice that the NIST evaluation scenario allows having an independent system for each subtask so the comparison conditions for the GMM and for the SVM are the same.

On the other hand, the off-diagonal elements of the GMM part from the Table 4 show lower error rates than the off-diagonal elements of the SVM part. That can be either an indication that the GMM is not so sensitive to the unmatching of the training and testing databases or can be the result of the fact that GMM used much larger amount of data for training.

Actually, the off-diagonal elements are not considered in NIST but here we include them to show the behavior of the GMM and SVM classifiers for the case when the characteristics of the training and testing databases do not match.

7. CONCLUSIONS

The presented work is oriented towards robust SVM-based Speech Activity Detection (SAD) systems for smart-room environments.

Two modifications of the usual training algorithm of the SVM-based classifier presented in [2] have been developed in order to cope with two problems of that classifier in our application: the very large amount of training data and the particular characteristics of the NIST metric. With those two modifications, the SVM system has reduced the error rate on the RT05 database from 11.45% to 8.03%, score comparable to the best GMM score of 8.47%. With the RT06 SAD evaluation task, the modified SVM system has achieved an error reduction with respect to the GMM system from 5.45% to 4.88% for the "confmtg" task, and from 7.1% to 6.16% for the "lectmtg" task.

Forthcoming work will be devoted to exploit one of the main advantages of SVM classifiers: to make use of a much longer feature set, e.g. preserving 2 or more LDA measures for each frame.

8. ACKNOWLEDGEMENTS

This work has been partially sponsored by the EC-funded project CHIL (IST-2002-506909) and the Spanish Government-funded project ACESCA (TIN2005-08852).

9. REFERENCES

[1] J. Padrell, D. Macho, C. Nadeu, "Robust Speech Activity Detection Using LDA Applied to FF Parameters", Proc. ICASSP'05, March 2005.

[2] D. Macho, C. Nadeu, A. Temko, "Robust Speech Activity Detection in Interactive Smart-Room Environment", 3rd Joint Workshop on MLMI, May 2006, to appear in LNCS.

[3] B. Schölkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[4] L. Lu, H-J. Zhang, S. Li, "Content-based Audio Classification and Segmentation by Using Support Vector Machines", *ACM Multimedia Systems* 8 (6), pp. 482-492, March 2003.

[5] J. Ramirez, P. Yelamos, J. Gorriz, C. Puntonet, J. Segura, "SVM-Enabled Voice Activity Detection", LNCS v. 3972, pp. 676-681, 2006.

[6] S. Zhang, H. Jiang, S. Zhang, B. Xu, "Fast SVM Training based on the Choice of Effective Samples for Audio Classification", Proc. Interspeech, 2006.

[7] G. Bakr, L. Bottou, J. Weston, "Breaking SVM Complexity with Cross-Training", Proc. NIPS, v. 17, pp. 81-88, 2004.

[8] S. Tong, D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, v.2, pp. 45-66, 2001.

[9] G. Fung, O. Mangasarian, "Proximal Support Vector Machine Classifiers", Proc. KDDM, pp. 77-86, 2001.

[10] SVMlight: http://svmlight.joachims.org/.

[11] G. Lebrun, C. Charrier, H. Cardot, "SVM Training Time Reduction using Vector Quantization," Proc. ICPR, v.1, pp. 160-163, 2004.